
Секция по компютърна лингвистика

Институт за български език

Българска академия на науките

Дисертация за присъждане
на образователната и научната степен „доктор”

---

# Автоматично разпознаване и тагиране на съставни лексикални единици в българския език

---

**Автор:**
Ивелина Стоянова

**Научен ръководител:**
Проф. Светла Коева

София
6 март 2012 г.

# Съдържание

# 1. Увод

## 1.1. Актуалност на проблематиката, свързана със съставните лексикални единици

Статистическите данни показват, че лексикалните единици, състоящи от две или повече графични думи (англ. **multiword expressions**), които в настоящата разработка ще наричаме несвободни фрази, представляват значителна част от лексикалната система на езика. Има редица свидетелства за мащабността на това езиково явление. Според Фелбаум (1998) в WordNet (версия 1.7) 41% от единиците са несвободни фрази, а според Коева (2006а) такива са и 24.49% от единиците в българската семантична мрежа БулНет. Джакендоф (1997) твърди, че броят на несвободните фрази в речника на човек е от същия порядък като единичните думи.

Други оценяват значението на несвободните фрази според според мащабността на тяхната употреба. Мелчук (1998) твърди, че несвободните фрази (които той нарича *фраземи*) преобладават значително и броят им се съотнася към този на единичните думи, както 10 : 1.

В същото време несвободните фрази поставят редица проблеми пред автоматичния анализ на естествените езици, тъй като техните синтактични и семантични характеристики затрудняват автоматичното им разпознаване и анотиране. Решаването на проблемите, свързани с обработката на несвободните фрази, ще подобри значително работата и резултатите на различни приложения за обработка на естествен език като системи за тагиране, машинен превод, автоматично отговаряне на въпроси, автоматично резюмиране на текст. Някои проблеми, свързани с машинния превод на несвободни фрази са демонстрирани в Пример 1.

**Пример 1** (Проблеми при машинен превод на несвободни фрази)**.**

- Избор на правилното значение при многозначни несвободните фрази

  *Това е* ***битово разложение****.* (конструиран пример)
  *This is a* ***bit decay****.*

  *Има редица свидетелства за* ***битовото разложение*** *у господстващата турска класа през XIX век.* (БНК © ИБЕ)
  *There are many testimonies of* ***domestic corruption*** *in Turkish ruling class in XIX century.*

- Коректно разпозната и преведена несвободна фраза

  *Когато числителят на дробта е по-малък от знаменателя, тя се нарича* ***правилна дроб****.* (Уикипедия)
  *When the numerator of the fraction is less than the denominator, it is called a* ***proper fraction****.*

- Некоректно преведена несвободна фраза

  *Полученият резултат е* ***правилна дроб****.* (конструиран пример)
  *The result is* ***right lung****.*

- Несвободна фраза, некоректно преведена като свободна

  *Рационално число, при което числителят е по-малък от знаменателя, е* ***правилна дроб****.* (конструиран пример)
  *Rational number in which the numerator is smaller than the denominator is* ***correct fraction****.*

- Неразпозната несвободна фраза (превод: английски → български)

  *Не* ***took a walk****.* (конструиран пример)
  *Той взе една разходка.*

(Google Translate)

В българската лингвистика и компютърна лингвистика досега не е обръщано голямо внимание на несвободните фрази. Разработките в тази област са съсредоточени основно върху устойчивите словосъчетания. Малкото разработки върху проблемите на несвободните фрази, ги разглеждат като цяло, без да обръщат специално внимание на категориите несвободни фрази.

Настоящата дисертация има за цел да очертае основните теоретични и практически проблеми, които несвободните фрази като цяло и категорията на съставните лексикални единици в частност представят пред компютърния анализ и обработка на езика. Най-общо казано, съставните лексикални единици са отделна категория несвободни фрази, които се отличават с това, че значението им е композирано от значенията на компонентите на фразата (детайли следват в текста).

Дисертацията също така има за цел да направи преглед на основните методи, използвани за разпознаване на съставни лексикални единици, като същевременно представи идеи и експерименти, свързани с приложението им за български език. Разработването на успешна методология за разпознаване на съставни лексикални единици е трудоемка задача, за която все още не е намерено задоволително решение дори за езици като английски.

Разпознаването на съставните лексикални единици заема специфично място в лингвистичния анализ на текста. Автоматичното обработване на текстове включва токънизация, лематизация, тагиране по части на речта, отстраняване на лексикалната и

граматичната многозначност, синтактично парсиране и др. Използваните в обработката компютърни речници обикновено съдържат прости лексеми (от една графична дума), а в някои случаи и ограничен брой устойчиви словосъчетания, най-често фиксирани и полуфиксирани фразеологични съчетания. Тъй като съставните лексикални единици са композирани както семантично, така и синтактично, т. е. характеризират се със синтактична структура, тяхното идентифициране се извършва след тагирането с част на речта и лематизацията. По същество процесът на разпознаване се свързва с отстраняването на многозначността, тъй като определянето на дадена конструкция като съставна лексикална единица я свързва с определено значение и отхвърля другите така наречени 'свободни' значения на компонентите.

Разпознаването на съставните лексикални единици може да се извършва с помощта на лингвистични методи и статистически методи. Най-често се прилагат хибридни методи, съчетаващи лингвистичната информация със статистически тестове. Идентифицирането на съставните лексикални единици е комплексна задача, която е свързана с тяхното разграничаване от другите категории несвободни фрази от една страна, а от друга – от свободните фрази, с които проявяват сходство в структурно и семантично отношение, тъй като фразата е конструирана по синтактичните правила на езика и значението ѝ е композирано от това на конституентите.

Особеност на съставните лексикални единици, която ги приближава до другите категории несвободни фрази, е, че те са с определена конвенционална форма. При тях парафразирането е ограничено (понякога недопустимо); в случаите, когато един или повече конституенти могат да бъдат заместени със синоними, получените изрази са с ниска честота, което показва, че формата е потенциално възможна, но не е реализирана в езика. Ето защо статистическите резултати са от особено значение в анализа на съставните лексикални единици.

Автоматичното разграничаване на съставните лексикални единици от свободните фрази също е трудна задача, която допълнително се усложнява от случаите, когато фразата е многозначна и може да се реализира и като свободна, и като несвободна. Затова за решаването на тази задача особено значение има анализът на контекста.

Правилното разпознаване и тагиране на съставните лексикални единици ще подобри значително работата и резултатите на редица приложения за обработка на естествения език. За целта е необходимо да се дефинира точно обхватът на явлението, обстойно да се изследват особеностите на съставните лексикални единици, да се изгради система за тяхното лингвистично описание и класификация, както и да се разработи методология за разпознаването и тагирането им с релевантна граматична и семантична информация.

## 1.2. Цел, задачи и принципи на изследването

Настоящото изследване се занимава с проблемите на съставните лексикални единици в българския език с оглед на тяхното автоматично разпознаване и тагиране като част от процеса на лингвистичен анализ и анотация на български текстове. Успешното решаване на задачата ще допринесе за подобряване на ефективността на различни компютърни програми за обработка на естествен език.

### 1.2.1. Цел на изследването

Основната цел на дисертацията е изграждането на теоретически модел и неговото практическо приложение за целите на автоматичното разпознаване и тагиране на съставните лексикални единици.

Целта може да се разглежда в теоретичен и практически аспект, които се характеризират с еднаква важност. В теоретично отношение се цели точност и издържаност на модела, а в практическо отношение – качество на методологията и нейното проложение.

### 1.2.2. Задачи на изследването

Във връзка с реализирането на поставената цел се поставят следните основни научни задачи:

1. Да бъде дефиниран обхватът на понятието за съставна лексикална единица, както и да бъдат анализирани граничните случаи, които представляват проблем за автоматичното идентифициране на съставните лексикални единици.
2. Да бъде изграден задълбочен теоретичен модел на съставните лексикални единици, който да описва техните характеристики на различни езикови нива – морфо-синтактични, семантични, синтактични, прагматични.
3. Да се изработи детайлна класификация на категорията на съставните лексикални единици, която да има практическо приложение за тяхното идентифициране, описание и анализ.
4. Да бъде разработена методология за разпознаване на съставните лексикални единици и отделните им категории.
5. Методологията да бъде приложена в практиката, като бъдат анализирани резултатите и възможностите.

### 1.2.3. Основни принципи на изследването

Изследването се основава на ясно поставени принципи, имащи за цел да осигурят успешното провеждане на изследването и да гарантират качеството на получените резултати.

1. **Независимост на изследването**.

   Изследването не е обвързано с конкретна теоретична школа или теория. Същевременно е използван опитът от изследванията на учени от различни области и течения. Целта е да бъде изграден непротиворечив общ модел, който след това може да намери различно място в отделните лингвистични теории.

2. **Обзорност**.

   Изследването има обзорен характер. Тъй като явлението в българския език не е добре изследвано и описвано досега, дисертацията има за цел са постави основните проблеми и да очертае възможни решения.

3. **Последователност**.

   Изследването има ясно формулирани задачи, последователното решаване на които води до изпълнението на целта. Изложението в дисертацията също следва последователното поставяне и решаване на задачите.

4. **Практически характер на изследването**.

   Изследването има практическа насоченост. Теоретичните задачи на изследването имат за цел да осигурят практическата реализация на дейностите, свързани с разпознаването и тагирането на съставните лексикални единици.

   С оглед на провеждането на конкретни практически експерименти в някои случаи се налага ограничаването на обекта на изследване – до съставни лексикални единици именни фрази или такива, отговарящи на определена синтактична структура.

   В теоретичното описание на явлението са включени множество примери от различни корпуси, които имат за цел да демонстрират описваните характеристики и поведението на съставните лексикални единици. Примерите са от реални текстове и подчертават значимостта на проблема, както и практическите измерения на поставените задачи.

   Примерите в текста са на български език, в редки случаи са дадени и чуждоезични примери. В хода на изложението при представяне на чужди разработки чуждите примери са заменени с български, близки до цитираните в чуждите източници. По този начин, когато е възможно, нагледно се показва валидността на представените твърдения и анализи за българския език.

5. **Изпълнимост**.

Важно условие е поставянето на изпълними цели и задачи. Съставните лексикални единици обхващат голяма и разнообразна група явления, поради което в голяма част от изследването изследването обектът беше ограничен до съставните лексикални единици именни фрази. В глава 2 се разглеждат особеностите на всички категории несвободни фрази и се обръща внимание и на другите класове съставни лексикални единици.

## 1.3. Използвани ресурси

### 1.3.1. Едноезикови и многоезикови корпуси за целите на изследването

#### 1.3.1.1. Български национален корпус

Българският национален корпус (БНК © ИБЕ) е разработен в Института за български език и представлява голям репрезентативен корпус на българския език, съдържащ 470 милиона думи (Коева, Благоева, и др., 2011). Текстовете са създадени между 1978 и 2011 година. Корпусът е снабден с подробни метаданни, които съдържат информация за автор и заглавие на текста, източник, година на създаване или издаване, стил, тематична област и жанр на текста и др. (Коева и Стоянова, 2009).

Корпусът е изцяло морфо-синтактично и отчасти семантично анотиран. Също така е създадена система за разширено търсене в корпуса (Тинчев и др., 2008; **Система за разширено търсене в Българския национален корпус** 2011), която е достъпна през интернет и позволява търсене по думи и регулярни изрази.

В настоящото изследване Българският национален корпус е използван основно за извличане за примери, които да илюстрират отделни характеристики на съставните лексикални единици или представени твърдения.

Корпусът също така може да намери приложение в практическото изследване на съставните лексикални единици, където честотните данни от него да се ползват за по-достоверен количествен анализ, който да се основава на по-големи по обем данни.

Част от Българския национален корпус е и Българският Браун корпус, който е създаден от Секцията по компютърна лингвистика (Коева, 2006b) съобразно методологията на известния Brown Corpus, разработен в университета Браун. Корпусът обхваща художествени и информативни текстове и наброява един милион думи. Българският Браун корпус и негови подкорпуси са използвани в някои предишни изследвания върху проблемите на съставните лексикални единици.

#### 1.3.1.2. Корпус от Уикипедия

Специално за целите на изследването беше създаден корпус със статии от Уикипедия (Wikipedia, 2011). Идеята за използването на Уикипедия като корпус не е нова и е намиерила приложение в различни изследвания в областта на компютърната лингвистика (Накаяма и др., 2007; Бекавац и Тадич, 2008; Винче и др., 2011).

Корпусът е съставен автоматично с помощта на компютърна програма (уеб краулер), която последователно обхожда всички страници от българската част на Уикипедия и запазва съответните документи. При запазване на файловете автоматично е създадено и тяхното описание, организирано във форма́та на метаданните на Българския национален корпус (Коева и Стоянова, 2009).

Предпочетено е свалянето на текстовете в XML формат поради съдържащата се допълнителна информация и възможностите за лесна обработка. С тази цел са използвани специалните страници `Специални:Изнасяне`, които се генерират от уикисофтуера при заявка. Първоначално бяха свалени 176,622 текста, които бяха намалени до 118,635, след като бяха премахнати текстовете, които съдържат по-малко от 30 думи (например такива, които сочат към друга страница). Корпусът от Уикипедия наброява общо 41 милиона думи.

Освен за български език бяха свалени и всички текстове на чужди езици, които са посочени като вариант на дадената страница на други езици (линкове към тези страници са дадени в лявото меню, под *На други езици*). Чуждоезичните страници може да бъдат: (1) оригиналът, от който е преведена българската; (2) превод на българската; (3) превод от текст на трети език, от който е преведена и българската страница; (4) независима статия под същото заглавие и на същата тема. Многоезиковият корпус от Уикипедия може да бъде определен като съпоставим (англ. **comparable**) и може да намери приложение в съпоставителни лингвистични изследвания, за тестване на различни езиково независими приложения и др. В настоящото изследване приложението на съпоставимия корпус е ограничено. Той е ползван само за извличане на многоезичен преводен речник на съставни лексикални единици. Настоящата разработка се съсредоточава върху изследването на явлението в българския език. Приложението на чуждоезикови ресурси в анализа остава като възможност за бъдеща работа.

За целите на изследването беше извлечен подкорпус на корпуса от Уикипедия, наречен *Уики1000+*. Тъй като обработването на голям брой малки по размер файлове се извършва бавно, беше решено да се намали броят на файловете, като за новия корпус бяха селектирани текстовете, съдържащи над 1000 думи. Това значително ограничи броя файлове – *Уики1000+* се състои от 6311 текстови единици, наброяващи общо 13.4 милиона думи.

Разпределението на текстовете в *Уики1000+* по тематична област е представено в Таблица 1.1.

| **Тематична област** | **Означение** | **Брой текстове** | **Брой думи** |
|---|---|---|---|
| Археология | A-Archeology | 5 | 10250 |
| Биология | B-Biology | 70 | 134115 |
| Химия | C-Chemistry | 25 | 56,127 |
| Физика | D-Physics | 23 | 47,786 |
| Икономика | E-Economics | 98 | 20,3368 |
| Философия | F-Philosophy | 157 | 342,099 |
| География | G-Geography | 1,102 | 2,267,690 |
| История | H-History | 505 | 1,048,621 |
| Литература | I-Literature | 37 | 66,902 |
| Медицина | J-Medicine | 58 | 117,123 |
| Астрономия | K-Astronomy | 20 | 59,418 |
| Езикознание | L-Linguistics | 18 | 34,649 |
| Математика | M-Maths | 34 | 61,622 |
| Социология | N-Sociology | 14 | 41,878 |
| Психология | O-Psychology | 17 | 31,970 |
| Образование | P-Education | 69 | 125,177 |
| Право | Q-Law | 17 | 34,341 |
| Технологии | R-Technology | 119 | 255,550 |
| Политика | S-Politics | 459 | 1,038,629 |
| Култура | T-Culture | 253 | 502,641 |
| Архитектура | U-Architecture | 12 | 31,116 |
| Спорт | V-Sport | 135 | 315,819 |
| Военно дело | W-Military | 250 | 497,445 |

| **Тематична област** | **Означение** | **Брой текстове** | **Брой думи** |
|---|---|---|---|
| Популярни | Y-Popular | 5 | 7,537 |
| Неопределен | Z | 2,809 | 610,1939 |
| Общо | | 6,311 | 13,433,812 |

Таблица 1.1.: Структура на *Уики1000+* – брой текстове и брой думи по тематична област.

### 1.3.2. Лексикални ресурси

В изследването намериха приложение лексикални ресурси, някои от които са предварително разработени от Секцията за компютърна линвистика, а други бяха автоматично или полуавтоматично извлечени от други ресурси.

За специфични задачи в хода на изследването бяха ползвани различни общи и специализирани лексикони. Такива са списъците с наименования от различни области: лични имена, топоними, дескриптори и наименования на организации и др.

Също така е използван списък с несвободни фрази, които са включени като такива към съответните речникови статии в **Български тълковен речник** (Попов, 1994). Бяха извлечени 2838 единици. От тях бяха филтрирани именните фрази, броят на които е 1050. Ръчно бяха верифицирани единиците и беше определена групата, към която принадлежи – разложими, идиосинкретично разложими или неразложими (повече детайли в 2).

Автоматично бяха извлечени и списъци и речници със съставни лексикални единици от *Уики1000+*. Този процес е описан в 5.2.3.

## 1.4. Структура на дисертацията

Изложението на дисертацията следва изпълнението на поставените задачи и е организирано по следния начин.

Глава 2 представя мястото на съставните лексикални единици в категорията на несвободните фрази. Представени са основните теоретични проблеми, свързани с несвободните фрази, които имат пряко отношение към разработването на методология за лингвистичния им анализ и автоматичното им разпознаване – тяхното място в лексикалната система на езика, особености, класификация. След преглед на основните възгледи е предложена дефиниция и е анализиран обхватът на понятието за състав-

на лексикална единица. Проучени са и връзките с други езикови явления, с които съставните лексикални единици се припокриват или имат сходни характеристики – сложните думи, колокациите, наименованията и терминологията.

В Глава 3 е представен модел за описанието на съставните лексикални единици именни фрази. В модела е включена комплексна лингвистична информация – формообразувателна, семантична, синтактична, прагматична. Представени са и различни класификации на съставните лексикални единици по семантични и синтактични, както и класификация според идиоматичността, която е определена като приложна класификация, тъй като отразява различията в подходите към третирането на отделните категории при обработката на езика.

Принципите и основните методи, използвани за разпознаване на съставните лексикални единици са разгърнати в Глава 4. Представени са някои основни лингвистични, количествени (статистически) и хибридни методи, като се изтъква, че третата категория дават най-добри резултати и намират най-широко приложение. В тази глава се анализират и факторите, които оказват влияние върху успешното прилагане на методите и качеството на резултатите.

Глава 5 описва серия от експерименти, които демонстрират отделните етапи на автоматичното разпознаване и тагиране на съставни лексикални единици в българския език. За различните подзадачи са приложени различни лингвистични и количествени методи. В главата също така е направен кратък преглед на налични компютърни програми, които включват във функционалностите си такива за разпознаване и определяне на несвободни фрази. След този преглед е представена новоразработената за целите на дисертацията компютърна програма `bgMWE`, която включва набор от инструменти за обработка на текстови корпуси, разпознаване на съставни лексикални единици по няколко различни метода и отчитане на резултатите.

Глава 6 е посветена на резултатите от изследването. Направени са някои основни изводи от работата, представени са оригиналните приноси на разработката в областта на компютърната лингвистика и са очертани някои насоки на бъдещата работа по темата.

Представена е и подробна библиографията, която съдържа 115 единици – 9 електронни източника за компютърни системи, 20 български и 86 англоезични източника.

Към дисертацията има четири приложения, които включват списък с примерите в текста, пример, демостриращ обработката на текстовете, списък с електронните ресурси, приложени към дисертацията на електронен носител и списък на допълнителните компютърни приложения, които са разработени за изпълнение на различни задачи и подзадачи на изследването.

# 2. Теоретично представяне и анализ на съставните лексикални единици

Настоящата глава разглежда теоретичните проблемите на съставните лексикални единици, които имат пряко отношение към разработването на методология за лингвистичния им анализ и автоматичното им разпознаване. Направен е преглед на използваната в литературата терминология и е уточнен терминологичният апарат, който ползва дисертацията. След това се разглежда мястото на несвободните фрази в лексикалната система на езика, като се обръща внимание на техните особености и класификация, а също така и на начина, по който задачата за разпознаване на съставни лексикални единици се вписва в по-общата задача за идентификация на несвободни фрази.

В 2.3 детайлно са представени съставните лексикални единици. След преглед на основните възгледи е предложена дефиниция и е анализиран обхватът на понятието. По-нататък са проучени връзките с други езикови явления, с които съставните лексикални единици се припокриват или имат сходни характеристики. Такива са наименованията, колокациите и терминологията, които поставят сходни проблеми за автоматичната обработка на естествените езици и за разпознаването на които се използват същите или подобни методи.

Обемът и разнообразието на представената в тази глава литература по въпросите на несвободните фрази и съставните лексикални единици не се подчинява на изискване за изчерпателност, а има за цел да очертае основните възгледи за разглежданите езикови явления. Ето защо по-детайлно внимание е обърнато на текстовете, които дават основата на възприетата дефиниция и разработената класификацията на съставните лексикални единици.

## 2.1. Преглед на основната терминология

Първите изследвания върху колокациите, контекста, в който думите се реализират, и комбинациите от думи, които системно се появяват в близост една до друга, датират от 50-те години на 20 век. Лингвисти от различни направления и области (теоретична и компютърна лингвистика, структурализъм и функционализъм и др.) обръщат внимание на тези явления и много често използват различен терминологичен апарат за тяхното описание. За целите на по-нататъшното изложение е необходимо да бъ-

де изработена единна в концептуално отношение и непротиворечива терминологична рамка, която да дава възможност за точно и ясно описание на изследваните езикови явления, като същевременно се вписва успешно в установените традиции. Подходът при описанието на терминологията ще бъде от по-общото към по-конкретното – от понятието за колокация към видовете съставни лексикални единици.

При компютърния анализ текстът минава през две основни фази на обработка – първата (предварителна) е обработването на текста в суров вид на графично или звуково ниво, а втората (основна) включва неговия лингвистичен анализ на различни равнища. При предварителната обработка писменият текст се разделя на графични думи и пунктуационни знаци, наречени **токъни** (англ. **token**). В общия случай комбинациите от токъни се наричат $N$**-грами** (англ. $N$**-grams**), като за комбинации от два токъна се използва названието **биграм** (англ. **bigram**), а за три – **триграм** (англ. **trigram**). $N$-грамите са произволни съчетания от токъни и не се характеризират с определено значение. Най-често срещаният биграм в Българския Браун корпус (вж. 1.3.1) например е *да се*, и двата компонента на което не са пълнозначни думи. За $N$-грамите може да се използва и названието **комплекс** или **съчетание** от думи.

След обработването на текста на графично ниво, се пристъпва към разглеждането му като езикова единица на ниво лингвистичен анализ, където съответствията на токъните са **думите** – всяка дума представлява единство от (слово)форма и лема, към която се отнася формата, като също така се характеризира с принадлежност към част на речта, набор от граматически характеристики и речниково значение. Синклер (1998), Стъбс (2002) и др. използват термина **лексикална единица** (англ. **lexical item, lexical unit**), за да разграничат думата, много често възприемана като графична дума, и лексикалната единица като комплекс от форма и значение.

В тази връзка е необходимо да изтъкнем и разграничението между теоретичен и практически подход към анализа на несвободните фрази и съставните лексикални единици в частност. Практическият подход изхожда от текста или корпуса от текстове и чрез анализ на различни равнища цели да идентифицира търсените единици – като етапите на обработка могат да включват: предварителна обработка на текстовете; идентифициране на кандидати; филтриране на кандидатите. Теоретичният подход от друга страна, изхожда от разглежданите единици, които са обект на изследване, и има за цел да очертае техните основни характеристики и да изгради система за тяхното описание, да проучи тяхната реализация в текста, с което да подсигури разработването на практически методи за тяхното идентифициране.

Тъй като основните разработки по въпросите на несвободните фрази могат да се поделят на теоретически и практически въз основа на техния подход към анализа на тези езикови явления, част от използваната терминология попада и в двете групи разработки, но с различно значение. Такъв е примерът с термина **колокация**.

Терминът **колокация** (англ. **collocation**) е използван широко в традицията на

Фирт (1957) и последователите му (Халидей, 1966; Синклер, 1991) и описва комбинация от (две или повече) думи, които проявяват склонност да се появяват заедно или в близост една до друга. Обикновено в колокацията се различават ядро (англ. **node**) и **колокати** (англ. **collocates**). Тук приемаме емпиричното разбиране за колокациите като последователност от думи, които директно могат да бъдат наблюдавани в текста, като го разграничаваме от теоретичната интерпретация на колокатите като лексикализирани идиосинкретични комплекси от думи (за подробности вж. Еверт, 2007). Мелчук (1995) също използва термина 'колокация', но в подчертано теоретично и стеснено значение за означаване на определена категория несвободни фрази, при които не са налице семантични ограничения, но се наблюдават прагматични и статистически. Това стеснено разбиране за 'колокация' не се прилага в настоящата разработка, в която ще използваме термина в неговото практическо значение.

Чуека (1988) дава дефиниция за **колокативен израз**: синтактично и семантично оформена единица, чието значение не може да бъде изведено пряко от значението на компонентите ѝ. За разлика от дефиницията за колокация, която се обляга единствено на емпирични наблюдения за съчетаемостта на думите, в дефиницията за колокативен израз присъства и лингвистичен компонент – съчетанието от думи трябва да отговаря на условието за синтактична и семантична оформеност, като същевременно демонстрира ограничения в определен аспект (семантична неделимост, недопустимост на модификации или на синтактични вариации, силни прагматични конотации и др.).

Най-общо на базата на отсъствието/наличието на синтактични, семантични и прагматични ограничения в реализацията различаваме **свободни фрази** (англ. **free phrases**) и **несвободни фрази** (англ. **non-free phrases**). При свободните фрази, които накратко могат да се наричат и само **фрази** или **словосъчетания**, не се наблюдават специфични ограничения – при тях се допускат вариации, синонимни замени на части от фразата, модификации и др. Несвободните фрази от друга страна търпят ограничения в един или повече аспекти. В англоезичната литература широко застъпен е терминът **единица от повече от една дума** (англ. **multiword expression**, **multiword unit**), но в дисертацията даваме предпочитание на термина 'несвободна фраза', въведен от Мелчук (1995), тъй като от една страна, експлицитно се противопоставя на понятието за свободна фраза, а от друга – има общ характер и е подходящ за описване на широкия набор от специфични явления, които се причисляват към тази категория. Не на последно място по важност е и фактът, че това е терминът, използван в някои съвременни български изследвания в областта (Тодорова, 2006, 2007, 2010; Тодорова и Обрешков, 2008). Мелчук (1995) използва също термините **устойчива фраза** (англ. **set phrase**) и **фразема** (англ. **phraseme**), а в предишна статия (Стоянова, 2008) е използван терминът **свързана фраза** като синоними на 'несвободна фраза'. В някои случаи може да се използва и **лексикална единица от повече от една дума** или **лексикална единица от повече думи** като превод на

**multiword expression**, но тези термини ще бъдат избягвани поради описателния си характер.

Бояджиев и др. (1998, част 2, Лексикология) използват термините **фразеологични изрази** и **устойчиви фрази** за съставните лексикални единици. В тази група се причисляват устойчивите съставни названия, описани като *'максимално точни наименования на характерните за предметите и явленията особености'*. Трябва да се подчертае, че класът на съставните лексикални единици е значително по-широк от този на съставните термини и включва и лексикални единици с общо, нетерминологично значение.

Според класификацията на Болдуин, Банард, Танака и Уидоус (2003) несвободните фрази се разделят на **неразложими** (англ. **non-decomposable**), **идиосинкретично разложими** (англ. **idiosyncratically decomposable**) и прости разложими (англ. **simple decomposable**). Българския превод на терминологията е заимстван от Тодорова (2007; 2010). Наред с тези термини ползваме съответно **некомпозирани**, **идиосинкретично композирани** и **прости композирани**.

Неразложимите фрази се характеризират с неделимост на значението, което се изразява в това, че значението на цялата фраза не може да се опише като формирано от значенията на конституентите. Неразложимите фрази в българската езиковедска литература традиционно се наричат **фразеологизми**, **фразеологични единици**, **фразеологични съчетания**, **устойчиви съчетания** (БАН, 1983; Ничева, 1987; Бояджиев и др., 1998).

Нунберг, Саг и Уасоу (1994) използват термина **идиом** (англ. **idiom**) за назоваването на несвободните фрази, при които се наблюдава неделимост на значението, и говорят за **идиоматичността** (англ. **idiomaticity**) като основен признак на идиомите, проявяващ се в три измерения: **конвенционалност** (англ. **conventionality**), **яснота** (англ. **transparency**) – противопоставено на **неяснота** (англ. **opacity**), и **композиционалност** (англ. **compositionality**). Терминът 'идиом', използван от Нунберг и др., 1994, обхваща неразложимите и идиосинкретично разложимите единици, докато в други разработки терминът 'идиом' се използва с по-тясно значение за назоваване само на неразложимите единици (Катц, 1973; Чомски, 1980). Поради многозначността на този термин, неговата употреба ще бъде ограничена.

Нунберг и др. (1994) разделят идиомите на **идиоматични фрази** (англ. **idiomatic phrases**) и **идиоматично комбинирани изрази** (англ. **idiomatically combining expressions**), които покриват съответно категориите на неразложимите и идиосинкретично разложимите несвободни фрази от класификацията на Болдуин и др. (2003).

Обект на изследване в настоящата дисертация са **съставните лексикални единици**, които Болдуин и др. (2003) определят като разложими несвободни фрази. При тях значението на фразата е композирано от значенията на конституентите и семантичната разложимост/композиционалност е основният критерий за разграничаване

| **Основен термин** | **Английски** | **Синоними** |
|---|---|---|
| токън | token | графична дума |
| дума | word,<br>lexical item,<br>lexical unit | лексикална единица |
| $N$-грам | $N$-gram | комплекс,<br>съчетание |
| колокация | collocation | |
| свободна фраза | free phrase,<br>phrase | фраза,<br>словосъчетание,<br>свободно словосъчетание |
| несвободна фраза | non-free phrase,<br>phraseme,<br>set phrase,<br>multiword expression,<br>MWE,<br>multiword unit | лексикална единица от повече от една дума |
| неразложима фраза | non-decomposable phrase,<br>idiomatic phrase,<br>frozen expression | идиоматична фраза,<br>некомпозирана фраза,<br>фразеологизъм,<br>фразеологично съчетание,<br>устойчиво съчетание |
| идиосинкретично разложима фраза | idiosyncratically decomposable phrase,<br>idiomatically combining expressions | идиоматично комбиниран израз,<br>идиосинкретично композирана фраза |
| съставна лексикална единица | decomposable phrase,<br>compound,<br>extended lexical item,<br>pragmateme,<br>pragmatic phraseme | разложима несвободна фраза,<br>съставна дума,<br>композирана фраза,<br>прагматема,<br>прагматична фразема |
| свободна колокация | free collocation<br>purely statistically marked phrase | |

Таблица 2.2.: Основна терминология, използвана в дисертацията.

на съставните лексикални единици от останалите групи несвободни фрази – неразложимите и идиосинкретично разложимите. Същевременно обаче съставните лексикални единици се различават от свободните фрази по това, че притежават единно и неделимо значение и търпят лексикални, синтактични, семантични и прагматични ограничения.

Терминът **съставна лексикална единица** е предпочетен пред **разложима фраза** или **фразема**, тъй като подчертава статута на явлението като част от лексикалната система на явлението, но от друга страна изтъква съставния им характер, което има значение за изграждането на подходящи модели на тяхното изследване и описание. **Съставна лексикална единица** е предпочетен и пред срещания в литературата термин **съставна дума** най-вече поради многото значения на термина **дума** и асоциирането му с графичната дума.

Отделно се разглежда и една гранична категория несвободни фрази – категорията на **свободните колокации**. Терминът се въвежда за първи път тук и основната му роля е да разграничи онези колокации, при които не се наблюдават езикови ограничения (семантични, синтактични или прагматични) и в този смисъл не могат да бъдат определени като съставни лексикални единици, а се проявяват само като статистически маркирани – което се изразява във висока честота на поява на съчетанието спрямо други възможни съчетания със същото значение, недопустимост на синонимни замени на конституентите и др. Статистическите ограничения, които могат да бъдат определени и като проява на признака институционализираност, обособяват тази категория като несвободни фрази, но по лингвистични характеристики те могат да бъдат определени като свободни. Новоконструираният термин **свободна колокация** изтъква характера им на колокации, като същевременно насочва и към това, че са свободни фрази.

Саг и др. (2002) наричат свободните колокации **институционализирани** (англ. **institutionalized**), но тук не приемаме тази употреба на термина, тъй като институционалността е признак на всички несвободни фрази и не е специфичен за тази категория.

В теорията на Мелчук (1995; 1998) фраземите се разделят на **семантични** (англ. **semantic phrasemes**) и **прагматични** (англ. **pragmatic phrasemes**) или **прагматеми** (англ. **pragmatemes**). Класът на прагматемите се характеризира с рестриктивност, но регулярност на означаваното, както и с рестриктивност, но регулярност на означаващото. В този смисъл прагматемите обхващат класа на съставните лексикални единици, заедно със свободните клокации.

Разграничаването на двата класа е от съществено значение за изграждането на сполучливи методи за автоматично разпознаване и тагиране на съставните лексикални единици, тъй като те проявяват сходни характеристики в текста (честота на колокациите, ограничаване на синонимни замени), но са две отделни категории от лингвистична гледна точка поради семантичните, синтактичните и прагматичните

си характеристики.

Таблица 2.2 представя в обобщен вид терминологията, използвана в дисертацията. Анализ и подробни дефиниции на основните понятия ще бъдат представени понататък в 2.2 и 2.3.

## 2.2. Място на несвободните фрази в лексикалната система на езика

За да бъде очертано мястото на несвободните фрази в лексикалната система на езика, ще представим основните виждания в българската и чуждестранната литература за тяхната позиция, особености и класификация, след което ще обърнем внимание на значението на несвободните фрази и техните характеристики за определянето и описанието на съставните лексикални единици.

### 2.2.1. Основни възгледи за несвободните фрази

В българската езиковедска литература многократно са поставяни въпросите за несвободните фрази, но най-задълбочено внимание се обръща на фразеологизмите.

В **Граматика на съвременния български книжовен език** (1983, том 3, Синтаксис) се говори за свободни и устойчиви (фразеологични) словосъчетания. Устойчивото словосъчетание се описва като:

> *... завършена и устойчива фразеологическа единица, равнозначна на една дума и най-често с преносно значение. В устойчивите словосъчетания думите изгубват или замъглят не само конкретните си лексикални значения, но и синтактичните си отношения, затова на тези словосъчетания трябва да се гледа като на неразчленени синтактични цялости.*

Изтъква се, че фразеологичните съчетания са обект на лексикологията, а не на синтаксиса. Детайлно описание и анализ на устойчивите съчетания са представени от Ничева (1987).

В **Съвременен български език: Фонетика, лексикология, словообразуване, морфология, синтаксис** (Бояджиев и др., 1998, част 2, Лексикология) фразеологичните съчетания се описват като част от лексикална система на езика. Още повече, се разграничават три вида значения на думите – фразеологично свързани, синтактично обусловени и конструктивно обусловени значения. Това хвърля светлина върху механизмите на образуване на фразеологичните съчетания – реализирането на думите в съчетание с други думи с определена семантика и в определена синтактична позиция, като компонентите не се реализират с типичните си значения и отношенията между тях не са обусловени от логическите отношения. Употребата на фразеологичните съчетания се установява от традицията и широката многократна

употреба. Тези единици се характеризират с определено значение, което не се или само частично се определя от значението на компонентите. Изтъква се също така, че в състава на фразеологичните единици думите не изгубват лексикалното си значение напълно, а само частично се десемантизират, като значението на цялата единица е обобщено и преносно, с богати смислови оттенъци и сложен характер. Граматическите им характеристики се отнасят до цялата единица, а не до отделните компоненти, съставът и формата им е установен и в повечето случаи не се допускат вариации.

Част от мотивацията за разглеждане на фразеологичните единици като предмет на лексикологията е тяхната прилика с думата като основна лексикална единица. При тях се наблюдават същите семантични явления – например полисемия, както и семантични отношения – синонимия, омонимия и др. (Пример 2)

**Пример 2.**
(Бояджиев и др., 1998, част 2 Лексикология, 8.2)

- полисемия

  *бия си главата*

  1. мъча се да разбера, да проумея, да разреша нещо;
  2. съжалявам много за нещо, случило се по моя вина; горчиво се кая, разкайвам се за нещо.

- омонимия

  1. *вдигам ръка* (давам знак, че искам да говоря)
  2. *вдигам ръка* (посягам да ударя)

- синонимия

  1. *въртя по гайдата*
  2. *играя по свирката*

  (изпълнявам безропотно заповедите на някого)

След логически анализ на дефинициите за дума, графична дума, понятие, Коева (2006а) достига до следното определение за несвободна фраза:

*A MWE is a sequence of two or more graphical words that denotes a unique and constant concept... we can accept that a MWE denotes a concept iff a relation of equivalence exists between it and a single word from a natural language.*

*Лексикална единица от повече думи е последователност от две или повече графични думи, която означава уникално и еднозначно определено понятие. ... Можем да приемем, че лексикалната единица от повече думи означава едно понятие тогава и само тогава, когато*

*съществува отношение на еквивалентност между нея и дума от естествен език.*

Това твърдение отразява една от основните насоки, в която се проявяват проблемите с разпознаването и третирането на несвободните фрази – машинния превод от един на друг език. Ако едно понятие се описва в езика на оригиналния текст със словосъчетание, а в езика, на който превеждаме, съществува като единична лексема, за да се гарантира сполучлив превод, е необходимо да се установи съответствието между двете, като за целта словосъчетанието в оригиналния текст е разпознато като една цялост.

Примерите от различни езици показват, че несвободните фрази са езиково специфично явление. Това е логично следствие от факта, че те се утвърждават по прагматични и социолингвистични механизми. Например, в ескимоските диалекти инуктитут (англ. **Inuktitut**) се срещат над десет думи за сняг и лед: *pukak* (сняг на кристали), *masak* (мокър падащ сняг), *qiqumaaq* (сняг със замръзнала повърхност) и др. (Нунавут, 2000). На български език тези понятия се изразяват с описателни словосъчетания, които могат да се определят като съставни лексикални единици, защото означават едно понятие.

Многобройни са и примерите за липса на лексикализация в български на понятия от WordNet (3).

**Пример 3.**

| | | |
|---|---|---|
| flagellate:1, scourge:2 *'whip'* | ↔ | бичувам:2 |
| birch:1 *'whip with a birch twig'* | ↔ | няма лексикализация *'бия с брезова бръчка'* |
| cat:1 *'beat with a cat-o'-nine-tails'* | ↔ | няма лексикализация *'бия с бич, камшик от девет върви'* |

(Примерът е от Коева, 2007a.)

Други съвременни разработки в българската литература по въпросите на несвободните фрази са изследванията на Тодорова (2006, 2007, 2010), които разглеждат глаголните фразеологизми в българския език, но също така представят подробна класификация на несвободните фрази, като се очертават и техните семантични и синтактични особености. Голяма част от терминологията в настоящата разработка е координирана с тези изследвания.

Българската езиковедска литература върху несвободни фрази е ограничена – както върху теоретични те проблеми, които те поставят, така и в приложна насока – върху въпросите за автоматичното им разпознаване и тагиране. Значително по-широко застъпени са разглежданите проблеми в чуждоезиковата литература, на която ще бъде обърнато внимание по-нататък.

Круз (1986, стр. 49) въвежда разграничение между **лексеми** (англ. англ. **lexemes**), които се дефинират в речника, и **лексикални единици** (англ. англ. **lexical units**), за които дава следната дефиниция:

*... form-meaning complexes with (relatively) stable and discrete semantic properties which stand in meaning relations ... and which interact syntagmatically with contexts in various ways...*

*... комплекс от форма и значение с (относително) стабилни и обособени семантични особености, който участва в семантични релации ... и взаимодейства синтагматично с контекста...*

Като основна характеристика на лексикалната единица се изтъква това, че се състои от поне една дума, а думата се определя от два основни признака – тя е минимална единица, която притежава мобилност в изречението, и максимална единица, която не позволява вмъкване на други думи между частите ѝ.

В семантичното описание на лексикалните единици Круз (1986) въвежда признака **семантична яснота** (англ. **semantic opacity**), който се описва със степени между яснота и неяснота в значението. Въз основа на този признак се отделят от една страна нефразеологичните съчетания, които се характеризират с яснота на значението (т.е. значението е съставено пряко от значения на конституентите), и от друга страна фразеологичните съчетания или идиоми, които се характеризират с едно неделимо значение.

По-нататък Круз (1986) въвежда и синтактични елементи в описанието на лексикалните единици. На първо място изтъква, че синтактичното поведение на фразеологизмите се определя от синтактичната структура на фразата, разглеждана като пълнозначна (т.е. приемайки, че значението е ясно), и от друга страна от факта, че синтактичните елементи не са семантични конституенти и затова не допускат модификация или друга семантична трансформация, както и не участват самостоятелно в семантични релации като синонимия, омонимия и др. Степента на яснота в значението зависи също от това, кои конституенти и до колко са семантични индикатори. Поради характеристиките на идиомите Круз (1986) заключава, че те трябва да бъдат включени в речника.

Синтактичните съчетания, които имат ясно значение, което е композирано от значенията на конституентите, често също показват характеристики, сходни с тези на идиомите. При колокациите например, макар да има яснота на значението, много често се наблюдава семантична цялост и взаимозависимост между компонентите, т.е. те не могат да се заменят или модифицират. Някои от тези колокации, наречени

свързани колокации, проявяват и синтактична константност – не допускат или силно ограничават вмъкването на други единици между частите. (Круз, 1986)

Синклер (1998) използва термина **лексикална единица** (англ. **lexical item/unit**), за да подчертае разликата между дума и единица, носеща значение, която може да съдържа повече от една дума. Основен момент в разработката е представянето на връзката между парадигматичното и синтагматичното ниво на анализ на думите. Извън контекста думата се разглежда като една парадигматична възможност, натоварена с широко общо значение, а от друга страна като компонент в синтагматична структура тя притежава тясно значение, което дава еднозначна информация.

Ето защо Синклер предлага модел, който описва лексикалните единици чрез пет основни компонента: два задължителни и три незадължителни. Задължителните са **ядрото** (англ. **core**), което е постоянно, и **семантичната прозодия** (англ. **semantic prosody**), която определя значението на цялата единица. Чрез незадължителните категории се реализират координирани вторични промени и настройки в значението на единицата, като по този начин се осъществява връзката между парадигматичното и синтагматичното ниво. Незадължителните категории са **колокация**, **колигация** и **семантично предпочитание** (Синклер, 1998).

На синтагматично равнище колокацията е съвместната поява на две думи на определено разстояние една от друга в текста. На парадигматично равнище колокацията се анализира като възможността две думи от една парадигма да се реализират в един и същи контекст. Колигацията е съвместната поява на граматически явления – по-конкретно на класове думи и части на речта, с определена дума или фраза. На парадигматично ниво, подобно на колокацията, колигацията се изразява във възможността за представяне с конструкция с различни граматически характеристики, например *читателски дневник – дневник на читателя*. Семантичното предпочитание се свързва с ограниченията, които се наблюдават върху съвместната поява на думи, които имат сходна семантика, и също се отнася както към синтагматичното, така и към парадигматичното ниво. Синклер, 1998

В труда на Стъбс (2002) анализът изхожда от колокацията, дефинирана по следния начин Стъбс, 2002, стр. 29:

*... a node-word co-ocurring with collocates in a span of words to left and right... 'Collocation' is a frequent co-occurrence.*

*... дума ядро, появяваща се с колокати, които са в определен спан наляво и надясно от ключовата дума... 'Колокация' е честа съвместна поява.*

По-нататък авторът подчертава, че графичната дума невинаги е основната единица, носител на значение. Той въвежда понятието **разширена лексикална единица** (англ. **extended lexical unit**), с което се описват несвободните фрази, представящи единица значение. За характеризиране на лексикалните единици се изтъкват тех-

ните семантични, лексикални, психолингвистични и социолингвистични особености, като наред с това се набляга на статистическите показатели за свързаност между отделните елементи на лексикалните единици.

Стъбс (2002) също така изгражда модел на разширените лексикални единици, който е представен в 2.2.2.

В същата насока са наблюденията на Нунберг, Саг и Уасоу (1994). Те описват идиомите според степента на проява на три основни семантични признака:

(1) конвенционалност;

(2) яснота/разбираемост на значението; и

(3) композиционалност.

Конвенционалността се изразява в степента на несъответствие между идиоматичното значение на фразата и буквалното ѝ значение, ако тя бъде приета за свободна. Яснотата е признакът, който показва доколко е възможно да се види мотивацията за употребата на фразата в дадена ситуация. Третият признак, композиционалността, описва степента, до която фразеологичното значение може да се представи чрез значенията на частите на фразата.

Трите признака в комбинация определят доколко значението на разширената лексикална единица може да бъде анализирано като функция от значенията на конституентите. Авторите изтъкват факта, че не всички конституенти губят самостоятелното си значение, а и тези, които го правят, не го губят напълно.

Основно твърдение в статията е, че значението на идиома е композирано от значенията на съставните му части, като в някои случаи конституентите са реализирани с нетипично, идиосинкретично значение. Идиомите наследяват граматическите характеристики от фразата като цяло и конституентите определят възможностите за формообразуване (където може да се наложат семантични, прагматични и други ограничения), както и възможностите за модификации, добавяне на квантификатори и др. (Нунберг и др., 1994)

Мелчук (1995) въвежда понятието **несвободна фраза**, като използва също и названията **устойчива фраза** и **фразема** (вж. 2.1). Свободните фрази се означават с $A \oplus B$ и се отличават с **нерестриктивност** и **регулярност** (англ. **unrestrictedness** и англ. **regularity**) във формата и значението.

Мелчук (1995, стр. 175) описва свободните фрази по следния начин:

1. *Its signified $'X' =' A \oplus B'$ is unrestrictedly and regularly constructed on the basis of a given ConceptR ... out of the signified $'A'$ and $'B'$ of the*

1. *Означаваното $'X' =' A \oplus B'$ е нерестриктивно и регулярно конструирано на базата на дадено ConceptR ... от означаваните $'A'$*

*lexemes* **A** *and* **B** *of the language* $\mathcal{L}$.

2. *Its signifier* $/X/ = /A \oplus B/$ *is unrestrictedly and regularly constructed on the basis of the SemR ... out of the signifiers* $/A/$ *and* $/B/$ *of the lexemes* **A** *and* **B**.

*и* $'B'$ *на лексемите* **A** *и* **B** *от езика* $\mathcal{L}$.

2. *Означаващото* $/X/ = /A \oplus B/$ *е нерестриктивно и регулярно конструирано на базата на SemR ... от означаващите* $/A/$ *и* $/B/$ *на лексемите* **A** *и* **B**.

**ConceptR** представлява извънезиковата понятийна репрезентация на израза.

**SemR** от друга страна е семантичната репрезентация на единицата и включва семантичната ѝ структура и отношенията на елементите в нея.

За разлика от свободните фрази, при несвободните значението не се формира нерестриктивно и регулярно, а се наблюдават по-сложни отношения в семантичната структура на фразата. Мелчук (1995, стр. 179) въвежда също понятието за **доминантен семантичен възел** (англ. **dominant semantic node**), до който семантичната структура може да се редуцира. В 2.2.2 е обърнато повече внимание на описанието и класификацията на фраземите, представени в Мелчук, 1995.

Мелчук (1998) подчертава, че фраземите не са композирани – че те не могат да бъдат конструирани от ConceptR чрез прилагане на общите езикови правила. Поради това фраземите се считат за лексикални единици и е необходимо да се описват в речника. Същевременно, според Мелчук в езика преобладават фраземите и броят им се съотнася към думите, както $10 : 1$, но преобладаващата част от тях се причислява към категорията, която той назовава 'колокации', съответстваща на приетия в дисертацията термин 'свободни колокации'.

Муун (1998) описва същността на несвободните фрази като комплекс от признаци, които си взаимодействат по различни начини. Представлява интерес и твърдението, че несвободните фрази представляват широк континуум от некомпозирани (идиоматични) и композирани лексикални единици. Това твърдение насочва към подходи, които не търсят ясно определени граници между езиковите явления, обобщени под 'несвободни фрази', а по-скоро търсят мярка за нивото на композираност на фразите.

Еверт (2005, 2007) дефинира несвободните фрази, като изхожда от колокациите. Неговият подход е подчертано практически и е свързан с основната задача за автоматично извличане на колокации и несвободни словосъчетания. Различават се две значения на названието 'колокация' – по-широко (**емпирична колокация** – всяка комбинация от две или повече думи, появяваща се с определена честота в текст/корпус) и по-тясно значение (специфична категория несвободни фрази). Отбелязани са и основните характеристики на колокациите в тясното значение на термина: липса на композиционалност, недопускане на модификации, устойчива структура.

В обобщение може да се каже, че се очертават два основни подхода към описанието

и анализа на съставните лексикални единици. Единият подход е подчертано теоретичен и изхожда от лексикалната система на езика, като се опитва да очертае мястото на несвободните фрази в тази система (БАН, 1983; Круз, 1986; Мелчук, 1995). Тези теоретични изследвания на несвободните фрази разглеждат характерните особености, които отличават несвободните от свободните фрази – особености на семантичната и синтактичната структура, различни степени на десемантизация на компонентите, ограничена вариативност, конвенционална употреба.

Другият подход е ориентиран към употребата на несвободните фрази и идентифицирането им в текстове и езикови корпуси. При този подход се изхожда от колокациите като емпирично установимо явление и се разглеждат типовете колокации и лексикалните единици в контекст, като след това се пристъпва към разглеждане на семантичните и синтактичните им характеристики (Синклер, 1998; Стъбс, 2002; Мелчук, 1998; Еверт, 2005; Еверт, 2007).

### 2.2.2. Класификация на несвободните фрази

За да очертаем мястото на съставните лексикални единици в системата на несвободните фрази бяха разгледани основните виждания в научната литература за несвободните фрази и техните основни особености, които определят позицията им в лексикалната система на езика. Следващата стъпка е да бъдат представени виждания за класификацията на несвободните фрази въз основа на техните семантично-синтактични особености, описвани често като семантични и синтактични ограничения или идиоматичност (вж. 2.1).

В Бояджиев и др. (1998, част 2, Лексикология) е представена класификация на устойчивите съчетания. Основните признаци, на базата на които се извършва класификацията, са степента на лексикална и структурна неделимост и граматическата им свързаност. На тази основа се различават следните категории:

1. Фразеологични словосъчетания
    - Фразеологични сраствания – лексикално и семантично неделими и неразложими с обобщено цялостно значение, което не се определя от съставките. Характеризират се с устойчив непроменен словоред и синтактична структура. Пример: *излизам от кожата си.*
    - Фразеологични единства – образувани чрез преносна употреба на техния общ смисъл. Могат да се съотнесат със свободно словосъчетания, за да се анализа метафоричното им значение. Пример: *чешат си езиците.*
    - Фразеологични съчетания – при тях една от думите може да се употреби само в съчетание с една или няколко определени думи. Семантиката на единицата е мотивирана от тази на съставните части, но се отделят с ясно разграничима фразеологична употреба, както и не се допускат синонимни

замени на думата с фразеологично свързано значение. Пример: *мъртво пиян* (думата *мъртво* е използвана с фразеологично свързано значение).

2. Фразеологични изрази (устойчиви фрази) – устойчиви по състав и употреба възпроизводими и семантично членими словосъчетания. Могат да функционират като отделни изречения. Пример: *всяка жаба да си знае гьола.*

3. Фразеологични съставни названия

   *Устойчивите съставни названия не са образни и изразителни, а максимално точни наименования на характерните за предметите и явленията особености, определени на понятийно-логическа основа. Характерна структурно-семантична особеност на фразеологичните термини е да уточняват понятието, изразено със съществително име, например атомна (слънчева, електрическа) енергия.*

Мелчук (1995, стр. 179) представя следната класификация на фраземите:

| *Phrasemes* | *Фраземи* |
|---|---|
| • *Pragmatic phrasemes* | • *Прагматични фраземи* |
| 1. *Pragmatemes* | 1. *Прагматеми* |
| • *Semantic phrasemes:* | • *Семантични фраземи* |
| 1. *Idioms* | 1. *Идиоми* |
| 2. *Collocations* | 2. *Колокации* |
| 3. *Quazi-idioms* | 3. *Квази-идиоми* |

Тази класификация се основава на признаците нерестриктивност и регулярност на означаваното и означаващото. В съпоставка със свободните фрази (вж. дефиницията в 2.2.1) се разграничават следните случаи:

1. Условия 1 и 2 са нарушени – означаваното не е нерестриктивно конструирано, дори да е регулярно, като същото важи и за означаващото. Тук попадат прагматичните лексеми.

2. Условие 1 е нарушено, но не и условие 2 – $'A \oplus B'$ е единственото възможно означавано, но то е нерестриктивно конструирано. Тук попадат също част от прагматичните лексеми.

3. Условие 1 не е нарушено, но условие 2 е – **AB** е нерестриктивно конструирано, но не е регулярно.

$$'X' =' A \oplus B' \qquad \Longrightarrow \qquad /X/ \neq /A \oplus B/.$$

Тук попадат семантичните фраземи.

От своя страна семантичните фраземи се разделят на три групи:

- Това са фраземи, при които идиомът **AB** има означавано $'C'$, което не съдържа нито $'A'$, нито $'B'$.

$$\mathbf{AB} = \langle 'C'; /A \oplus B/ \rangle \mid 'C' \neq' A' \,\&\, 'C' \neq' B'$$

  Тук се включват същинските фраземи, или идиомите. Пример: *изплювам камъчето.*

- Означаваното на едната съставна лексема се запазва, но другият компонент е проблемен.

$$\mathbf{AB} = \langle 'A \oplus C'; /A \oplus B/ \rangle \mid$$
$$'C' \text{ е изразено с } \mathbf{B}, \text{ така че } /A \oplus B/ \text{ е нерестриктивно конструиран.}$$

  Тук попадат полу-фраземите или колокациите.

- Фраземи, при които означаваното на **AB** включва означаваното и на двата конституента, но съдържа и допълнението $'C'$.

$$\mathbf{AB} = \langle 'A \oplus B \oplus C'; /A \oplus B/ \rangle \mid 'C' \neq' A' \,\&\, 'C' \neq' B'$$

  Авторът определя тази група като квази-фраземи.

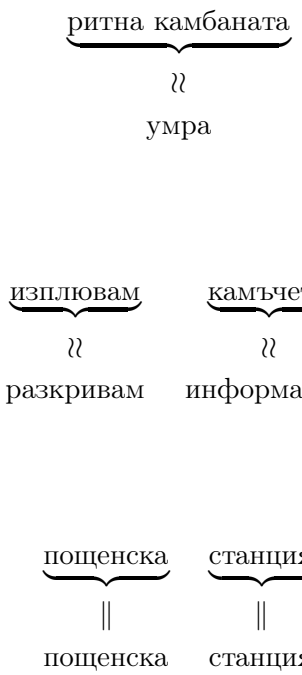
Относно колокациите Мелчук (1995) разглежда четири отделни случая:

1. или $'C' \neq' B'$, т.е. **B** не разполага с такова означавано (не присъства в речника); и
   (а) $'C'$ е празно, т.е. **B** е селектирано от **A**, за да участва в определена синтактична конструкция – колокации с опора, пример: *правя услуга*; или
   (б) $'C'$ не е празно, но лексемата **B** изразява $'C'$ само в комбинация с **A** или ограничен брой подобни лексеми – пример: *черно кафе*;

   или

2. $'C' =' B'$, т.е. **B** разполага с означавано (в речника) и
   (а) $'B'$ не може да се изрази с **A** и друг иначе възможен синоним на **B** – колокации с модификатори за интензитет, пример: *силно кафе* – **мощно кафе*; **страстен пушач* – **увлечен пушач* – ?*пристрастен пушач.* или
   (б) $'B'$ съдържа (важна част от) означаваното на $'A'$, което го прави по-специфично и поради това **B** е обвързано с **A** – пример: *конско цвилене.*

В статията на Саг, Балдуин, Бонд, Коупстейк и Фликинджър (2002) е представено по-общото разделяне на несвободните фрази на институционализирани и лексикализирани съчетания. Лексикализираните единици са тези, които изцяло или частично се характеризират с неделимост в семантично или синтактично отношение; те биват фиксирани, полуфиксирани и синтактично вариращи съчетания.




Фигура 2.1.: Композиционалност на значението при: (а) неразложими; (б) идиосинкретично разложими и (в) прости разложими несвободни фрази. (Болдуин и др., 2003)

Институционализираните единици, от друга страна, се описват като семантично и синтактично съставни, които обаче се появяват с относително голяма честота, поради което се определят като статистически идиоматични. Важно е също така да се очертае границата между институционализираните фрази и други свободни фрази, които се появяват със сравнително голяма честота, по предвидими извънезикови причини (например *продавам* и *къща*).

Болдуин, Банард, Танака и Уидоус (2003) разделят несвободните фрази на няколко основни типа: неразложими, идиосинкретично разложими и разложими. Фигура 2.1 и Пример 4 представят композиционалността и механизма на съотнасяне на значението при трите типа: ≀ означава непряко (преносно) съотнасяне, а ‖ – пряко (непреносно) съотнасяне. Авторите описват първата категория като лексикални единици, които не предлагат възможност за декомпозиционен анализ и в повечето случаи не допускат вътрешни модификации. Под вътрешни модификации се разбира например модификация или квантификация на конституентите. Втората категория представляват лексикални единици, които могат да се декомпозират, но (някои от) конституентите са реализирани със значение, което не е допустимо извън рамките на лексикалната единица. При тях се наблюдават синтактични вариации, макар и в силно ограничена степен. При третата категория се допуска декомпозиране на значението на това на конституентите. При тях има известна свобода на синтактичните вариации, както и

се допускат някои вътрешни модификации.

Границата между разложимите лексикални единици и свободните словосъчетания се изразява в това, че разложимите лексикални единици не допускат композиционни алтернативи, например синонимни (антонимни) замени на конституенти (Пиърс, 2001; Болдуин и др., 2003).

Пиърс (2001) разделя единиците в едно синонимно множество по отношение на дадена дума опора на три групи:

(1) единици, които се използват с тази опора;

(2) единици, които е допустимо да се използват, но не се срещат или са с много малка честота; и

(3) единици, използването на които води до неестествени изрази.

Например спрямо опора *въздух* синонимният ред *свеж, пресен, млад, опреснен, освежен, незастоял* се разделя на:

(1) *свеж*;

(2) *освежен*, *незастоял*; и

(3) *пресен*, *млад*, *опреснен*.

Употребата на определена конструкция е институционализирана и вариациите обикновено са силно ограничени.

Пример 4 има за цел да демонстрира различните ограничения, налагащи се върху отделните типове несвободни фрази, като дава възможност за сравнение между тях. Примерите включват следните явления:

- Недопустимост/допустимост на вътрешни модификации на частите: 1а), 1г); 2а), 2в), 2й); 3д); 4а), 4б), 4г).
- Ограничено словообразуване: 1а).
- Недопустимост/допустимост на синтактични вариации: 1б), 1д), 1е); 2б), 2г), 2д), 2и); 3г), 3е); 4в).
- Недопустимост/допустимост на замени със синонимни или описателни изрази: 2е); 3в); 4в).
- Недопустимост/допустимост на експлицитно приписване на признака към опората (при именни фрази): 2з); 3б), 3ж); 4е).

Трябва да се отбележи обаче, че допускането на модификации и замени (синонимни или антонимни) се определя и от синтактичната категория на фразата. При глаголни несвободни фрази с допълнение (например *изплюя камъчето*, *образувам дело*) промените в словореда са много често допустими при идиосинкретично разложимите и простите разложими, но несвободните именни фрази търпят силни ограничения не толкова поради особеностите на несвободна фраза, колкото поради синтактичните си

особености, които притежават и свободните фрази (Пример 4).

**Пример 4.**

1. Неразложими

   *ритна камбаната*; *от дъжд на вятър*

   а) **ритна голямата камбана*

   б) **Камбаната ли ритна Иван?*

   в) **Беше ли ритната камбаната.*

   г) **от много дъжд на вятър*

   д) **от дъжд и на вятър*

   е) **на вятър от дъжд*

2. Идиосинкретично разложими

   *изплюя камъчето*; *играя по свирката*; *дървен философ*; *работна ръка*; *дам под съд*

   а) **Изплюй едно камъче!*

   б) ?*Камъчето ще го изплюеш накрая!*

   в) **играя по голямата/дървената свирка*

   г) *По свирката на Иван ще ми играе!*

   д) ?*Философ дървен ли си?*

   е) **философ от дърво*

   ж) **трудова ръка*

   з) **Ръката е работна.*

   и) *Бих могъл под съд да ви дам за обида!*

   й) **дам под Софийския градски съд*

3. Прости разложими

   *пощенска станция*; *образувам дело*; *коефициент на полезно действие*

   а) *Градът е голям и това е пощенската му станция.*

   б) ?*Станцията е пощенска.*

   в) **формирам/правя/съставям/създавам дело*

   г) *Решено беше дело да не се образува.*

   д) *Ще бъде образувано дело за оперативна проверка.*

   е) **Трябва да се изчисли на полезно действие коефициентът.*

ж) **Коефициентът е на полезно действие.*

4. Свободни фрази

*висока станция*; *чиста кърпа*; *чист въздух*

а) *висока пощенска станция*

б) *високата 25 метра станция*

в) *?Станцията висока приближаваме.* (стилово ограничена употреба)

г) *чиста носна кърпа*

д) *Кърпата е чиста.*

е) *свеж въздух*

Саг и др. (2002) също отчитат факта, че отделните видове несвободни фрази се отличават с различно ниво на синтактична вариативност. Те описват три групи несвободни фрази, които допускат различни синтактични вариации: конструкции от глагол и частица, разложими идиоми (те се припокриват с идиосинкретично разложимите фрази в нашата класификация, вж. 2.1) и така наречените 'леки' глаголи (англ. **lignt verbs**). Докато в английски **фразовите глаголи** са често срещано и разнообразно явление, в българския език се срещат глаголи реципрока тантум и рефлексива тантум само с частиците *се* и *си*, които причисляваме към категорията на простите думи (в.ж. Класификация 1, стр. 33, 1.Б.; Коева, 2006b). Както се вижда от Пример 4, идиосинкретично разложимите несвободни фрази допускат някои синтактични вариации, които зависят и от синтактичния тип на фразата. Третата група на така наречените леки глаголи включва фрази като *полагам клетва*, *вземам решение* и др., при които глаголът има общо значение, което се конкретизира от съществителното, реализирано с нормалното си лексикално значение. Идиоматичността на тези конструкции се състои в ограниченията, които търпят – например ограничени възможности за синонимни замени: *давам клетва*, *полагам клетва*, но **правя/заявявам/излагам клетва*; *вземам решение*, но **сдобивам се/получавам/правя решение*, а също така могат да бъдат изразени с единична лексема, съответно *заклевам се* или *решавам*. От друга страна, значението им е разложимо в голяма степен, макар и понякога не напълно, поради което попадат в границите на съставните лексикални единици. В някои случаи може да става въпрос за свободни колокации, а границата между тях и съставните лексикални единици е неясна. В повечето случаи при тази група несвободни фрази се допуска и модификация на компонентите: *полагам клетва за вярност*, *вземам твърдо решение*.

Болдуин (2004) представя комплексен модел за описание на лексикалните единици, който се базира на следните типове маркираност:

- **Лексикална маркираност** – лексикално-граматическа фиксираност, ограничения в реализацията, ограничения във формообразуването (например *рит-*

*на камбаната*), специфичност на реализацията (например различно ударение: *Дóбър ден!* – прозодична маркираност);

- **Синтактична маркираност** – проява на синтактични аномалии (например *Добър вечер!*, липса на съгласуваност от съвременно гледище; институционализирана фраза, която запазва историческите си характеристики, въпреки граматичните промени в езика – промяната на рода на съществителното *вечер*);
- **Семантична маркираност** – некомпозираност на значението, участва в семантични релации с прости лексикални единици;
- **Прагматична маркираност** – прагматическите характеристики на частите не съвпадат с тези на цялата единица или изразът се асоциира с определена прагматична отправна точка;
- **Статистическа маркираност** – конвенционалност, изразява се във висока честота на поява на определени съчетания и нулева или маркирано ниска честота на други, които имат синонимно значение (например *строго секретен* срещу **стриктно секретен*).

| Лексикални единици | Маркираност | | | | |
|---|---|---|---|---|---|
| | Лекс. | Синт. | Сем. | Прагм. | Стат. |
| де юре | ✓ | ✓ | ? | ? | ✓ |
| Стара Загора | ✓ | ✗ | ✓ | ✓ | ✓ |
| на първо време | ✗ | ✓ | ✗ | ✗ | ✓ |
| първа помощ | ✗ | ✗ | ✓ | ? | ? |
| Добро утро! | ✗ | ✗ | ✗ | ✓ | ✓ |
| Дóбър вечер! | ✓ | ✓ | ✗ | ✓ | ✓ |
| сляпа баба | ✓ | ✓ | ✓ | ? | ? |

Таблица 2.4.: Типове маркираност на лексикалните единици, съставени от повече от една дума: лексикална, синтактична, семантична, прагматична и статистическа. ✓– маркиран; ✗– немаркиран; **?** – в известна степен.

Според Болдуин (2004) свободните съчетания не са маркирани, докато лексикалните единици са маркирани по един или няколко признака. В Таблица 2.4 са представени примери за изрази, маркирани по някои от признаците.

Болдуин (2004) обаче противопоставя лексикалните единици от този тип на ко-

локациите, които определя като *'случайни повтарящи се комбинации от думи'* за разлика от Стъбс (2002), който говори за релация в колокацията. Освен това, според Болдуин колокацията се характеризира с композирано значение и относителна синтактична свобода.

### 2.2.3. Значение на несвободните фрази за анализа на съставните лексикални единици (обобщение)

Направеният преглед на възгледите относно несвободните фрази беше необходим, за да се очертае мястото на съставните лексикални единици в системата на езика. В настоящата разработка приемам класификацията на несвободните фрази, представена от Болдуин и др. (2003), според която те се разделят на неразложими, идиосинкретично разложими и прости разложими (съставни лексикални единици). Класификацията обаче се допълва от още две категории, при които липсва семантична маркираност (Класификация 1).

Първата добавена категория е категорията на съставните лексикални единици със служебна функция (съставни съюзи и предлози), при които отсъстват или се наблюдават абстрактни, обобщени семантични отношения, поради което не е възможно да се определи наличието на семантична маркираност, но пък именно тези единици са силно маркирани по другите признаци. Служебните съставни лексикални единици са описани като отделна категория несвободни фрази и от Тодорова (2006).

Втората добавена категория в класификацията е тази на фразите, които са единствено статистически маркирани, без да са маркирани по другите показатели. Тази категория е обособена и от Мелчук (1998), който ги нарича 'колокации' (в тази разработка терминът има друго значение, вж. 2.1), както и от Саг и др. (2002), които ги определят като институционализирани фрази и ги отделят от лексикализираните фрази – каквито са останалите категории несвободни фрази.

**Класификация 1** (Лексикални единици в българския език)**.**

**А.** Прости лексикални единици (състоящи се от една дума):

- **А.1.** думи – състоящи се от една графична дума със значение, формирано от един пълнозначен компонент (*пиша*, *преписвач*);
- **А.2.** думи с частици – състоящи се от няколко графични думи, но една от тях пълнозначна, а другите служебни думи (*смея се*, *спомня си*);
- **А.3.** сложни думи – състоящи се от една графична дума, но чието значение е формирано от повече от един пълнозначен компонент (*ветропоказател*, *кораборемонтен*, *заместник-директор*).

**Б.** Несвободни фрази (състоящи се от повече от една дума):

**Б.1.** служебни съставни думи (*въпреки че*, *за да*, *благодарение на*);

**Б.2.** неразложими несвободни фрази (*от дъжд на вятър*);

**Б.3.** идиосинкретични несвободни фрази (*изплювам камъчето*);

**Б.4.** съставни лексикални единици (*компютърна система*).

**Б.5.** свободни колокации (*чист въздух*).

Обособяването на втората категория има важно значение за определянето на обхвата на понятието **съставни лексикални единици**. Според класификацията на Болдуин и др. (2003) простите разложими несвободни фрази включват и свободните колокации. Авторите слагат в една група примери като *kindle excitement* (бълг. *изгарящо/изпепеляващо вълнение*, на български език съчетанието не демонстрира същото явление) и *traffic light* (бълг. *светофар*). Това е допустимо, тъй като класификацията се основава на наличието на ограничения – недопускане или силно ограничаване на синонимни замени и вътрешни модификации.

При съставните лексикални единици обаче ограниченията са значително по-големи, отколкото при свободните колокации. Причината за това е, че съставните лексикални единици означават просто и неделимо понятие (например *минерална вода*), докато при свободните колокации става въпрос за модификация или описание на качество на понятието (например *пресен хляб*) и те допускат повече възможности за вътрешни модификации, синтактични вариации и др. (Пример 5).

**Пример 5** (Свободни колокации)**.**

*пресен хляб*; *глутница вълци*; *стадо овце*

| **Пример** | **Честота в БНК** |
|---|---|
| *пресен хляб* | 127 |
| *пресен бял хляб* | 5 |
| ?*нов хляб* | 7 |
| ?*свеж хляб* | 1 |
| *глутница вълци* | 148 |
| *глутница гладни вълци* | 11 |

| | |
|---|---|
| ?*група вълци* | 5 |
| ?*стадо вълци* | 5 |
| *стадо овце* | 214 |
| ?*група овце* | 1 |
| **глутница овце* | 0 |

Свободните колокации от лингвистична гледна точка са свободни съчетания, тъй като ограниченията, които търпят, се основават на извънлингвистични признаци – не е налице семантична, синтактична или прагматична маркираност. Ограниченията им се основават единствено на социолингвистични фактори – степента на конвенционалност на фразата, изразяваща се в честотата ѝ на употреба спрямо други възможни съчетания кандидати със същата опора и със същото значение (Пример 5).

## 2.3. Представяне на съставните лексикални единици

### 2.3.1. Композиционалността като основна характеристика на съставните лексикални единици

Принадлежността на даден израз към категорията на несвободните фрази се определя на базата на наличието или отсъствието на признака идиоматичност на фразата (вж. 2.1 и Нунберг и др., 1994). Идиоматичността има три измерения: конвенционалност, яснота и композиционалност, и определя доколко значението на фразата се осъзнава като формирано от значението на конституентите и доколко се повлиява и от външни, прагматични и социолингвистични фактори.

Композиционалността е една от основните характеристики на несвободните фрази. Тя определя континуум без категорична граница между различните категории несвободни фрази и свободните фрази, което усложнява задачата за класификацията на несвободните фрази.

Една от ключовите идеи на Нунберг и др. (1994) е, че не всички конституенти на несвободната фраза изгубват значението си и още повече, че дори тези, които го загубват, не го правят напълно. Ето защо авторите изтъкват, че при повечето несвободни фрази могат да се идентифицират отделните компоненти, синтактично, както и семантично, при което могат да се изследват семантичните връзки между компонентите и да се проследят механизмите, по които се формира значението на несвободната фраза и реализацията ѝ като определена синтактична конструкция.

Банард и др. (2003) дават следната практически ориентирана дефиниция на композиционалността:

*... for practical NLP purposes we are forced to adopt a rather straightforward definition of compo- sitionality as meaning that the overall semantics of the MWE can be composed from the simplex semantics of its parts, as described (explicitly or implicitly) in a finite lexicon.*

*... по практически причини сме принудени да приемем ясна дефиниция за композиционалността като значение, при което цялостното значение на несвободната фраза може да бъде композирано от простите значения на нейните части, както са описани (експлицитно или имплицитно) в речник с краен брой единици.*

Реди и др. (2011) стесняват горната дефиниция, като приемат, че дадена лексикална единица е композирана, ако значението ѝ може да бъде разбрано от простите (буквални) значения на нейните части. Този подход описва композиционалността като буквалност на значението. Реди и др. (2011) изследват степента на композиционалност – степен на буквалност на значението на цялата фраза, но и буквалност на значението на отделните компоненти в състава на фразата, в контекста на фразата и текста.

Използват се различни методи за установяване на композиционалността, като условно могат да се разделят на две групи:

- методи, разчитащи на лексикалната маркираност и синтактическите характеристики на несвободните фрази;
- методи, разглеждащи семантичните отношения и близостта между конституентите и цялата фраза.

При първите се използват статистически мерки за оценка на свързаността между компонентите на фразата, докато при вторите се измерва семантичното подобие между конституентите и несвободната фраза. Примери и за двата подхода могат да бъдат намерени в Глава 4.

Реди и др. (2011) разработват метод за емпирично установяване на степента на композираност на несвободни фрази от две думи с опора съществително име, при който хора оценяват доколко фразата като цяло и отделните конституенти са реализирани с буквално значение. В изследването се включват четири групи несвободни фрази:

1) при които и двете думи са използвани буквално;
2) при които първата дума е използвана буквално, но не и втората;
3) при които втората дума е използвана буквално, но не и първата;
4) при които и двете думи не са използвани с буквалното си значение.

По този начин авторите целят да изследват връзката между конституентите и композиционалността на фразата. Установява се, че това, доколко е буквално значенията на конституентите може да бъде намерено добро приближение за степента на композиционалност на цялата фраза, както и че и двата конституента допринасят за това – което посочва като слабост изложени в други разработки методи, използващи само определени конституенти и пренебрегващи други.

Болдуин (2006) дава представената по-долу дефиниция за понятието **композиционалност**.

**Дефиниция 1.**
Композиционалността е степента, до която характеристиките на частите на несвободни фрази комбинирано обясняват характеристиките на цялата фраза.

Болдуин (2006) описва различни видове композициалност – семантична, лексикална, синтактична, прагматична, и изтъква, че композиционалността покрива всички аспекти на маркираността (с изключение на статистическата маркиранст).

**Дефиниция 2.**
Разложимостта е степента, до която характеристиките на несвободната фраза могат да се обяснят чрез характеристиките на нейните части.

Докато композиционалността описва процеса на конструиране на значението чрез съчетаване на значенията на конституентите, разложимостта характеризира анализа по обратния път – от единното значение на цялото се прави опит за възстановяване на отделните значения на конституентите.

### 2.3.2. Дефиниция и обхват на понятието за съставна лексикална единица

Основните характеристики, отличаващи съставните лексикални единици от останалите категории несвободни фрази са следните:

- Съставните лексикални единици се възприемат като композирани и техните компоненти могат да бъдат идентифицирани (т.е. не са лексикално маркирани, вж. Таблица 2.4).
- Значението им се формира от значението на конституентите и е разбираемо. По тази характеристика се доближават до сложните думи (Класификация 1, стр. 33, категория А.3.) и често се разглеждат във връзка с тях (Болдуин, 2004; Джакендоф, 1997).
- Съставните лексикални единици се разграничават от свободните словосъчетания по следните признаци: Те се появяват със значителна честота в езика, като се противопоставят на синонимни/ антонимни съчетания, появяващи се със значително по-ниска (или дори нулева) честота, т.нар. антиколокации (Болдуин,

2004, Пиърс, 2001). В този смисъл те са институционализирани (конвенционализирани).

- Макар като цяло да се твърди, че проявяват относителна синтактична / структурна свобода, в много случаи те се характеризират със синтактични ограничения.

Ако се опитаме да опишем съставните лексикални единици чрез системата на Болдуин (2004) (Таблица 2.4), стигаме до обобщената характеристиката, представена в Таблица 2.6: те не са маркирани лексикално и семантично, често са синтактично маркирани, не е изключено да са прагматично маркирани и задължително са статистически разпознаваеми.

| Лексикални единици | Маркираност | | | | |
|---|---|---|---|---|---|
| | Лекс. | Синт. | Сем. | Прагм. | Стат. |
| Съставни лексикални единици | ✗ | Рядко | ✓(слабо) | ✓ | ✓ |

Таблица 2.6.: Типове маркираност при съставните лексикални единици.

Необходимо е обаче да въведем изискването да е налице повече от чиста статистическа маркираност, за да причислим единицата към категорията на съставните лексикални единици (вж. 2.3.1). Това изискване е важно, защото определя важната граница между лексикалните единици (означаващи просто, неделимо понятие) и свободните фрази. Свободните колокации са гранично явление между несвободните фрази (търпят ограничения) и свободните фрази (не търпят лингвистични ограничения, а само извънезикови, наложени от характера им на институционализирани фрази).

От направените дотук наблюдения можем да изведем следната дефиниция.

**Дефиниция 3** (Теоретична дефиниция)**.**
Съставна лексикална единица е всяка лексикална единица, която притежава следните характеристики:

(1) Състои се от две или повече думи;

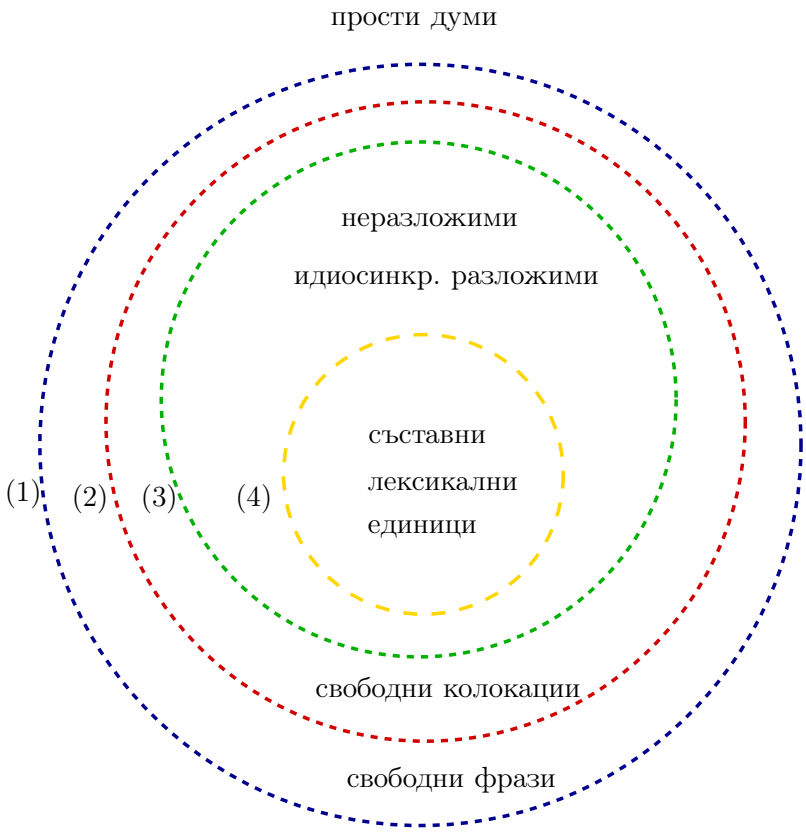
(2) Маркирана е поне по един признак: лексикално-граматически, семантично, синтактично, прагматично и/или статистически (но не само статистически);

(3) Означава цялостно и неделимо понятие;

(4) Фразата е семантично разложима – значението на цялата единица може да се представи чрез значенията на конституентите.

Лексикално-граматическата маркираност се изразява във фиксираност на формата, например прозодична маркираност в устната форма (*Дóбър вечер!*) или ограничени възможности за формообразуване (**Добри вечери!*). Синтактичната маркираност се изразява в силно ограничаване на синтактичните вариации, замените и вътрешните модификации на части от лексикалната единица. Прагматичната маркираност представлява обвързаността на лексикалната единица с определена комуникативна ситуация, участници и дискурс (например поздрави, термини и др.).

Предложената дефиниция има за цел да обособи категорията на съставните лексикални единици в системата от несвободни фрази, като очертае макар и условни граници с останалите категории, както и със свободните фрази (Фигура 2.2).




Фигура 2.2.: Приложение на условията от Дефиниция 3 за обособяване на категорията на съставните лексикални единици от другите езикови единици.

Посочената дефиниция има теоретичен характер и произлиза от наблюденията върху езиковите характеристики на несвободните фрази. Следвайки примера на Саг и др. (2002), Банард и др. (2003), Еверт (2007) и др., е направен опит за изграждане и на практическа (емпирична) характеристика на съставните лексикални единици, която произтича от поведението им в текстове и корпуси от текстове и е специално насочена към решаването на задачата за автоматичното разпознаване и тагиране на съставните лексикални единици.

**Дефиниция 4** (Практическа характеристика)**.**
Съставните лексикални единици в текста показват следните особености:

(1) Те са колокации – с ядро (опората на фразата или друга дума) и колокати (в определен спан от ядрото), и по това се различават от свободните фрази;

(2) Значението на всеки компонент на фразата е сходно или съвпада частично с това на цялата фраза, тъй като компонентите се реализират с лексикалното си значение, и по това се различават от неразложимите и идиосинкретично разложимите;

(3) Фразата най-често е регулярно и нерестриктивно конструирана (по терминологията на Мелчук (1995)), т. е. не се наблюдават синтактични особености, несъответстващи на правилата в българския език, синтактичната маркираност се изразява в ограничаването на възможностите за синтактични вариации;

(4) Семантичната маркираност се изразява в това, че фразата означава единно, неделимо понятие – по което се отличава от свободните колокации. Поради неделимостта на понятието не се допуска или е силно ограничено модифицирането на части от фразата;

(5) Фразата е институционализирана – което налага ограничения върху възможностите за синонимни замени и използването на описателни изрази.

Основната разлика между Дефиниция 3 и Дефиниция 4 е в това, че втората описва характеристики на съставните лексикални единици, които могат да бъдат наблюдавани в речта и за които могат да бъдат намерени методи за количествено измерване и качествена оценка. Повече детайли за това са изложени в 4.

Участието на значението на отделните конституенти във формирането на значението на съставната лексикална единица не е еднакво. Опората на фразата носи ядрото на значението, като същевременно предава и граматичните характеристики (на първо място принадлежност към част на речта). Най-общо казано, опората определя по-голям клас, към който понятието принадлежи, а подчинената част конкретизира значението.

Определянето на обхвата на понятието и изработването на дефиниция с практически характер има голямо значение за изграждането на методология за идентифициране на съставните лексикални единици и тяхното лингвистично описание. Глава 3 отделя повече внимание на класификацията на съставните лексикални единици и тяхното лингвистично описание.

### 2.3.3. Отношения на съставните лексикални единици с други езикови явления

#### 2.3.3.1. Съставните лексикални единици и сложните думи

Сложната дума представлява една графична дума, която обаче съдържа два или повече пълнозначни елемента, т. е. съдържа повече от един корен (например *светофар*, *хидроелектроцентрала*). В **Граматика на съвременния български книжовен език** (1983, том 2, Морфология, стр. 75–80) са описани видовете отношения между частите на сложната дума – съчинително и подчинително, като е обърнато особено внимание на типовете **семантико-синтактична зависимост**:

а) подложна – първият член съответства на подлог, например *водопад = [явление, при което] вода пада*

б) допълнителна – подчинената съставка съответства на пряко допълнение, например *кръводарител = [човек, който] дарява кръв*

в) атрибутивна – подчинената съставка съответства на определение, например *висококачествен = [който има] високо качество*

г) обстоятелствена – подчинената съставка съответства на обстоятелствено пояснение, например *бързоходец = [човек, който] ходи бързо*

Описанието на отношенията в сложната дума чрез синтактични трансформации е подходящо само за случаите, когато значението е формирано регулярно и нерестриктивно (по терминологията на Мелчук (1995)) – за сравнение, *водопад* и *вертиопашка* (*[някой, който] върти опашка*). Втората категория в **Граматика на съвременния български книжовен език** е определена като изключение и на нея е обърнато незначително внимание.

Традиционните правописни речници на българския книжовен език не включват съставните лексикални единици, а единствено сложните думи. В предговорите се обръща внимание на правилата за разделно, полуслято и слято изписване в някои случаи, при които има колебания в начина на изписване. Мурдаров (2008) анализира колебанията при изписване при *бизнесцентър – бизнес център*, *офистехника – офис техника* и обяснява причините за приемането им като правописни дублети. Тези примери демонстрират тънката граница между сложните думи и съставните лексикални единици и преминаването от едната в другата категория.

Изглежда продуктивно да се подходи към сложните думи, както към несвободните фрази, като се анализират въз основа на композиционалността в значението си. При подобен подход ще се различават неразложими, при които значението не може да бъде представено чрез значението на компонентите, частично или идиосинкретично разложими, при които само частичен и условен анализ е възможен, и разложими, при които значението на сложната дума може да се анализира като съчетание от

значенията на двата корена.

При подобен анализ се виждат сходните характеристики на съставните лексикални единици и разложимите сложни думи, които се проявяват в следните насоки, някои от които са описани и от Джакендоф (1997, стр. 164–167):

- Разложимите сложни думи са продуктивен клас – при създаване на нови термини (**Граматика на съвременния български книжовен език**, 1983), образуване на прилагателни от именни фрази (*руса коса* → *русокос*), поради което пълното им описание в речника е невъзможно;
- Характеризират се с недопустимост на варианти, например размяна на позицията на корените (**фаросвет*, **електрохидроцентрала*).
- Характеризират се с различна степен на композиционалност в значението и различни видове отношения между отделните компоненти:
  - В някои случаи значението на цялото е напълно конструирано от това на частите (*русокос*);
  - Понякога цялото е придобило по-конкретно значение от описаното от частите (*червеношийка* – не всяко същество / животно с червена шийка, а конкретен вид птица);
  - Много често липсва компонент от значението (зададен в квадратни скоби, [.]), който в някои случаи може частично да бъде възстановен от морфологични елементи, например наставки (*кръводарител*, наставка *-тел* за деятел), но в други случаи не може да бъде възстановена от компонентите;
- Характеризират се с институционализираност – имат установено значение и употреба, което е свързано с еднозначното им тълкуване дори в случаите, когато липсва компонент.

Въпреки приликите между сложните и съставните думи, сложните думи остават извън обсега на настоящата разработка, тъй като от практическа гледна точка се различават от съставните лексикални единици. Сложните думи, за разлика от съставните лексикални единици, се изразяват с една графична дума, поради което не представлява проблем да бъдат разпознати като една цялост в текста.

#### 2.3.3.2. Съставните лексикални единици и колокациите

Колокацията според приетото тук значение на термина в емпиричното му значение (вж. 2.1) представлява композиция от ядро (основна дума) и колокати, появяващи се в определен обсег (спан) от ядрото, които се срещат в текст или корпус от текстове с честота, която показва, че съвместната им поява не се основава на случайността.

Още през 50-те години на 20 век започва да се обръща внимание на контекста, в който думите се реализират. Фирт (1957) поставя началото на цяло течение в съв-

ременната лингвистика, което се занимава с контекстуалната теория за значението. Фирт (1957, стр. 192) разглежда значението по следния начин:

*To make statement about meaning in terms of linguistics, we may accept the language event as a whole and then deal with it at various levels, sometimes in a descending order, beginning with social context and proceeding through syntax and vocabulary to phonology and even phonetics, and at other times in the opposite order …*

*За да се опише значението от лингвистична гледна точка, може да се приеме езиковото явление като едно цяло и след това да се разглежда на различни нива, понякога от по-голямото към по-малкото, като се започне със социалния контекст, през синтаксиса и речника и се стигне до фонологията и даже до фонетиката, а друг път в обратния ред …*

По-нататък Фирт разглежда значението като разпределено между различни езикови равнища и въвежда термините **колокативно значение** (англ. **meaning by 'collocation'**) и **колокативна способност** (англ. **collocability**), чрез които се описва значението като абстракция на синтагматично ниво, без да се търси пряка връзка със значението от концептуална гледна точка (Фирт, 1957). Подходът на Фирт се основава на това, че дадена дума се реализира само с ограничен брой колокати, като във всеки отделен случай на думата се придава различно значение. В тази връзка е и известното изказване на Фирт (1957, стр. 11):

*You shall know a word by the company it keeps.*

*Ще разберете думата от компанията, в която се намира.*

Представеният по-долу Пример 6 демонстрира различни колокативни значения на съществителното име *човек*, в колокация с предходно прилагателно. Данните са от БНК © ИБЕ. Отделните колокати не се появяват с еднаква честота – най-популярно е съчетанието *млад човек* (5068 срещания), следвано от *стар човек* (1627 срещания), *непознат човек* (567 срещания) и *първобитен човек* (237 срещания), докато *правоверен човек* се среща само веднъж.

**Пример 6.**

1. *Страстите засипват живота и не позволяват на един* ***млад човек*** *да задели нещичко за черни дни!*
2. *Дори ми хрумна да помоля един* ***стар човек*** *с очила с телена рамка да ми заеме за малко тесничката си, с която ломеше камони по пътя, за да мога да пробия пътя си към Дора през гранитни скали.*
3. *Първите религиозни вярвания се формират с появата на* ***първобитния човек****, при това в резултат на неговия страх пред могъщите природни стихии.*
4. *Мисля си, че може да е негов колега или роднина, макар че е допустимо да е*

*съвсем* ***непознат човек****, който си доставя удоволствие, като вреди на другите.*

5. *Генерал дьо Лафланел беше* ***правоверен човек****, нещо твърде обичайно за генералите, хвърлили в огъня и убили много хора, които без сами да знаят това, са умрели по този начин съвсем по християнски благодарение на доблестните убеждения на дивизионния си командир.*

(БНК © ИБЕ)

Последователите на Фирт приемат възгледа, че значението на думите се определя от заобикалящата ги среда – контекста на социалната ситуация (за разлика от изолираното разглеждане на идеализирани участници в общуването), контекста на дискурса (писмения или устен текст за разлика от изолирани изречения) и непосредствения контекста в рамките на изречението, и задълбочават изследванията в тази насока (Халидей, 1966; Синклер, 1991 и др.)

Изследванията на контекста се стимулират след 60-те години на 20 век и от развитието на технологиите, с помощта на които става възможно събирането и обработването на сравнително големи корпуси от текстове.

(Барони и Еверт, 2007) и (Еверт, 2007) изследват колокациите – техните структурни характеристики и редица количествени мерки за асоциацията между компонените. Видовете съвместна поява са: повърхнинна (определен брой думи преди и след ядрото, ограничен от край на изречение), текстуална (честотата на поява в различни изречения или други текстови единици в зависимост от начина на дефиниране на текстова единица) и синтактична (поява на определени конструкции).

Отношението между колокациите и съставните лексикални единици в настоящата разработка се изразява по следния начин:

Емпирично ↔ теоретично;

Общо ↔ конкретно.

Колокациите са явление, което се наблюдава в корпуса, след което се пристъпва към техния лингвистичен анализ, докато съставните лексикални единици се разглеждат по обратния път – като лексикални единици, описани на теоретично ниво, които след това се идентифицират в текста. Още повече, като колокации се проявяват всички категории несвободни фрази, поради което колокациите съставни лексикални единици са подклас на всички колокации.

#### 2.3.3.3. Съставните лексикални единици и наименованията

Съставните лексикални единици се припокриват в определена степен с категорията на така наречените именувани същности или наименования (англ. **named entities**), които са широко разглеждани в съвременната лингвистика. Тази лексикална катего-

рия е обособена на базата на семантични и прагматични характеристики и включва еднозначно определени обекти (например *Иван Петров*, *Иван Петров*, *Стара планина*, *Организация на обединените нации*), названия на растителни и животински видове и др. Голяма част от наименованията са съставени от повече от една дума и по тази причина попадат и в настоящия анализ.

Тук приемаме класификацията на наименованията, представена в Лесева и Стоянова (2008), при която на базата на структурата им се обособяват същински, дескриптивни и каскадни наименования. Отделно се разглеждат дескрипторите – думи или фрази, които назовават общия клас, докато названията се отнасят към конкретен, еднозначно определен референт.

В настоящия анализ се включват категориите на дескрипторите и на така наречените дескриптивни наименования, които в преобладаващата си част са съставни лексикални единици – те назовават неделимо понятие, т. е. са лексикални единици, но значението им може да се представи като формирано от значенията на конституентите (например *Българска асоциация за електронна търговия*). Тези наименования са разгледани по-подробно в 3.1.4.

Тук представяме в обобщен вид основните категории дескриптори и дескриптивни наименования, представени в класификацията на BBN (Брунщайн, 2002), които влизат в обхвата на нашия анализ, т. е. съдържат примери за съставни лексикални единици, като същевременно обобщаваме и допълваме някои категории според особеностите на българския език. (Класификация 2).

**Класификация 2.**

**А.** Дескриптори на хора – титли, професии, занимания и др.

*доктор на математическите науки*, *автомобилен състезател*, *телевизионен говорител*

**Б.** Дескриптивни наименования на организации

**Б.1.** Религии и религиозни движения

*Църква на адвентистите от седмия ден*

**Б.2.** Партии и партийни организации

*Български земеделски народен съюз*

**Б.3.** Органи на управлението

*Министерство на здравеопазването*, *Варненски общински съвет*

**Б.4.** Образователни организации

*Университет за национално и световно стопанство*

**Б.5.** Икономически, търговски и фирмени организации

*Българска народна банка*

**Б.6.** Други

*Бургаско ловно-рибарско дружество*, *Българска федерация по тенис*

**В.** Дескриптори за организации

**В.1.** Религии и религиозни движения

*протестантска църква*

**В.2.** Партии и партийни организации

*комунистическа партия*

**В.3.** Органи на управлението

*общински съвет*, *областна комисия*

**В.4.** Образователни организации

*средно училище*, *висше училище*

**В.5.** Икономически, търговски и фирмени организации

*търговско дружество*

**В.6.** Други

*ловна дружинка*, *футболен клуб*

**Г.** Дескриптивни наименования на обекти, дело на човека

**Г.1.** Сгради – къщи, църкви, манастири, лаборатории и др.

*Национален природонаучен музей*

**Г.2.** Съоръжения – пътища, мостове, магистрали, летища

*Военновъздушна база към военноморските сили*

**Г.3.** Други

*Панагюрско златно съкровище*

**Д.** Дескриптори за обекти, дело на човека

**Д.1.** Сгради – къщи, църкви, манастири, лаборатории и др.

*къща музей*, *художествена галерия*, *химическа лаборатория*

**Д.2.** Съоръжения – пътища, мостове, магистрали, летища

*Т-образно кръстовище*, *военно летище*

**Д.3.** Други

*златно съкровище*

**Е.** Дескриптори за места

**Е.1.** Природни обекти

*планинска верига*, *алувиална равнина*, *солено езеро*

**Е.2.** Административни обекти

*областен център*

**Е.3.** Други

*междублоково пространство*

**Ж.** Дескриптори за продукти

*огнестрелно оръжие*, *дебитна карта*

**З.** Изрази за дати, време, числа, пари (Макар да попадат в разглежданата тук категория, те няма да бъдат включени в анализа, защото проблемите с тяхното разпознаване са по-специфични и се различават от тези на съставните лексикални единици.)

**И.** Названия за събития

*Втора световна война*

**Й.** Дескриптори за произведения на изкуството

*театрално представление*, *концертна пиеса за симфоничен оркестър*

**К.** Названия на закони и други административни актове

*Правилник за прилагане на закона за народното здраве*

**Л.** Терминология *блатен хвощ*, *гребенест крокодил*, *химично съединение*, *атмосферно налягане*

Същинските наименования са собствени имена, а каскадните се състоят от дескриптор и същинско наименование. При тях значението се формира с участието на референции към обекти от извънезиковата действителност (комуникативната ситуация). Макар да не приемаме, че дескрипторите са наименования, тяхната класификация може да бъде използвана като отправна точка в изработване на класификацията на съставните лексикални единици по семантични признаци в 3.2.1.

Съществуват и други класификации на наименованията като например представената от Тодорова (2010), според която те се делят на:

- Антропоними – поликомпонентни лични имена, прякори, псевдоними: *Ана-Мария*, *Поразяващата ръка*;
- Топоними: *Велико Търново*, *Луда Камчия*;
- Космоними: *Млечен път*, *Слънчева система*;
- Хрононими – имена на исторически епохи и събития: *Втора Световна война*, *Ден на славянската писменост и култура*;

- Хрематоними – собствени имена на предмети от материалната култура имена на картини, статуи; търговски марки: *"Кока кола"* и *"Видима идеал стандарт"*;
- Ергоними – собствени имена на учреждения, организации, съюзи и др.

Според класификацията на несвободните фрази, представена от Тодорова (2010), съставните имена са подкатегория на съставните фраземи, наред със съставните названия и съставните термини. В настоящата разработка се приема различна класификация на несвободните фрази (вж. 2.2.2) и по-нататък и на съставните лексикални единици (вж. 3.2), при която се разглежда идиоматичността на несвободните фрази в семантичен, синтактичен и прагматичен аспект.

Различните класификации на наименованията, и по-конкретно на съставните наименования, както и анализът на различни методи, използвани за тяхното автоматично разпознаване, могат да послужат за изграждането на методологията за класификация и идентифициране на съставните лексикални единици, които ще бъдат разгледани в 3 и 4.

Близостта между съставните лексикални единици и наименованията е разглеждана от практическа гледна точка, например общата методология за идентифициране, но не е правен опит за теоретичното изясняване на степента, до която двете групи се припокриват и начините, по които се съотнасят едни към други. Очевидно е, че след като могат да бъдат използвани сходни методи за тяхното идентифициране, те притежават сходни характеристики във формално, а вероятно и в семантично отношение. Повече детайли за отношенията между двете явления са представени в 3.1.4.

#### 2.3.3.4. Съставните лексикални единици и терминологията

Голяма част от терминологията се състои от несвободни фрази, поради което задачата за автоматично идентифициране и извличане на термини в определени аспекти е близка до тази за съставни лексикални единици (Франтци и Ананиаду, 1999; Да Силва и Лопез, 1999; Бонин и др., 2010 и др.). Терминологията също така се приема за част от именуваните същности (вж. 2.3.3.3).

Задачата за идентифициране на термини се характеризира със следните особености:

- Термините обхващат не само несвободни фрази, но и голям процент прости думи;
- Терминологията включва единици от различни типове несвободни фрази: неразложими (*божа кравичка*), идиосинкретично разложими (*морско конче*), съставни лексикални единици (*покритосеменни растения*);
- Появата на дадени термини много често се ограничава до определена, понякога тясно специализирана тематична област;

- Основните приложения на разпознаването на термини са за целите на автоматичното извличане на информация, класификацията на текстове по тематични области, изработване на тематични онтологии и др.

Задачата за откриване на съставни лексикални единици от друга страна, не е ограничена в рамките на научните текстове, тъй като съставните лексикални единици са широко разпространено общоезиково явление. Също така, от гледна точка на езиковите характеристики съставните лексикални единици са по-еднородна група, отколкото термините, които обхващат различни категории несвободни фрази.

За целите на извличането на информация например, често се предпочита използването на термини, които са несвободни фрази, тъй като при простите думи се наблюдава повече многозначност, което затруднява задачата. Някои използвани методи за откриване на несвободни фрази термини могат да бъдат приложени и ползвани за идентифициране на съставни лексикални единици.

Етапите на процеса на извличане на термини се припокриват с тези при разпознаване на съставни лексикални единици – необходима е предварителна обработка на корпуса (тагиране с част на речта и чънкиране, вж. 4) за извличане на кандидати, т. е. синтактични конструкции, например именни фрази, след което се извършва филтриране на кандидатите с помощта на статистически методи и лингвистичен анализ.

# 3. Лингвистично описание и класификация на съставните лексикални единици

Глава 2 беше посветена на несвободните фрази, тяхната класификация и мястото на съставните лексикални единици в лексикалната система на езика. Бяха дадени различни примери,демонстриращи разнообразието на несвободните фрази от гледна точка на тяхната структура и композиционалност. Примерите илюстрират и факта, че има несвободни фрази, както и съставни лексикални единици, които принадлежат към различни части на речта – съществително име (например *будка за вестници*), прилагателно име (например *еднозначно определен*, *жизнено необходим*), наречие (например *много малко*, *също така*), глагол (например *правя предложение*).

Най-значителен дял от съставните лексикални единици са именните фрази – , затова в тази и следващите глави ще бъде обърнато внимание специално на тази категория. Ограничаването на разглежданите видове съставни лексикални единици прави задачата за изграждане на теоретичен модел по-целенасочена и обозрима. Голяма част от наблюденията, класификацията и изводите обаче са приложими както за именните фрази, така и за цялата категория на съставните лексикални единици.

Настоящата глава има за цел да представи система за детайлно лингвистично описание на съставните лексикални единици, като се опита да ги разграничи от останалите категории несвободни фрази. Основната цел на предложената система за описание е практическа – да се изгради теоретичната рамка, която да позволи разработването на ефективни методи за автоматичното разпознаване и тагиране на съставните лексикални единици.

Втората част на главата представя няколко различни подхода към класификацията на съставните лексикални единици, които също имат подчертано приложно значение. В 3.2.3 като обобщение е представена практическа класификация, която ще служи като основа на разработените методи за автоматичен анализ, представени в следващите глави.

## 3.1. Лингвистично описание на съставните лексикални единици

### 3.1.1. Основни подходи към лингвистичното описание на съставните лексикални единици

Стъбс (2002, стр. 87-89) представя модел за формално описание на несвободните фрази (наричани от него 'разширени лексикални единици'), който използва елементи от анализа на Синклер (1996, 1998). Моделът се основава върху анализ на възможните конституенти на разширената лексикална единица и отношенията между тях. Моделът е представен в обобщен вид в Таблица 3.1.

| **Релация** | ***конституент*** |
|---|---|
| (1) Колокация | колокат: самостоятелна форма или дума |
| (2) Колигация | граматическа категория |
| (3) Семантично предпочитание | лексикално множество: клас от семантично близки форми или думи |
| (4) Дискурсна прозодия | описание на отношението на говорещия и неговата роля в общуването |
| (5) Сила на привличане | вероятността за поява на даден колокат, граматическа категория или лексикално множество |
| (6) Позиция и мобилност | насочена релация, относителната позиция на конституентите един спрямо друг |
| (7) Разпределение по текстови типове | дали се ограничава до определена тематична област или е общоезиково явление |

Таблица 3.1.: Конституенти на несвободни фрази и отношения между тях

Моделът дава приоритет на лексикалната информация и оставя на заден план синтактичните особености на лексикалната единица.

Болдуин (2004) представя комплексен модел за описание на лексикалните единици, който се базира на различни видове маркираност (вж. 2.2.2). Отделните групи несвободни фрази се характеризират с различна по тип и степен маркираност. Ако бъде приложена тази схема към съставните лексикални единици, може да се изгради следната система за описание:

- **Лексикална маркираност** – проявява се рядко и в степен, при която фор-

мата се губи частично и единицата остава разбираема (например *Дóбър ден!* – прозодична маркираност);

- **Синтактична маркираност** – не се проявява или рядко и в малка степен;
- **Семантична маркираност** – семантичната маркираност при съставните лексикални единици се изразява в ограниченията, които се налагат при конструирането на означаваното в някои случаи;
- **Прагматична маркираност** – съставните лексикални единици често са свързани с определена комуникативна ситуация, а в някои случаи значението се формира и чрез референции към извънезиковата действителност (например при наименованията);
- **Статистическа маркираност** – съставни лексикални единици имат приета конвенционална форма; при тях възможностите за синонимни замени, вътрешни модификации и др. са силно ограничени

По-долу да представени в повече детайли отделните аспекти от лингвистичното описание на съставните лексикални единици, като е направен опит за изграждане на концептуално единно описание с подчертано приложна насоченост.

### 3.1.2. Семантични особености на съставните лексикални единици

Подходът на Мелчук (1998) към описанието на несвободните фрази се опира на това, че те са лингвистични знаци (Дефиниция 5).

**Дефиниция 5.**
**Лингвистичният знак** е наредена тройка от вида:

$$\mathbf{X} = X;\ \ /X/;\ \ s_{\mathbf{S}},$$

където $X$ е означаваното на знака $\mathbf{X}$ (англ. **signified**), $/X/$ е означаващото (англ. **signifier**, т. е. формата – например фонетична) и $s_{\mathbf{S}}$ е неговата съчетаемост (множеството от данни за съвместната му поява с други знаци).

(Мелчук, 1998)

В лингвистичното описание на съставните лексикални единици като част от несвободните фрази ключова роля играят понятията за **нерестриктивност** и **регулярност** на формиране на означаваното и означаемото на базата на даден ConceptR (Мелчук, 1998; дефинициите са представени в 2.2.1). Достигаме до следната характеристика на съставната лексикална единица $\mathbf{X} = \mathbf{AB}$:

- $X$ е регулярно формирано от означаваното $A$ и $B$ на лексемите $\mathbf{A}$ и $\mathbf{B}$, но е рестриктивно конструирано, т. е. съществуват ограничения върху подбора на компонентите на означаваното. Например при *бяла мечка* няма избор за признака, чрез който да бъде конкретизиран този животински вид – той задължително

е цветът на козината (дори начинът, по който изглежда козината, тъй като тя е всъщност прозрачна), не мястото на живеене или друг характерен признак. Рестриктивността във формиране на означаваното се съчетава с институционализираност на значението, което прави единицата разбираема – *бяла мечка* означава не просто *мечка с бяла козина*, а конкретен животински вид.

- Означаващото $/X/$ също е регулярно формирано от $/A/$ и $/B/$, но е рестриктивно конструирано, т. е. съществуват ограничения върху подбора на компонентите на означаващото, което се изразява в невъзможността за парафразиране – *мечка с бяла козина* $\neq$ *бяла мечка*, тъй като цветът на козината не е единствената характеристика на този вид мечки; **светла мечка*, **белокосместа мечка*.

Проявата на рестриктивността е езиково зависима. В английски например *бялата мечка* се нарича *polar bear* (полярна мечка), а латинското название е *Ursus maritimus* (морска мечка) – този животински вид се характеризира по три различни признака в различни езици. Мелчук (1998) разглежда заедно случаите, при които са налице ограничения и върху означаваното, и върху означаващото, както в горния пример с *бяла мечка*, и случаите, при които има само ограничения върху означаваното, като означаващото е регулярно и нерестриктивно конструирано, като нарича двете групи прагматеми. Към втората група причислява прагматично маркирани фрази като поздрави, клишета, стандартни съобщения и други, например *Паркирането (е) забранено!* (пътен знак), които допускат варианти на изразяване, но без да се променя означаваното, *Не паркирай!* (променено прагматично значение – знак на гараж за разлика от официалния пътен знак), ?*Паркирането не е разрешено!*, **Забранено е оставянето на спряно превозно средство!*. Втората група има ограничено значение за настоящия анализ, поради което няма да бъде разглеждана в детайли.

В отделна група Мелчук (1998) обособява семантичните фраземи, при които означаваното е нерестриктивно конструирано, но не и регулярно. Нерегулярността може да се проявява по три начина:

- $\mathbf{AB} = C;\ /A\,B/\,|\,C \neq A\,\&\,C \neq B$ (неразложими и идиосинкретично разложими)
- $\mathbf{AB} = A\,C;\ /A\,B/\,|\ C$ се изразява чрез $B$ и $/A\,B/$ не е нерестриктивно (свободни колокации)

  Например при израза *силно кафе*, думата *кафе* е свободна и реализирана с лексикалното си значение, докато определението *силен* не е нерестриктивно избрано, а се налага над други синонимни еквиваленти – **мощно кафе*, **интензивно кафе*.

- $\mathbf{AB} = A\,B\,C;\ /A\,B/\,|\,C \neq A\,\&\,C \neq B$ (голяма част съставни лексикални единици)

  Мелчук (1998) определя последната група като квази-фраземи.

Последното формално описание по-горе за случаите, в които несвободната фраза

съдържа значенията на конституентите си, но съдържа и допълнително значение $C$, което не се предава нито с $A$, нито с $B$. По-точно би било изразяването във вида

$$\mathbf{AB} = A\,B\,C;\ /A\,B/\,|\,C \not\subseteq A \,\&\, C \not\subseteq B,$$

за да се подчертае, че означаваното на $C$ не се съдържа в $A$ и $B$ (дори имплицитно), освен че не съвпада с нито едно от тях. Например *свидетелство за сѣдимост* съдържа неизразен компонент *липса на [осѣдителни присѣди]*, докато ако се използва ?*свидетелство за неосѣждане*, липсата на присъда би се изразявала имплицитно от компонента *неосѣждане*.

По мое мнение гореописаната характеристика на съставните лексикални единици е една от основните им семантични характеристики, като обаче трябва да се отбележи, че нерегулярността в случая, както и преди това рестриктивността се компенсират от институционализираността на фразата, т. е. в много случаи това е достатъчно условие, за да се възстанови липсващият компонент. В случая можем да допуснем $C$ да бъде и $\varnothing$, при което това ограничение ще важи за всички случаи на съставни лексикални единици, тъй като не във всички случаи се проявява неизразен компонент на значението, например *асансьорна кабина*, *автомобилен състезател*.

В такъв случай може да се обобщи следната семантична характеристика на съставните лексикални единици:

- Означаваното $X$ е регулярно формирано от означаваното $A$ и $B$ на лексемите **A** и **B**, но не е нерестриктивно конструирано.
- Означаващото $/X/$ също е регулярно формирано от $/A/$ и $/B/$, но не е нерестриктивно конструирано .
- Нерегулярност на формиране на означаваното $\mathbf{AB} = A\,B\,C;\ /A\,B/\,|\,C \not\subseteq A \,\&\, C \not\subseteq B$, като се допуска и $C = \varnothing$ (т. е. се свежда до регулярност).

Една от представените класификации на съставните лексикални единици по нататък (вж. 3.2.1) се основава именно на характера и степента на възстановимост на липсващия компонент.

Леви (1978, стр. 1–5) разглежда съставните съществителни имена (англ. **complex nominals**)[1], като описва три начина за образуването им – чрез 'изтриване' на предикат, чрез номинализация на предикат и с непредикатни прилагателни определения. Изтритите предикати са възстановими и се описват като краен брой категории. Това е в съответствие и с въведения по-горе признак на съставните лексикални единици – наличието на неизразен компонент $C$. Номинализацията представлява трансформиране на предикат, заедно с прилежащите аргументи, в номинална структура,

[1] Разглеждат се именни фрази от две съществителни или прилагателно и съществително име, но наблюденията върху семантичната структура могат да се пренесат и към настоящото изследване, което включва и именни фрази, съдържащи предложна фраза. Още повече, често фрази с несъгласувано определение предложна фраза в български език се превеждат на английски с фраза от две съществителни, например 'синдром на Даун' – англ. 'Down syndrome'.

например *[човек, който] продуцира филми → филмов продуцент*. Непредикатните прилагателни представят вътрешно присъщи качества на обект. Характерно за тях е, че не могат да заемат предикатна позиция (с глагола *съм*) и често променят значението си при съчетаване с различни опори, например *слънчева батерия*; ?*батерия, която е слънчева*; *слънчев часовник*. Тук също има място по-горното условие за неизразения компонент, с който се показва различното значение в различни единици – в *слънчева батерия* значението е *източник на енергия*, а в *слънчев часовник* е *средство за изпълняване на функцията*. Категориите на Леви (1978) са представени в 3.2.1.

По подобен начин е характеризитана композиционалността на съставните лексикални единици от Круз (2000, стр. 67).

*The meaning of a grammatically complex form is a compositional function of the meanings of its grammatical constituents. This incorporates three separate claims:*

*(i) The meaning of a complex expression is completely determined by the meanings of its constituents.*

*(ii) The meaning of a complex expression is completely predictable by general rules from the meanings of its constituents.*

*(iii) Every grammatical constituent has a meaning which contributes to the meaning of the whole.*

*Значението на граматически комплексна форма е функция, която композира значенията на граматическите ѝ конституенти. Това включва три отделни твърдения:*

*(i) Значението на съставната единица е напълно определено от значението на конституентите.*

*(ii) Значението на съставната единица може да бъде изведено от значенията на конституентите с помощта на общоезикови правила.*

*(iii) Всеки граматически конституент има значение, с което допринася към значението на съставната единица.*

Значението на съставната единица може да се формира по два начина от значенията на конституентите – чрез събиране (англ. **additive mode**) или чрез съчетаване (англ. **interactive mode**). Най-общо казано събирането се отнася до съчинително свързване между конституентите. Макар и в някои случаи такова свързване да е допустимо в рамките на съставните лексикални единици (например *Съвет за радио и телевизия*), то не е обект на настоящия анализ. Съчетаването от друга страна се изразява в подчинително свързване и може да бъде ендоцентрично или екзоцентрично, при което се формират съответно ендоцентрични или екзоцентрични съставни лексикални единици (Круз, 2000; Болдуин и др., 2003). Ендоцентричните единици са хипоним на опората си или фразата, от която са образувани, например *число – просто число*, *химична реакция – анаеробна химична реакция*. При екзоцентричните се формира ново значение, различно от това на опората или изходната фраза, например

*конче – морско конче.*

Според теорията на Болдуин (2006) композиционалността на семантично ниво е основният признак на несвободните фрази, които се разделят на класове според степента на композициалност, най-общи са класовете на неразложимите фрази, на идиосинкретично разложимите и на простите разложими фрази, като съставните лексикални единици попадат в последната група. В нея обаче се причисляват и свободните колокации (вж. 2.1), които представляват многобройна група, но притежават повече характеристики на свободни фрази, отколкото на несвободни.

В рамките на категорията на съставните лексикални единици композиционалността може да варира. Пример 7 представя фрази с различна опора и подчинен конституент *морски*, които се характеризират с различна степен на композиционалност.

**Пример 7.**

- *морска разходка* (свободна фраза)

  = *разходка по море*, *разходка в морето*

- *морска риба* (граничен случай със свободните колокации)

  ?*риба от морето*; *рибата е от морето*; *рибата е морска*

- *черноморска риба* (граничен случай със свободните фрази)

  *риба от Черно море*; *рибата е от Черно море*; *рибата е черноморска*

- *морска костенурка* (съставна лексикална единица)

  означава семейство водни животни; значението е композирано от *костенурка* и *морски* (който живее в морето)

  **костенурката е от морето*; **костенурката е морска*

- *морско конче* (граничен случай с идиосинкретично композираните)

  означава род морски риби; значението е частично композирано от *конче* (животно, което има външна прилика с конче) и *морски* (който живее в морето). Подчертаните части означават приноса, който са дали конституентите за значението на цялата фраза. Както се вижда, неизразената част $C$ е значителна и нетривиална и е спорно доколко се подразбира и може да бъде възстановена.

  **кончето е от морето*; **кончето е морско*

- *морска звезда* (граничен случай с идиосинкретично композираните)

  означава клас безгръбначни морски животни; значението е частично композирано от *звезда* (обект, чиято форма прилича на звезда) и *морски* (който живее в морето), като неизразената част $C$ е голяма и е спорно, доколко се подразбира и може да бъде възстановена.

  **звездата е морска*

Съставните лексикални единици се характеризират с композиционалност на значението, което се изразява в това, че значението на цялата единица се формира като комбинация от значенията на конституентите. Трябва да се отбележи обаче, че след формирането на единицата, тя започва да се развива и е подвластна на отношенията в лексикалната система на езика, в които може да влезе независимо от конституентите, при което може да промени и първоначалния си статут от гледна точка на композиционалността си. Именната фраза *пощенска кутия* първоначално се формира като разложима лексикална единица, при което прилагателното *пощенски* (предназначен за поща) и *кутия* се реализират с лексикалното си значение.

**Речник на СБЕ** (1977–2008, т. 8, стр. 400) дава следната дефиниция за *кутия* (значение 1):

> *Неголям съд от картон, тенекия, дърво, пластмаса и под. с плоско дъно и обикновено с капак, в който се събират и съхраняват различни дребни предмети или продукти.*

Към това значение в края на речниковата статия са изброени съставни лексикални единици с опора *кутия*, измежду които е и *пощенска кутия*. В периода след издаването на **Речник на СБЕ** (т. 8 излиза през 1995) в българския език навлизат редица нови понятия, свързани с компютърните технологии и интернет комуникациите, при което и някои традиционни български лексикални единици придобиват нови значения, като например *мишка* за означаване на компютърното периферно устройство и *пощенска кутия* за мястото на сървър, където се съхраняват получени имейли (Пример 8).

**Пример 8** (Нова употреба на *пощенска кутия*)**.**
*Размерът на самата* ***пощенска кутия*** *пък е 3000 мегабайта, а възможността за прикачване на файлове в писмо позволява прикачването да е до 8 мегабайта.*

(БНК © ИБЕ)

Въпреки отдалечеността от първоначалното чисто композирано значение на съставната лексикална единица, *пощенска кутия* в значението си на понятие от интернет технологиите продължава да бъде възприемана като единица с композирано значение. Едно възможно обяснение на композираността в този случай е развитието на значенията на всеки от компонентите – *пощенски* (свързан с електронна поща) и *кутия* (място, контейнер, който има някакво съдържание и определена функция).

Друг възможен подход към описанието на значението и композиционалността на съставните лексикални единици е този, при който се използват множество от семантични примитивни – прости и универсални значения, които са интуитивни и отразяват универсалните езикови способности, т. е. могат да бъдат изразени на всички езици. Подобни методи се описват от Круз (2000), Ханкс (2000) и др. Ханкс (2000) например описва значението на думата извън контекста като множество от потенциали

на значения, като при всяка реализация думата се реализира с едно или комбинация от значения.

Ако се декомпозира значението на примитивни значения, можем да изследваме проявата на композиционалността, защото много често компонентите на съставната лексикална единица се свързват един с друг на базата на част от тези значения, а не на всички. В тази посока на анализ, горният пример с *пощенска кутия* може да получи различно обяснение.От дефиницията на *кутия* могат да се обособят няколко отделни примитивни значения, които заедно описват понятието – физическа форма, материал на изработване, размери, функция. Компонентът *кутия* се свързва с *поща, пощенски* най-вече чрез семантичния елемент *функция* – да служи за съхранение на поща. Тогава значението *функция* може да се окачествти като **свързано** значение на *кутия* в рамките на *пощенска кутия*. Свързаното значение е това значение, с което компонентът участва в образуването на фразата, а останалите значения остават свободни. Възможно е и съставната единица да се формира и на базата на повече от едно значение, както и всички да са свързани и да няма свободни.

Свързаното значение след това проявява стабилност в рамките на несвободната фраза, докато другите значения могат да търпят промени. По този начин *пощенска кутия* запазва основното си свързано значение, но останалите значения може да се загубват в определен контекст, като в случая с Пример 8.

Различен подход, особено актуален за целите на автоматичното обработване на съставните лексикални единици, е използваният от Наков (2008), при който се използва множество от парафразиращи глаголи, чрез които се описва приблизителното значение на съставната единица. Този подход отразява особеностите на съставните лексикални единици, при които в значението участва и неизразен елемент $C$. Именно този неизразен елемент може да бъде приближен с помощта на парафразиращи глаголи. В голяма част от случаите $C$ е именно предикат, така че подходът с парафразиращи глаголи изглежда основателен. Повече детайли върху приложението на този метод ще бъдат представени в глава 4.

Друг тип многозначност се получава, когато несвободната фразата има значение и като свободна фраза (Пример 9). Интересно е да се отбележи, че този тип многозначност е по-рядко срещан при съставните лексикални единици, отколкото при другите несвободни фрази, тъй като композираното значение е близко до значението на свободната фраза и не е възможен ярък контраст между двете значения – свободното и свързаното. Те се различават чрез наличието или липсата на институционализираност.

**Пример 9.**
*Купи на сина си голяма* ***кафява мечка****, а на дъщеря си розова.* (конструиран пример)

*Появата на* ***кафява мечка*** *в горите в Бавария, за първи път от 170 г. насам, пре-*

*дизвика смесени чувства сред местните жители, съобщи местната преса.* (БНК © ИБЕ)

Едно по-слабо изследвано явление е граматическата многозначност на синтактично ниво (Пример 10). Тя е подобна на многозначността, при която едното значение е като съставни лексикални единици, а другото като свободна фраза, но в този случай не става въпрос за многозначност на ниво фраза, а на ниво синтактична последователност от компоненти, тъй като в първото изречение на Пример 10 *съгласен* и *звук* не формират фраза, а случайно попадат в съседство.

**Пример 10.**
*Той е* ***съгласен звук*** *да не издаде.* (примерът е по Коева, 2006а)

*Богата е римата, при която съвпадат повече от един* ***съгласен звук*** *в непосредствена близост до ударения гласен.* (БНК © ИБЕ)

Многозначността на съставните лексикални единици представлява проблем за системите за автоматичен анализ и обработка на естествен език и системите за автоматичен превод. Повече на тези проблеми е отделено внимание в следващите глави.

Съставните лексикални единици влизат и в собствени семантични отношения с прости думи и други съставни лексикални единици – синонимия, омонимия, хипонимия, хиперонимия и други. Те заемат и важно място в лексикално-семантични мрежи като Wordnet (**Българският WordNet**; **WordNet: An Electronic Lexical Database**). Тези отношения остават извън обсега на настоящата разработка, макар да се отчита фактът, че подобни семантични отношения могат да се ползват за подпомагане на процеса на автоматично идентифицирае на съставните лексикални единици.

### 3.1.3. Синтактични особености на съставните лексикални единици

Както вече беше обяснено, тук ще бъдат разглеждани само съставните лексикални единици, които са именни фрази ($NP$). Те представляват разнообразна и нееднородна група от гледна точка на синтактичната структура. Налагаме ограничение за максимална дължина на фразата пет думи. Има три основни начина за образуване на съставна лексикална единица от базовата единица $NP \rightarrow N$ (проста дума), като всеки от начините може да се прилага рекурсивно, а редът на конституентите е фиксиран:

1. Чрез присъединяване на предпоставено определение прилагателно име

$$NP \rightarrow AP\ N', \quad AP \rightarrow A, \quad AP \rightarrow Adv\ A$$

(*бяла мечка*)

2. Чрез присъединяване на задпоставена предложна фраза

$$NP \rightarrow N'\ PP, \quad PP \rightarrow P\ NP$$

(*вдигане на тежести*)

3. Чрез приложение

$$NP \rightarrow N\ N', \qquad NP \rightarrow N'\ N$$

(*къща музей*, *бизнес център*)

При формирането на съставна лексикална единица често се налагат определени ограничения върху подчинената част. На първо място това са граматически ограничения – в случаите на съгласувано определение подчинената част се съгласува с опората. Освен това обаче в някои случаи се наблюдават допълнителни ограничения върху подчинената част, които могат по-нататък да се използват за успешното разпознаване на съставните лексикални единици. Подчинената част в случая на несъгласувано определение предложна фраза обикновено означава клас обекти, а не конкретен обект, например *паста за зъби* е съставна лексикална единица, докато *паста за зъбите на Иван* не е, а **паста за зъб*, **паста за зъбите* са неприемливи (с честота 0 в БНК срещу 417 срещания на *паста за зъби*).

Възможностите за модифициране на конституентите в рамките на съставната лексикална единица не са толкова силно ограничени, колкото при другите категории несвободни фрази. Макар и да означават едно понятие, в някои случаи се допуска модифицирането на конституентите, което води до конкретизиране на понятието и изменение на значението, което съответно води до образуването на нова съставна лексикална единица (например *паста за млечни зъби*, което се разглежда като отделна единица хипоним на *паста за зъби*), или води до конкретизиране на значението на конституента, при което съчетанието се окачествява като свободно (например *торта с орехи* → *торта с орехите от дървото в градината*).

Словоредните изменения в рамките на съставните лексикални единици са сравнително ограничени, още повече при именните фрази. Влияние върху възможностите за синтактични вариации играе и типът фраза – при предпоставените съгласувани определения възможностите за промяна на позициите на комплемента и опората са силно ограничени (обикновено са ограничени до поетичната реч, например *Мечка бяла днес видях*). Тяхната форма и значение са конвенционализирани, затова са и относително фиксирани. В някои случаи е възможно подчинената част да бъде изразена с прилагателно или с предложна фраза, като двете лексикални единици могат да се окачествят като варианти или синоними, тъй като имат едно и също значение, например *автомобилна гума – гума на автомобил*.

Силно ограничени са и възможностите за вмъкване на фраза между конституентите на съставните лексикални единици. Изключение прави кратката форма на притежателното местоимение, като и тя не се допуска в някои съставни лексикални единици, например *пощенската* ***ми*** *кутия*, *пастата* ***му*** *за зъби*, но **Организацията* ***ни*** *на обединените нации*, а единици като *офис оборудване* категорично не допускат вмъквания между конституентите.

Заслужават внимание и две други явления при съставните лексикални единици, които затрудняват автоматичното им разпознаване. Едното явление е елипсата на опората на фразата в случаи, когато тя се подразбира, например *съгласна* _ или *венечна съгласна* _. Тези конструкции остават неразпознати в практическия анализ, тъй като не отговарят на описаните допустими конструкции за именна фраза. Другото явление, което остава изолирано, е съчинителното свързване на две съставни лексикални единици, като съчинителното свързване се изразява на ниво подчинени части, например *алергия към краве мляко* ***и*** *глутен*, *главни* ***и*** *второстепенни части на изречението*, или на нивото на опората, например *лице* ***и*** *число на глагола*.

### 3.1.4. Прагматични особености на съставните лексикални единици

Прагматичната идиоматичност на съставните лексикални единици се проявява в тяхната обвързаност с определена комуникативна ситуация – дискурс на общуване, отношения на изказването с контекста, участници в ситуацията и техните роли в нея и т. н. Болдуин (2004) описва като прагматично маркирани несвободни фрази, които са обвързани със ситуацията – като поздрави (например *Добър вечер!*) и др.

Тук обхватът на прагматичната идиоматичност се разширява в посока към включването на референтността. По този начин се установяват и отношенията, които съществуват между наименованията (англ. **named entities**) и съставните лексикални единици и на които не е обръщано задълбочено внимание досега с изключение на практическия анализ и използването на общи методи за тяхното автоматично идентифициране в текст.

Донякъде проблематични са и дефиницията и обхватът на понятието 'наименование'. Крипке (1980) дефинира наименованията на базата на тяхната способност да означават референта точно, еднозначно и непротиворечиво. В групата на точните означаващи се включват личните имена, както и термини от естествените науки – например биологически видове, химически вещества и др. Изразите за време и повечето числени изрази, например валутни означения, мерки за дължина, тегло и др. също традиционно се причисляват към наименованията. Крипке (1980) описва наименованията като означаващи един и същи обект във всеки възможен свят (едозначни означаващи). Случайните означаващи от друга страна не притежават тази способност. Добавянето на описания или дескриптори към имената се използва за фикси-

ране на референцията, при което дескрипторите, които могат да бъдат изразени и от фраза, стават неделима част от наименованието.

Тук се приема класификацията на Лесева и Стоянова (2008) на наименованията, при която те се разделят на:

- същински наименования (англ. **true NEs**), например *Иван Петров*;
- описателни наименования (англ. **descriptive NEs**), например *Световна тенис федерация*;
- каскадни наименования (англ. **cascading NEs**), например *Софийски университет "Св. Климент Охридски"*.

Същинските наименования са собствени имена, а каскадните се състоят от дескриптор и същинско наименование. При тях значението се формира с участието на референции към обекти от извънезиковата действителност (комуникативната ситуация). Описателните наименования представляват съставни лексикални единици, при които значението е композирано от това на конституентите, но често са налице неизменяемост в синтактичната структура и недопустимост на алтернативни форми и синонимни замени (например *Национален съвет за радио и телевизия*, но **Национален съвет за телевизия и радио*; *Българска асоциация на туристическите агенции*, но **Българска асоциация на туроператорите*) и имат ограничена парадигма (например ?*Световните федерации по тенис*).

Отделно се разглежда групата на дескрипторите, които в преобладаващата си част са съставни лексикални единици – те назовават неделимо понятие, т. е. са лексикални единици, но значението им може да се представи като формирано от значенията на конституентите (например *информационна агенция*).

Прагматичните особености на съставните лексикални единици могат да се разглеждат във връзка със степента, до която значението на съставната лексикална единица се формира с участието на референции към обекти от извънезиковата действителност. Същинските и каскадните наименования са силно прагматично маркирани, тъй като ще означават конкретен референт, чрез което стават пряко обвързани с дадената комуникативна ситуация. Описателните наименования ще бъдат също силно маркирани, тъй като посочват еднозначно определен референт, като същевременно дават и неговото описание.

Различните родови и видови имена, например биологични видове, химически вещества и др. ще имат по-слаба обвързаност с контекста и се характеризират с по-слаба прагматична маркираност, която се основава на тяхната относителна обвързаност с определена тематична област или дял на науката. Също в тази група могат да се причислят и съставни лексикални единици, които имат подчертано битов характер, например *снежен човек*, *столче за кола*, *паста за зъби* и др., или подчертано административен характер, например *работна среща*, *човешки ресурси* и др. Тук влизат и поздравите (*Добро утро!*), както и някои означения и стандартни надписи (*Пушене-*

*то забранено!*) и др. Обхватът и разнообразието от категории на групата на "средно маркираните" съставни лексикални единици зависи от възприетото разбиране за комуникативна ситуация, нейните измерения и характеристики, както и аспектите на отношенията между език и реалност.

Като прагматически немаркирани или слабо маркирани могат да се окачествят съставни лексикални единици от общата лексика, за които не може да се намери връзка между употребата им и контекста, например *пощенска кутия*, *слънчева светлина* и др.

### 3.1.5. Формообразувателни особености на съставните лексикални единици

Тъй като опората на фразата е съществително име, съставната лексикална единица приема лексикално-граматическите характеристики на съществителното име. Лексикалните единици се характеризират с род и образуват форми за число, бройни форми (при мъжки род, неодушевени), членувани форми.

При редица лексикални единици съществуват ограничения в парадигмата. От една страна те може да се налагат от опората, напр. *братска обич* няма форма за множествено число, защото *обич* няма такава форма. При други ограниченията са обусловени от прагматично гледище, при което *Бългaрска социалистическа партия* е единствена същност и поради това формата в множествено число ?*Бългaрски социалистически партии* е нетипична. Има и случаи, при които макар и да се образува пълна парадигма, определени форми се възприемат като нетипични, тъй като не се срещат или се срещат с много малка честота в речта (Пример 11).

**Пример 11** (Редки форми)**.**

| **Пример** | **Честота в БНК** |
|---|---|
| *информационна услуга* | 10 |
| *информационни услуги* | 339 |
| *информационна технология* | 67 |
| *информационни технологии* | 2647 |
| ?*мобилна комуникация* | 0 |
| *мобилни комуникации* | 288 |

Коева (2006а) представя формален модел за описанието на морфологичните особе-

ности на несвободните фрази в българския език. Техните морфологични и синтактични особености зависят от конституентите им. Описанието на всяка единица в речника включва следните елементи[2]:

- граматическите характеристики на опората – съществително собствено или нарицателно, род, одушевено или неодушевено;
- брой и характеристика на конституентите – означенията от вида $aN+Npn+m$, в което главните букви са характеристиките на опората и са използвани стандартните означения за частите на речта: $A$ – прилагателно име, $N$ – съществително име, $V$ – глагол, $P$ – предлог и т. н., и граматическите характеристики: $M$ – мъжки род, $N$ – среден род и т. н. Знакът + въвежда гламатически определения към последната дума, например $pn+m$ съответства на конструкция от предлог ($p$), следван от съществително ($n$) в мъжки род ($+m$).
- информация за характера на всеки интервал между частите на несвободната фраза – различават се интервал тип 1, който е винаги празен, пример *висши бозайници*, **висшите ни бозайници*; тип 2, който допуска само клитики, пример *пощенска кутия*, *пощенската си кутия*; и тип 3, където може да се появи неопределен компонент, но този тип засяга само глаголните несвободни фрази, например *правя снимка*:

  **Пример.**
  *Според Балева и Брунбауер през 1889 г. полякът преоткрива кървавата история на Батак и* ***прави*** *[на място]* ***снимки*** *на инсцинировки на клането, както и на мнимите злодеи, служещи за студии за монументалната му творба.* (БНК © ИБЕ)

- флективен тип – флективният тип трябва да определи напълно всички словоформи на съставната лексикална единица.
- словоред – различават се фиксиран словоред, например *товарен влак*, и относително свободен словоред, който се наблюдава при част от несвободните фрази глаголи, например *правя снимка*.

  **Пример.**
  *Снимки ли правите още?* (Конструиран пример)

- съгласуване между компонентите – зависимите компоненти може да се променят независимо от опората (неизменяеми думи, ограничена парадигма, пълна парадигма) или да се съгласуват с опората (ограничена или пълна парадигма). В Пример 12 са илюстрирани основните възможности за отношения между опората и конституентите.

[2] Тук разглеждаме само съставни лексикални единици, принадлежащи към категорията на съществителното име. Някои оригинални примери, които не са именни фрази, са заменени, както са добавени и някои нови илюстративни примери.

**Пример 12.**

- *горски пожар* – *горски* се съгласува с опората *пожар*
- *къща музей* – не се съгласуват по род (приложение), но се съгласуват по число, *къщите музеи*
- *паста за зъби* – *за зъби* не се съгласува (предложна фраза) и не се изменя
- *правя снимка* – значително разнообразие по форми, пълна парадигма на *снимка*; може да се модифицира, квантифицира и др.

Глаголите съставни лексикални единици поставят редица въпроси за изследване, свързани с границите между свободните и несвободните фрази, проблемите на лингвистичното описание и анализ на несвободни фрази, тяхната класификация, както и пред разработването на методи за тяхното автоматично разпознаване и обработване. Тъй като са много широк и разнообразен клас, те заслужават отделно внимание и остават извън рамките на настоящата разработка.

### 3.1.6. Необходимо ли е включването на съставните лексикални единици в речника?

Правописните речници на българския книжовен език не включват съставните лексикални единици, но съдържат информация за наличието на правописни дублети като *бизнесцентър – бизнес център*. **Български тълковен речник** (1994) и **Речник на СБЕ** (1977–2008) обаче включват всички видове несвободни фрази – както неразложими и идиосинкретично разложими (например *девета дупка на кавала сам* в статията за *кавал*, **Речник на СБЕ**, 1977–2008, т. 7, стр. 17), така и съставни лексикални единици и свободни колокации (например *лека кавалерия* и *тежка кавалерия* в статията за *кавалерия*, **Речник на СБЕ**, 1977–2008, т. 7, стр. 19).

Съставните лексикални единици са представени по два начина в **Речник на СБЕ** (1977–2008). В някои случаи те са представени в отделен раздел на речниковата статия на опората, отбелязан със символа ◊, например *високочестотен кабел* (**Речник на СБЕ**, 1977–2008, т. 7, стр. 11). Те са съпроводени с маркер за тематичната област, в която се срещат (*високочестотен кабел*, *Техн.*), и речниково значение. В други случаи съставните лексикални единици са изредени в края на речниковата статия, без да се съпровождат с допълнителна информация, например *асансьорна кабина*, *командна кабина*, *параходна кабина*, *пасажерска кабина*, *прожекционна кабина* (**Речник на СБЕ**, 1977–2008, т. 7, стр. 13). Различният подход изглежда обоснован от степента на композиционалност и декомпозиционалност – доколко значението на несвободната фраза може да се изведе от значението на конституентите.

Шон и Джурафски (2001) изследват начините за автоматично генериране на лексикални ресурси, в които са представени несвободни фрази, като същевременно е спазено условието за минимална повторяемост на информацията. Освен това тряб-

ва да се отчитат и възможните приложения на подобни лексикални ресурси, което ще предедели количеството и качеството на включената в описанието им информация. Същевременно не е препоръчително обвързването на разработените ресурси с конкретна дейност или задача, за да се осигури по-широката им приложимост.

Саг и др. (2002) изтъкват необходимостта от генерализиране на еднотипни конструкции и явления, тъй като често те могат да бъдат обособени в групи (англ. **families**) със сходни характеристики и поведение. Авторите наблягат на необходимостта от изучаване на отделните типове несвободни фрази, за да се изгради качествен модел за тяхното описание.

Грос (1986) очертава различен подход към речниковото описание на несвободните фрази, при който те се характеризират като последователност от синтактични категории. По този начин те се разделят на класове и се осигурява по-бърз достъп до речниковата информация.

С практическа значимост на настоящото изследване е моделът, който Болдуин (2006) очертава за представяне и анализиране на композиционалността. Обърнато е внимание на възможностите за представяне и моделиране на композиционалността на лингвистично ниво, както и възможностите за нейния емпиричен анализ. Болдуин (2006) представя следните подходи към нейното изследване:

- Речников подход – бинарна оценка (да – не), основана на предположението, че неразложимите фрази са зададени в речника;
- Онтологичен подход – относителното подобие на частите и цялата фраза;
- Подход, анализиращ следствените връзки – бинарна оценка на това, дали цялото "влече" частите;
- Степенен подход – описание на композиционалността по непрекъсната или дискретна скала;
- Класов подход – интерпретация на съставните лексикални единици по отношение на принадлежността им към краен брой семантични класове, формирани на базата на нивото на разложимост.

Болдуин (2006) представя и практическа оценка на възможностите на отделните методи да представят композиционалността, като прави впечатление, че от една страна речниковият метод, а от друга страна степенният и класовият подход са най-успешни.

Очертават се три възможности за мястото на съставните лексикални единици в речника:

- Да не бъдат включвани;
- Избирателно включване – според степента на идиоматичност;
- Да бъдат вкючвани.

При съставните лексикални единици се оформят няколко съществени довода против включването им в електронните речници, използвани за обработка на естествения език.

1. Съставните лексикални единици не са краен брой, а са продуктивен клас; допълването с нови единици не изисква дълъг период на употреба (като например при *всяка жаба да си знае гъола*), а институционализирането в много случаи става бързо (например въведената в тази разработка нова терминология).
2. Допълнителното включване на съставните лексикални единици в речника е свързано с дублиране на информация, тъй като те се характеризират с композиционалност на значението и формата.
3. Ще се увеличи значително обемът на лексикалните ресурси, разработването им е скъпа дейност от гледна точка на човешки труд и ресурси за съхранение, организация и обработка.
4. Ако съставните лексикални единици се определят и анализират на ниво лексикален анализ заедно с другите лексикални единици, това ще увеличи значително времето за обработка на този етап.
5. Избирателното включване изисква ясно поставяне на граница между тези съставни лексикални единици, които да бъдат и които да не бъдат описани в речника. Към момента не съществува точен количествен или качествен критерий, чрез който да се класифицират съставните лексикални единици.

Ето защо е препоръчително да се проучат възможностите за разработване на други методи за анализ на съставните лексикални единици, като се обърне специално внимание на спетенния и класовия подход, описани от Болдуин (2006). В следващата глава 4 се търсят начини за установяване на подходящи качествени и количествени критерии, които да позволяват автоматичното идентифициране на съставните лексикални единици.

## 3.2. Класификация на съставните лексикални единици

Могат да бъдат очертани няколко различни подхода към класификацията на съставните лексикални единици по различни признаци – семантична класификация (според семантичните отношения между опората на фразата и останалите конституенти), синтактична класификация (според синтактичната структура на фразата) и идиоматична класификация (според типа и степента на идиоматичност на фразата). Литературата вурху класификацията на съставните лексикални единици е ограничена в сравнение с тази за класификацията на несвободните фрази, но голяма част от анализите и подходите за класигикация са валидни както за несвободните фрази, така и за съставни лексикални единици.

С оглед на автоматичното разпознаване и тагирането на съставните лексикални единици синтактичната и семантичната класификация имат широко приложение. След тяхното представяне ще бъде направен опит за изграждане на подчертано практическа класификация, която е залегнала в основата на разработената методология за идентифициране и тагиране на съставните лексикални единици в българския език, която е представена в глава 4.

### 3.2.1. Класификация по семантични признаци

Към класификацията на съставните лексикални единици може да се подходи по различни начини – според семантичните отношения между опората и конституентите, както и според характера и проявата на рестриктивност и нерегулярност при формиране на означаваното и означаващото.

#### 3.2.1.1. Преглед на някои подходи към семантичната класификация на съставните лексикални единици

Представената Класификация 2 на наименованията, стр. 44, обхваща голяма част от съставните лексикални единици и може да бъде ползвана като отправна точка към тяхното описание и класификация.

Друг възможен подход към семантичната класификация е според семантичните релации между съставната лексикална единица и по-общия клас, към който понятието принадлежи, изразено в значението на опората на фразата. Този подход важи само за ендоцентричните съставни лексикални единици (вж. 3.1.2), тъй като при екзоцентричните имаме промяна в значението на опората.

Като отправна точка тук можем да използваме представените от Гиржу и др. (2007) семантични релации, които са представени и в WordNet (**WordNet: An Electronic Lexical Database**, **Българският WordNet**):

- Причина – резултат;
- Инструмент – агент;
- Продукт – производител;
- Произход – същност;
- Тема – инструмент;
- Част – цяло;
- Съдържание – съдържаща същност.

Гиржу и др. (2007) изтъкват, че не е постигнат консенсус за множеството от релации, които да се използват, както и за алгоритмите за тяхното изследване. Поради това определят като малко вероятно дадена система за описание да има универсален

характер. Системата, която те създават, покрива примерите, с които разполагат.

Леви (1978) описва образуването на съставни лексикални единици именни фрази по три начина – чрез изтриване на предикат, чрез номинализация на предикат и с непредикатни прилагателни определения (вж. 3.1.2). Представени са девет обобщени предиката, които могат да се изтриват и по този начин да се получават съставни лексикални единици. Предикатите са възстановими, т. е. се съдържат в семантичната структура на съставната единица. Тези предикати са наречени **изтрити с възможност за възстановяване** (англ. **Recoverably Deletable Predicate**s, RDP) и са представени и илюстрирани с примери в Таблица 3.2.

| **Предикат, RDP** | **Примери** |
|---|---|
| CAUSE<br>ПРИЧИНЯВАМ | *малариен комар*<br>*вирусна инфекция*<br>*родилни болки* |
| HAVE<br>ИМАМ | *шоколадова торта*<br>*болнично легло*<br>*бисерна мида* |
| MAKE<br>ПРАВЯ | *копринена буба*<br>*птиче гнездо* |
| USE<br>ИЗПОЛЗВАМ | *парен локомотив* |
| BE<br>СЪМ | *пчела работничка*<br>*къща музей* |
| IN<br>СЪМ_В | *водна лилия*<br>*морска костенурка* |
| FOR<br>СЪМ_ПРЕДНАЗНАЧЕН_ЗА | *почивна станция*<br>*паста за зъби*<br>*плувен басейн* |
| FROM<br>СЪМ_ОТ | *слънчева светлина*<br>*морска сол* |
| ABOUT<br>ОТНАСЯМ_СЕ_ЗА | *закон за авторското право*<br>*коефициент за полезно действие* |

Таблица 3.2.: Изтрити предикати с възможност за възстановяване (англ. **Recoverably Deletable Predicate**s, RDP), (Леви, 1978, стр. 75–118).

В някои случаи подчинената част на съставната единица е в позиция на подлог, а понякога на допълнение в изречението с предиката преди изтриването му (Пример 13).

**Пример 13.**

- *Тортата има шоколад.* → *шоколадова торта*
- *Болницата има легла.* → *болнично легло*
- *Бубата прави коприна.* → *копринена буба*
- *Птицата прави гнездо.* → *птиче гнездо*

Относно номинализациите и по-специално синтактичната позиция на подчинената част в оригиналната конструкция Леви (1978, стр. 167–184) обособява типовете субективни (англ. **subjective nominalizations**), обективни (англ. **objective nominalizations**) и многосъставни (англ. **multi-modifier nominalizations**), всеки от които допълнително разделя на четири подтипа според синтактичната роля на опората в оригиналната конструкция преди номинализацията. Ролите на *пациент*, *продукт* и *агент* трябва да се възприемат в тяхното обобщено значение. Категориите са представени с примери в Класификация 3.

**Класификация 3.**
(Номинализации спрямо синтактичната позиция на подчинената част (1, 2, 3) и на опората (а, б, в, г) в оригиналната конструкция.)

1. Субективни номинализации
   (а) Опората е действие: *съдебно решение*;
   (б) Опората е продукт: *съдебна грешка*;
   (в) Опората е агент: – (подчинената част е агент);
   (г) Опората е пациент: *народен избранник*.
2. Обективни номинализации
   (а) Опората е действие: *сърдечен масаж*;
   (б) Опората е продукт: *информационна услуга*;
   (в) Опората е агент: *научен работник*;
   (г) Опората е пациент: – (подчинената част е пациент).
3. Многосъставни номинализации

(а) Опората е действие: ?*радиоактивно замърсяване на почвата* (граничи със свободно съчетание; 2 срещания в БНК), ?*въоръжена борба за освобождение* (граничи със свободно съчетание; 6 срещания в БНК);

(б) Опората е продукт: *общинска наредба за чистота(та)* (2 срещания в БНК);

(в) Опората е агент: *активен борец против фашизма* (47 срещания в БНК, 33 от които са част от *активен борец против фашизма и капитализма*);

(г) Опората е пациент: –

Някои единици могат да бъдат описани с повече от един предикат, например *родилни болки – раждането ПРИЧИНЯВА болки* или *болките СЕ ОТНАСЯТ до (са свързани с) раждане*; *машинно масло – машината ИЗПОЛЗВА масло* или *маслото Е ПРЕДНАЗНАЧЕНО за машини*. Изясняването на основните принципи за класифициране на единиците и създаването на единни и непротиворечиви категории е необходимо условие за приложимостта на класификацията.

Спорно е и твърдението, че тези предикати, въпреки общия си характер, имат потенциала да опишат всички възможни отношения между опората и подчинената част в случаите на изтрити предикати. Другото спорно твърдение е, че веднъж изтрити предикатите са възстановими.

Нека разгледаме съставните лексикални единици *пещерен лъв* и *пещерен протей* (Пример 14). Пещерният лъв е изчезнал вид лъв, останки от който са открити в пещери в Европа, Азия и Северна Америка и от това произлиза името му, а дали е обитавал пещери е спорен факт. Пещерният протей от друга страна получава името си, защото живее в пещери. Примерът илюстрира факта, че еднотипни съставни лексикални единици може да са образувани от различни предикати и възстановяването на предиката не е тривиално и не се съдържа (дори имплицитно) в съставната единица.

**Пример 14.**

- *Останки от пещерен лъв у нас досега са открити единствено от палеолита в Северна България — пещерата Бачо Киро.*

  лъв, [останките на който са открити в] пещери

- *Дишат с бели дробове; ларвите и някои възрастни (пещерен протей) - с хриле.*

  протей, [който обитава] пещери

  (БНК © ИБЕ)

Възможно е анализът да бъде ограничен до най-общото значение на предиката (изразено с В, англ. **IN**), като включва и случаи като *пещерен лъв* и да бъде прието,

че предикатът може да бъде възстановен. Спорно става тогава обаче, доколко толкова общи категории имат приложение за класификацията и описанието на съставните лексикални единици. Тук стои и въпросът, дали *пещерен лъв* не трябва да се причисли към категорията на идиосинкретично разложимите несвободни фрази с обяснението, че *пещерен* се проявява с нетипично значение, но от Пример 14 може да се види, че и двата конституента участват в конструирането на значението с лексикалното си значение, а идиоматичността се дължи на изтрития предикат.

#### 3.2.1.2. Класификация на съставните лексикални единици по семантични признаци

Проведени бяха обстойни наблюдения върху съставни лексикални единици, например от **Български тълковен речник** (1994), които са представени в **??**, и от Уикипедия и други източници; начините за тяхното извличане са описани в глава 4. Анализът и съпоставката им спрямо категориите на Гиржу и др. (2007) показа, че тези категории са недостатъчни за описанието на съставните лексикални единици в български език.

Това наложи адаптирането и допълването на класификацията за целите на описанието на съставните лексикални единици в български език. Бяха направени и някои допълнения и преработки в съответствие с предложенията на Леви (1978).

При разширяване на класификацията трябва да се отчита и фактът, че не всички семантични релации, които служат за образуване на синтактични фрази, служат за образуване и на съставни лексикални единици (например отношението *количество – същност*). При вторите се проявяват редица семантични ограничения, а във връзка с това се наблюдават и синтактични особености в съпоставка със свободните фрази.

Класификация 4 обособява категориите според значението на подчинената част и отношението с главната част в значението на съставната лексикална единица. Представената класификация от една страна е обобщение на двата подхода, изложени в 3.2.1.1, а от друга страна е опит за изграждане на класификация, която да отговаря на българските съставни лексикални единици.

**Класификация 4** (Отношения между подчинената част и опората в съставните лексикални единици)**.**

| **Релация** | | **Примери** |
|---|---|---|
| **Подчинена част** | **Главна част** | |
| Причина | Резултат (събитие, действие, състояние) | *астматичен пристъп*<br>*вирусна инфекция* |

| | | |
|---|---|---|
| Деятел | Резултат (продукт) | *птиче гнездо*<br>*съдебно решение* |
| Деятел | Резултат (дейност, състояние) | *народен избранник* |
| Резултат | Причина, причинител (активна или неактивна същност) | *маларичен комар* |
| Дейност | Инструмент | *сълзотворен газ* |
| Дейност | Деятел | *лекуващ лекар* |
| Инструмент | Същност (действаща същност) | *лазерен принтер* |
| Инструмент | Същност (дейност, състояние) | *информационна услуга* |
| Метод, дейност | Същност (действаща същност) | *тревопасно животно*<br>*лекуващ лекар* |
| Продукт | Производител (действаща същност) | *млечна жлеза*<br>*копринена буба*<br>*научен работник* |
| Източник, произход | Същност | *морска сол*<br>*слънчева светлина* |
| Материал | Същност | *найлонова торбичка*<br>*шоколадов бонбон* |
| Местонахождение | Обитател | *морска костенурка*<br>*пещерен лъв* |
| Местонахождение | Същност (събитие, действие, процес, обект) | *белодробна инфекция*<br>*междуетажна площадка*<br>*сърдечен масаж* |

| | | |
|---|---|---|
| Тема | Същност (инструмент, предавател, посредник) | *телевизионна антена*<br>*новинарска емисия*<br>*закон за авторското право* |
| Цяло | Част | *автомобилна гума*<br>*компютърен процесор* |
| Съдържание | Съдържаща същност | *картофена супа*<br>*златна мина* |
| Цел (предназначение) | Същност | *чаена чаша*<br>*плувен басейн*<br>*паста за зъби* |
| Същност (допълнителен признак) | Същност (основна) | *пчела работничка* |
| Друга характеристика | Същност | |

Тази класификация приема, че е налице предикат, който свързва подчинената и главната част, който се е трансформирал или е изчезнал при образуване на съставната лексикална единица. Степента, до която този предикат е възстановим от съставната единица, определя степента на семантична идиоматичност. Наред с предиката трябва да се установят и семантичните роли на подчинените части и на опората.

Прилагането на подобна класификация повдига някои въпроси (някои от които споменати и от Гиржу и др., 2007). От една страна стои въпросът за това, дали възможно ли е да се изгради система от краен брой категории, които напълно да опишат всички съставни лексикални единици, или системата би трябвало да се допълва с нови категории при необходимост. Ако подлежи на допълване, е важно доколко е допустимо разрастването на системата, тъй като особено обширна система ще бъде практически неефективна или неприложима. Възможно решение е да бъдат подбрани ограничен брой категории за целите на конкретно изследване или съобразени с изследвания корпус, но това обвързва подхода и описанието с конкретно изследване или ресурс, с което значително ограничава тяхната приложимост за други изследователски цели.

С по-голямо практическо значение би била класификация, която обръща внимание на идиоматичния характер на съставните лексикални единици, особено от гледна точка на тяхното значение, защото често съставните лексикални единици се описват като семантично немаркирани (Болдуин, 2004), тъй като значението им се формира

от значението на конституентите. Макар значението да се формира регулярно и нерестриктивно, в процеса участва и единицата $C$ (предикат, допълнително значение), което не е изразено от компонентите.

### 3.2.2. Класификация по синтактични признаци

В настоящия анализ са включени само фразите с опора съществително име. В 3.1.3 са описани синтактичните особености и структура на съставните лексикални единици. Предложената Класификация 5 се основава на частичен структурен анализ на фразите като последователност от думи от определена част на речта. Тази класификация има подчертано практическа насоченост с основна цел да опише възможните конструкции, за да се осигури тяхното разпознаване в текста. За целта се отчитат и особеностите на разделителите между компонентите на съставните лексикални единици, т. е. интервалите (Коева, 2006а).

**Класификация 5** (Изследвани синтактични конструкции)**.**
Централната колона представя опората. Всички интервали се приемат за празни (недопускащи вмъкване) с изключение на интервалите, означени с "_". Означения за частите на речта: A – прилагателно име; N – съществително име; P – предлог; Adv – наречие; B – числително име.

| **Предпоставени** | **Опора** | **Задпоставени** |
|---:|:---:|---|
| N | N | |
| | N | _ N |
| A _ | N | N |
| ?A _ A | N | N |
| ?A _ A A | N | N |
| ?N | N | P N |
| A _ | N | |
| A _ A | N | |
| A _ A A | N | |
| ?A _ A A A | N | |
| | N | _ P N |
| | ?N | _ P N N |

| **Предпоставени** | **Опора** | **Задпоставени** |
|---|---|---|
| | N | _ P A N |
| | ?N | _ P A A N |
| A _ | N | P N |
| A _ | N | P A N |
| A _ A | N | P N |
| ?Adv A _ | N | |
| ?A _ Adv | N | |
| ?A _ Adv A | N | |
| ?A _ A Adv | N | |
| ?B | N | |
| ?B A | N | |

Всички конструкции са с максимална дължина пет графични думи, тъй като максималната дължина на съставни лексикални единици, извадени от **Български тълковен речник**, е четири. Конструкциите, които са малко вероятни, са отбелязани със символа '?'; това са конструкциите, при които само незначителна част от единиците са несвободни фрази и които допринасят с незначителен процент към всички несвободни фрази (честотните данни са представени в Таблица 5.2, стр. 106).

Тъй като тук разглежданите съставни лексикални единици се ограничават до именни фрази, това улеснява класификацията и описанието им, тъй като някои много обширни категории (глаголите) отпадат – а именно тези, при които се допускат връзки между компоненти, където може да се вмъкне неопределена фраза (т. е. не клитика), както и случаи със свободен словоред. В случая с именна фраза се допуска само вмъкване на клитики, при това само на кратката форма на притежателните местоимения, докато при глаголите може да се вмъкват и други клитики, например въпросителната частица *ли*. При именните фрази не се допускат словоредни вариации при предпоставено определение, а са много редки случаите на разместване при съставни лексикални единици с предложни фрази.

Извън класификацията остават съставни лексикални единици, при които се използва координирано свързване на компоненти (съчинителни връзки), например *Министерство на образованието, младежта и науката*. Те остават също извън настоящото изследване, тъй като се срещат със сравнително малка честота, а представляват допълнителен проблем за автоматичната обработка на езика и ще усложнят зада-

чата за идентифициране на съставните лексикални единици, тъй като ще въведат множество кандидати със сложна структура, само незначителна част от които ще бъдат несвободни фрази (повече детайли са представени в следващата глава 4).

Няма да се разглеждат и съставни лексикални единици, при които фразата включва пунктуационни знаци – запетая за съчинително свързване на компоненти или кавички при каскадните наименования (Лесева и Стоянова, 2008). Не са включени и единици, при които единият компонент е изписан на латиница, например *HD телевизия* – разпознаването на тези единици е извън обсега на изследването по практически причини, тъй като тагерът не разпознава думи на латиница (*HD*) и ги тагира като неразпознати думи, поради което тези конструкции отпадат след приложение на синтактичния филтър. Разпознават се обаче конструкции със съкращения, изписани на кирилица, например *ДНК анализ*.

### 3.2.3. Приложна класификация според характеристиките на идиоматичността (обобщение)

Необходимо е да се подчертае, че за идиоматичната класификация използваме признака **идиоматичност** (англ. **idiomaticity**) в смисъла на Нунберг и др. (1994) – като основен признак на несвободните фрази, който се изразява в степента на конвенционалност, разбираемост и композиционалност. Със същото значение терминът 'идиоматичност' е използван и от Болдуин (2006), който говори за семантична, прагматична, синтактична и т. н. маркираност (повече детайли са представени в 2.2 и 2.3).

От изложеното дотук може да се обобщи, че идиоматичността на съставните лексикални единици се изразява в следните признаци:

- Съставните лексикални единици се характеризират с прагматична маркираност, която може да варира от липса на маркираност до силна маркираност. Тук разбирането за прагматична маркираност се конкретизира до това, дали значението се формира чрез референции към извънезикови обекти (т. е. дали единицата е наименование), което има отношение и към степента, до която е възстановима липсващата част $C$. Като експлицитно изразена връзка е определена тази, при която са експлицитни всички компоненти.
- Съставните лексикални единици се характеризират и с различна степен и характер на семантична идиоматичност, като степента може да варира от експлицитно изразена връзка до трудно възстановима неизразена връзка.

На базата на тези признаци можем изградим практическа класификация на съставните лексикални единици въз основа на тяхната идиоматичност и от гледна точка на тяхното автоматично идентифициране и обработка.

**Класификация 6** (Според идиоматичността)**.**

(–) Прагматично немаркирани единици без връзка или с идиосинкретична връзка (по смисъла на Болдуин и др., 2003) между компонентите принадлежат към неразложимите и идиосинкретично разложимите несвободни фрази.

(1) Прагматично маркирани единици (същински наименования), при които не се търси връзка между компонентите – например лични имена *Иван Петров*, *[училище] "Васил Левски"*.

(2) Прагматично маркирани единици (същински наименования), при които има подчинителна връзка между компонентите – например *Стара Загора*, *Горна Оряховица*, *Северна Корея*. При тях, макар и цялата фраза да е същинско наименование, се съдържа и значещ елемент; *Стара* носи значение, свързано с възраст; съпоставя се и с *Нова Загора*.

(3) Прагматично немаркирани съставни единици, при които връзката е трудно възстановима – например *пещерен лъв*, *морско конче*. Те формират гранична група между съставните лексикални единици и идиосинкретичните несвободни фрази, тъй като проявяват и черти на синкретичност на значението, т. е. некомпозираност и неразложимост.

(4) Прагматично маркирани съставни единици (наименования), формирани от поне два пълнозначни елемента, при които връзката е нестандартна и/или трудно възстановима – например *Граждани за европейско развитие на България* ≉ политическа партия, още повече *дясноцентристка консервативна политическа партия в България*; *Черно море* – определението *Черно* е част от името, което е отчасти мотивирано – скитите го наричат *Тъмноцветно*, а през Средновековието се утвърждава името *Черно море*, но мотивираността на съставката *Черно* вече е трудно възстановима; дескрипторът *море* е реализиран с обикновеното си лексикално значение.

(5) Прагматично маркирани съставни единици (наименования), при които едната съставка също е същинско наименование или производно от същинско наименование, а другата съставка е пълнозначен дескриптор; връзката между компонентите не е изразена, но е стандартна и поради това е (лесно) възстановима – *Челопеченско езеро* ≈ езеро, което се намира до Челопечене; *Лондонския мирен договор* ≈ мирен договор, подписан в Лондон.

(6) Прагматично немаркирани съставни единици, при които едната съставка е същинско наименование или производно от същинско наименование; връзката между компонентите не е изразена, но е стандартна и поради това е (лесно) възстановима – *синдром на Даун* ≈ синдром, открит от Даун; *български език* ≈ език, който се говори от българи.

(7) Прагматично немаркирани съставни единици, при които връзката не е изразена, но е стандартна и поради това е (лесно) възстановима – *морски бозайник*, *планинска коза* – местонахождение + животно показва хабитат. При тази кате-

гория формално съвпада с ... (например *пещерен лъв*).

(8) Прагматично маркирани съставни единици (наименования), при които връзката не е изразена, но е стандартна и поради това е (лесно) възстановима – *Организация за икономическо сътрудничество* ≈ *организация* + предлог *за* показва цел, предназначение; *Съюз на изобретателите* ≈ *организация* + предлог *на* показва принадлежност, членство в организация.

(9) Прагматично немаркирани съставни единици, при които връзката е изразена напълно експлицитно (например *сълзотворен газ* ≈ *газ, който твори сълзи*).

(10) Прагматично маркирани съставни единици (наименования), при които връзката между компонентите е изразена напълно експлицитно (например *Специализирана болница за активно лечение по онкология*) ≈ *болница, която Е специализирана да лекува активно онкологични състояния.*

Класификацията е приложна с оглед на това, че предлага формални показатели, чрез които да се идентифицират категориите. Значението ѝ се изразява също в това, че отделните категории изискват специфичен подход при обработката, особено с оглед на машинния превод.

От една страна наименованията са силно институционализирани, което означава, че е необходимо да се търси не само валиден, но и установен превод, който не винаги съответства точно на източника, например при превод от български на английски *Организация на обединените нации* → *United Nations* (единият компонент не се предава). От друга страна единиците, които не са наименования, често се характеризират с повече свобода. В някои случаи се наблюдава еднотипно предаване на конструкциите при превод от езика източник на езика цел, например *синдром на Даун* → *Dawn syndrome* (при всички конструкции *N на N*).

Изследването на особеностите на отделните категории, както и анализът с оглед на приложимостта на класификацията за целите на автоматичната идентификация на съставните лексикални единици и за други научно-изследователски задачи ще бъдат засегнати частично в следващите глави 4 и 5.

# 4. Основни методи за разпознаване и тагиране на съставни лексикални единици

През последните двадесет години широко се дискутират проблемите, които несвободните фрази поставят пред системите за автоматичен анализ на естествения език. Автоматичното им разпознаване се възпрепятства от факта, че тези единици съдържат повече от една графична дума, а разпознаването им като цялостна единица се налага от това, че демонстрират неделимост в семантично, синтактично и прагматично отношение.

В предходната глава 3 бяха очертани основните лингвистични особености на съставните лексикални единици и бяха изложени няколко различни класификации. С най-голямо приложение е класификацията, основана на идиоматичността на фразата, тъй като има пряко отношение към начините, по които трябва да се анализира значението на фразата. Анализът на идиоматичността е базиран на следните показатели:

- доколко значението на несвободната фраза е композирано от значенията на конституентите;
- доколко е значеща опората;
- доколко значението е институционализирано (фразата е наименование или не);
- доколко във формирането на значението участват наименования.

Настоящата глава има за цел да представи някои от по-широко прилаганите методи за разпознаване на несвободни фрази и съставни лексикални единици в частност, както и да направи анализ на факторите, които влияят върху ефективността на методите. По този начин ще бъде осигурена основата за представянето на резултатите от изследването върху автоматичното разпознаване на съставните лексикални единици в глава 5.

Най-общо методите за разпознаване на съставните лексикални единици могат да се разделят на три основни вида въз основа на типа информация, която използват:

- **Лингвистични методи** – разчитащи на лингвистичен анализ и ресурси;
- **Количествени методи** – основани на количествена информация за езиковите

единици и статистически тестове;

- **Хибридни методи** – съчетаващи лингвистичен анализ и статистически тестове.

Тъй като изследванията, съсредоточени специално върху разпознаването на съставните лексикални единици като отделно обособена категория, са силно ограничени, горната класификация на методите се основава на разпознаването на несвободните фрази, но е разгледана приложимостта им с оглед на разпознаването и класификацията на съставните лексикални единици.

В настоящата разработка не се цели изчерпателно изложение за всички разработени методи за автоматично разпознаване на съставните лексикални единици, а по-скоро да се очертае разнообразието от методи, като се отбележи приложимостта им за целите на идентифицирането на несвободни фрази и съставни лексикални единици. Представени са и някои отделни техники и похвати, които не изпълняват цялостна задача, идентифициране на несвободни фрази, и по тази причина не могат да бъдат определени като методи, но са важна съставка в изграждането на методологичната рамка и поради това намират място в настоящия анализ.

## 4.1. Характеристика на лингвистичните методи

Лингвистичните методи се основават на прецизно разработени лингвистични ресурси и модели, които описват лексикалните и граматическите характеристики на езиковите единици – думи, фрази, изречения. Лингвистичните модели са разработени от специалисти езиковеди и имат за цел да опишат възможно най-точно изследваното явление и да осигурят успешното и точното му разпознаване в текста. Като най-обща оценка може да се каже, че те се отличават с висока точност, но от една страна са органичени върху специфично езиково явление или отделна характеристика, а от друга – са неикономични за разработване, тъй като изискват голямо количество ресурси като време и човешки труд.

### 4.1.1. Лингвистични ресурси

Лингвистичните ресурси включват различни по съдържание и обем речници и системи от правила. Някои от по-широко използваните видове ресурси са представени по-долу, като е обърнато внимание на тяхната приложимост за автоматичното идентифициране на съставните лексикални единици.

- **Списъци от думи** – без допълнително граматическо описание.

  Могат да се използват списъци както за единични думи (Пример 15), така и за съставни единици или смесени (Пример 16). Най-често се използват за разпознаване на неизменяеми думи като същински названия. В някои случаи включват

всички възможни форми като отделни единици, а понякога могат да се съчетаят със система от правила, чрез които да се дефинират отделните форми, например правило за образуване на женски род и множествено число на фамилни имена. Например, ако дадено собствено име завършва на (1) *-ов*, (2) *-ев* или (3) *-ски*, могат да се образуват следните допълнителни форми: (1) *-ова*, *-ови*, (2) *-ева*, *-еви* или (3) *-ска*.

**Пример 15** (Извадка от списък с лични имена от една дума)**.**

| | | |
|---|---|---|
| Гергана | Гергинов | Геренов |
| Герганов | Гергинчев | Геренски |
| Герганчев | Гергов | Гереро |
| Гергелчев | Герд | Герзов |
| Гергиев | Герджиков | Гери |
| Гергина | Герджикович | Гериът |

Списък с лични имена © СКЛ

**Пример 16** (Извадка от смесен списък с топоними)**.**

| | | |
|---|---|---|
| Банкя | Баня | Баренцбург |
| Бановка | Баня Лука | Барешани |
| Баново | Баракли | Бари |
| Банска Бистрица | Бараково | Барун Урт |
| Бански дол | Баралевац | Батова река |
| Банско | Баранкиля | Батовско езеро |

Списък с географски имена © СКЛ

Подобни списъци могат да намерят приложение за разграничаване на отделните типове съставни лексикални единици – за откриване на същинските наименования и на категориите, чието значение се формира с участието на наименование.

- **Граматичен речник** – съдържащ лингвистично описание на включените в речника единици.

  Българският граматичен речник (Коева, 1998) се използва в тагера, чрез който бяха обработени текстовете от корпуса при предварителния етап на изследването (глава 5). В този речник се съдържа лемата на всяка дума, част на речта, както и граматическите й характеристики според частта на речта, към която принадлежи. Също така за съществителните имена има информация дали са собствени или нарицателни, което има отношение към целта за класифициране на съставните лексикални единици според степента на идиоматичност.

Българският речник на несвободните фрази (Коева, 2006а) използва сходна система за описание на несвободните фрази, както в Граматичния речник. Фразата наследява граматичния клас (част на речта) и основните граматически признаци (при съществително – род) на опората; указани са също броят на останалите конституенти и тяхната част на речта, както и вида на интервалите между отделните единици. Повече детайли за описанието са дадени в 3.1.3.

- **Лексикално-семантична мрежа** (онтология), например WordNet (**WordNet: An Electronic Lexical Database**; **Българският WordNet**) – съдържа семантична информация, която може да включва описание на значението (англ. **sense, gloss**), синонимни отношения в рамките на едно множество (англ. **synset**), други семантични отношения между единици от системата (хипонимия, хиперонимия, отношението *подобен на* при прилагателните и др.)

  В синонимните множества в БулНет (**Българският WordNet**) наред с простите думи са включени и несвободни фрази, които са 24.49% от литералите (отделните единици, членове на синонимните множества) според Коева (2006а).

  Освен за директно разпознаване на съставните лексикални единици, включени в БулНет, лексикално-семантичната мрежа може да намери и редица други приложения. Може да се използва семантична информация за определяне на категориите съставни лексикални единици, например разпознаването на организации и други отделни категории. Също така може да се определи дали състаната единица е едноцентрична или екзоцентрична според това дали съставната единица е хипоним на опората (или единицата от която е съставена, която може също да е съставна).

  Друго възможно приложение лексикално-семантичната мрежа може да намери при някои хибридни методи за идентифициране на съставни лексикални единици, като например метода, основан на латентния семантичен анализ.

- **Преводен речник** – съдържащ езикови единици и техните преводни съответствия на друг език.

  Намират приложение при изследване на междуезиковите отношения и превод от един език на друг. Преводните речници също могат да съдържат както единични думи, така и несвободни фрази. Традиционните преводни речници включват неразложимите и идиосинкретично разложимите лексикални единици, а някои са разширени и със съставни лексикални единици.

  Статутът на дадена фраза като несвободна или като съставна лексикална единица може да бъде определен и от наличието на преводен еквивалент, изразен с една дума (Коева, 2006а). Макар и този метод за разпознаване на съставните лексикални единици да не е ефективен, той може да помогне при изучаването на характеристиките и поведението на съставните лексикални единици на

междуезиково ниво.

- **Семантико-синтактичен речник**, като например ФреймНет – съдържащ информация за семантичните и езиковоспецифичните синтактични и лексикално-семантични ограничения при съчетаемостта на лексикалните единици (Коева, 2008).

  ФреймНет има за цел описанието на глаголите, като са включени техните задължителни и допустими обкръжения. Указват се също така семантичните и синтактичните ограничения върху обкръженията.

- **Системи от правила** – описващи изследваното явление.

  Прилагат се за изследване на конструкции със стандартна форма, като броят на привилата ще зависи от необходимата детайлност на модела – по-абстрактни модели ще се основават на по-малък брой, но по-общи оравила, докато по-детайлен модел ще включва повече по-детайлни правила, като ще позволява и по-фина класификация на изследваното явление.

  Например голяма част от наименованията на организации може да бъде описана с помощта на малък брой правила за възможните конструкции (Пример 17).

**Пример 17** (Правила за разпознаване на наименования на организации)**.**

- Общи правила
  - дескриптор за организация + предложна фраза
  - определение + дескриптор за организация
- По-конкретни правила
  - дескриптор за организация + предлог *за* + цел
  - дескриптор за организация + предлог *на* + членове
  - определение за цел + дескриптор за организация
  - определение за качество + дескриптор за организация
  - ...
- Още по-конкретни правила (според вида на организацията)
  - *организация* + предлог *за* + цел
  - *организация* + предлог *на* + членове
  - *съюз* + предлог *за* + цел
  - *съюз* + предлог *на* + членове
  - *асоциация* + предлог *за* + цел
  - *асоциация* + предлог *на* + членове

– ...

Често срещано явление са лексикални ресурси с комплексна структура, които съчетават различни видове информация. Лексикално-семантичната мрежа WordNet например, може да се ползва и като преводен лексикален ресурс, тъй като мрежите за отделните езици могат да се съпоставят на базата на междуезиковите индекси (англ. **Inter-Lingual-Index**, англ. **ILI**), които са уникално приписани на отделните значения, съответстващи на синонимните множества.

### 4.1.2. Приложение на лингвистичните методи

Лингвистичните методи се състоят в приложение на лексикалните и граматичните ресурси върху изследваните корпуси от текстове и маркирането на съставните лексикални единици, които се съдържат в речниците или отговарят на зададената система от правила.

В случая със съставните лексикални единици, както вече беше аргументирано в 3.1.6, само малка част от тях е оправдано да бъдат включени в лексикални ресурси – списъци или речници, като например същинските наименования. При останалите това не е оправдано, тъй като се получава дублиране на лингвистичната информация. Също така съставните лексикални единици са разнообразна и продуктивна категория, която не може да бъде обхваната изцяло и описана в статичен речник.

Поради тази причина чисто лексикалните методи имат ограничено приложение за целите на разпознаването на съставни лексикални единици. По-специално те се използват за изследване на отделни категории съставни лексикални единици и не са универсални. Системите от семантични и синтактични правила в общия си вид намират широко приложение в хибридните методи за разпознаване.

## 4.2. Характеристика на количествените методи

В областта на компютърната лингвистика и компютърната обработка на естествените езици отдавна се обръща внимание на комбинациите от думи, които проявяват сравнителна устойчивост и се появяват със значителна честота в езика. Говори се за **асоциация** (англ. **association**) между думите, както и за **колокации** (англ. **collocation**), които описват явлението на честа съвместна поява на дадена дума ядро, заедно с колокати.

Количествените методи разчитат на това, че от даден корпус може да бъде извлечена количествена информация, която да служи като показател дали дадено съчетание от думи е колокация. Колокациите се разглеждат като потенциални несвободни фрази и съставни лексикални единици, след което се търсят начини за определянето

им като свободни или несвободни, както и за причисляването им към определена категория според целите на конкретното изследване.

Голяма част от преводните еквиваленти на статистическата терминология за заимствани от Вындев (2003a) и Вындев (2003b). Математическата аргументация на методите и извеждането на отделни формули няма да бъде изложена в настоящата разработка, като в повечето случаи ще бъдат посочени основни източници по тези въпроси. Повече внимание ще бъде обърнато на методите, използвани за експериментите в глава 5.

### 4.2.1. Основни математически понятия и похвати

Манинг и Шутце (1999, стр. 153) и Еверт (2005, стр. 76–77) представят набор от разнообразни количествени методи, които разделят на няколко групи според подхода към количественото измерване на асоциацията между думите.

Други разработки, които представят обзор на различни математически (статистически) методи и похвати, както и сравнителен анализ между тях, са например работите на Да Силва и Лопез (1999), Еверт (2007), Ким и Болдуин (2007b), Рамиш и др. (2008), Ким (2008), Печина (2008), Жанг и др. (2009) и др.

В резултат на проучването на голям набор от методи, тук са подбрани някои основни видове информация за представяне на количествените данни и похвати за техния анализ, измежду които с особено значение са следните:

- Прост честотен анализ;
- Честотни таблици;
- Тестване на хипотеза;
- Мерки за асоциация между думите ($\chi^2$, логаритмично подобие и др.);
- Мерки от теория на информацията (ентропия и взаимна информация);
- Факторен анализ;
- Евристични техники.

#### 4.2.1.1. Прост честотен анализ

Най-простият метод за откриване на колокации се основава на честотата на тяхната поява. Ако комплекс от думи се среща с голяма честота, това сочи, че вероятността комплексът да изпълнява специфична самостоятелна функция е по-голяма (Манинг и Шутце, 1999). Непосредственото прилагане на метода обаче не дава добри резултати, тъй като с най-голяма честота се появяват съчетания със служебни думи (пример е представен в 5.3.2).

Джъстесон и Катц (1995) предлагат проста евристика за подобряване на резулта-

тите от честотния метод – използване на синтактичен филтър, чрез който се отстраняват неподходящите кандидати за колокации, като се разглеждат само тези, които отговарят на определена синтактична структура (вж. 4.3). Подобен метод използва и Смаджа (1993). Прилагането на синтактичен филтър разчита на лингвистична информация, което причислява този метод към групата на хибридните методи.

Методът е приложим за тематични области, в които некомпозираните и идиосинкретично композираните несвободни фрази се появяват с малка честота, тъй като при него се разпознават институционализирани (статистически маркирани) единици, без да е възможно да се диференцира между отделните категории.

#### 4.2.1.2. Представяне на честотните данни

Честотните данни се представят във формата на честотна таблица (англ. **contingency table**)[1]. За две думи $w_1$ и $w_2$, които се появяват в корпуса с честота съответно $f_1$ и $f_2$, се използва таблица от наблюдения с размери $2 \times 2$, в която $O_{11}$ означава броя на срещанията на $w_1$ и $w_2$ заедно, $O_{12}$ и $O_{21}$ – броя срещания съответно само на $w_1$ или само на $w_2$, а $O_{22}$ – броя случаи, при които не присъства нито едната дума (Пример 18).

След това се изчисляват очакваните честоти, като се използват сумите по ред ($R_1$ и $R_2$) и по колона ($C_1$ и $C_2$) в оригиналната таблица. Тези суми се наричат още маргинални честоти (англ. **marginal frequencies**). $N$ означава броя на всички думи в корпуса. (Еверт, 2007, стр. 24–28)

**Пример 18** (Честотни таблици за две думи $w_1$ и $w_2$)**.**

| | $w_2$ | $\neg w_2$ | |
|---|---|---|---|
| $w_1$ | $O_{11}$ | $O_{12}$ | $R_1 = \sum_j O_{1j} = f_1$ |
| $\neg w_1$ | $O_{21}$ | $O_{22}$ | $R_2 = \sum_j O_{2j}$ |
| | $C_1 = \sum_i O_{i1} = f_2$ | $C_2 = \sum_i O_{i2}$ | $N = \sum_{i,j} O_{ij}$ |

Таблица на наблюдаваните честоти

[1] Макар и да съществува малко несъответствие между английския термин и българския превод, е използван терминът 'честотна таблица', както е даден от Вындев (2003а, стр. 41–44).

| | $w_2$ | $\neg w_2$ |
|---|---|---|
| $w_1$ | $E_{11} = \frac{R_1 C_1}{N}$ | $E_{12} = \frac{R_1 C_2}{N}$ |
| $\neg w_1$ | $E_{21} = \frac{R_2 C_1}{N}$ | $E_{22} = \frac{R_2 C_2}{N}$ |

Таблица на очакваните честоти

При съчетания от три думи $w_1$, $w_2$ и $w_3$ се използва тримерна таблица, т. е. куб с размери $2 \times 2 \times 2$, осемте елемента на която са от следния вид:

$(w_1, w_2, w_3)$,
$(\neg w_1, w_2, w_3)$, $(w_1, \neg w_2, w_3)$, $(w_1, w_2, \neg w_3)$
$(\neg w_1, \neg w_2, w_3)$, $(\neg w_1, w_2, \neg w_3)$, $(w_1, \neg w_2, \neg w_3)$
$(\neg w_1, \neg w_2, \neg w_3)$

По-нататък са дадени формули за изчисляване на асоциацията между думите в съчетания от две и от три думи. При тях се използва информацията от честотните таблици.

#### 4.2.1.3. Проверка на хипотеза

Методите с проверка на хипотеза имат за цел да покажат дали появата на дадена комбинация от думи е случайна или не. За целта се формулира нулевата хипотеза

$$H_0 \ : \text{Думите не са свързани,}$$

т. е. не формират колокация, а се появяват в комбинация случайно. На нулевата хипотеза се противопоставя алтернативната хипотеза

$$H_1 \ : \text{Думите са свързани.}$$

Изчислява се вероятността $p$ комбинацията да се появява с наблюдаваната честота при положение, че $H_0$ е вярна. Ако тази вероятност е малка, то $H_0$ се отхвърля, а в противен случай $H_0$ се приема за истинна. Задава се **критично ниво** $\alpha$, което представлява вероятността да бъде отхвърлена вярна хипотеза, като обикновено се избира малка стойност, например $\alpha = 0.05$. Числото $1-\alpha$ се нарича **ниво на доверие**.

#### 4.2.1.4. Мерки за асоциация между думите

**Мярката за асоциация** е формула, по която се изчислява асоциацията между отделните елементи на комбинация от думи. Стойността е показател за това, колко силна е връзката между думите, като се отстранява случайният ефект.

В някои случаи се използва **мярка за разстоянието** между компонентите на съчетание от думи. Докато при измерване на асоциацията при по-голяма стойност думите са по-близки, отколкото при по-малка, то при мярката за разстояние е обратното – по-малката стойност показва по-голяма близост от по-голямата.

Едни от най-често прилаганите мерки за асоциация между думите са следните:

- $\chi^2$

$$\chi^2 \;=\; \sum_{i,\,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Логаритмично подобие (англ. log**-likelihood**)

$$\text{log-likelihood} \;=\; 2 \sum_{i,\,j} O_{ij} \;\log \frac{O_{ij}}{E_{ij}}$$

- Взаимна информация (англ. **mutual information, MI**)

  Оригиналната формула за взаимна информация между думите $x$ и $y$ е от вида

$$I(x,y) \;=\; \log_2 \frac{P(x,y)}{P(x)\,P(y)},$$

  където $P(x)$, $P(x)$ и $P(x,y)$ представляват вероятността за поява в корпуса съответно на $x$, $y$ и съчетанието $x,y$ (Чърч и Ханкс, 1990). Взаимната информация сравнява вероятността да се появи съчетанието $x,y$ с вероятностите $x$ и $y$ да бъдат наблюдавани като независими променливи.

  Ако $x$ и $y$ са независими, то ще се наблюдава

$$P(x,y) \;\approx\; P(x)\,P(y) \qquad \text{и} \qquad I(x,y) \;\approx\; 0,$$

  докато при съществуваща връзка между двете думи ще се получи

$$P(x,y) \;\gg\; P(x)\,P(y) \qquad \text{и} \qquad I(x,y) \;\gg\; 0$$

  Взаимната информация регистрира и случаите, при които има взаимно изключваща се поява на думите $x$ и $y$ (антиколокации), при която

$$P(x,y) \;\ll\; P(x)\,P(y) \qquad \text{и} \qquad I(x,y) \;\ll\; 0$$

  Вероятностите $P(x)$, $P(y)$ и $P(x,y)$ се получават като честотата на съответната дума $f_x$, $f_y$ или фразата в корпуса $f_{x,y}$ се нормализира по $N$ – големината на корпуса.

- Подобрена мярка за взаимна информация (англ. **enhanced mutual information, EMI**)

  Жанг и др. (2009) предлага подобрена мярка за взаимна информация, която се представя със следната формула

$$EMI(x,y) = \log_2 \frac{P(x,y)}{P(x)\,P(y)\left(1-\frac{P(x,y)}{P(x)}\right)\left(1-\frac{P(x,y)}{P(y)}\right)},$$

$$= \log_2 \frac{P(x,y)\,N}{(f_x - f_{x,y})(f_y - f_{x,y})}.$$

Тези мерки са използвани в програмата за български език, представена в глава 5.

Представените формули измерват зависимостта между две думи. Има няколко възможности мерките да се разширят за три и повече единици. Единият начин е да се използват някои готови форуми (Жанг и др., 2009). Друг възможен подход при съставните лексикални единици с дължина от три и повече думи е изчисляването на асоциацията последователно, като на всеки етап $n$ се разглежда присъединяването на една дума $w_n$ към комбинацията $w = w_1\,w_2\,\cdots\,w_{n-1}$, формирана на етап $n-1$, която също се разглежда като една дума $w$.

Комбинации от повече от две думи обаче няма да бъдат разглеждани, тъй като честотните анализи (вж. 5) показаха, че конструкциите с два пълнозначни елемента покриват 95.5% от всички съставни лексикални единици, и мерките ще бъдат приложени за изследване на асоциациите между два пълнозначни елемента.

### 4.2.2. Значение на количествените методи

Количествените методи се появяват като алтернатива на лингвистичните. Те разчитат на това, че при наличие на значително количество данни, езикови единици, демонстриращи сходни явления, ще проявяват и сходно поведение. Поведението може да се изразява в честота на появя в корпуса, системна поява в определено обкръжение и др.

Разработването и прилагането на количествени методи включва две основни задачи:

- Намиране на подходяща количествена мярка, която да е показателна за изследваното явление.

  Мярката много често се определя в зависимост от поставената задача и наличните корпуси (размер, тематична област и др.) При съставните лексикални единици се търсят начини за измерване на това, доколко отделните думи в рамките на съставната единица са свързани една с друга и затова се търсят количествени показатели, чрез които да се установи в кои случаи появата на

дадена комбинация от думи не се основава на шанс, а на наличието на определена закономерност между елементите.

- Установяване на обосновани и еднозначни начини за тълкуване на резултатите.

  Като резултати е необходимо е да бъдат формулирани научно обосновани изводи, които да не се базират на субективна преценка или произволно тълкувание.

Количествените методи не изискват наличието на езикова информация, с което се избягва нуждата от разработването на големи по обем езикови ресурси – речници и други, което е трудоемка и често икономически нерентабилна задача. От друга страна обаче се оказва, че чисто количествените методи не дават особено добри резултати и това налага все по-широкото използване на хибридни методи.

## 4.3. Хибридни методи

Количествените методи постигат ограничени резултати, когато се използват самостоятелно. Необходимо е те да бъдат базирани на лингвистичен анализ на изследваното явление, в случая съставните лексикални единици. Това ще позволи изследването да бъде съобразено със спецификата на проучваното явление, при което значително ще се намали броят на наблюдаваните фрази кандидати, което пък от своя страна ще доведе до подобряване на резултатите.

На първо място приложението на лингвистична информация се изразява в тагирането на корпуса с част на речта и граматична информация. Това позволява да бъде приложен синтактичен филтър, при което значително се ограничава броят на изследваните кандидати, като се разглеждат само тези, които отговарят на дадената синтактична структура.

Освен това кандидатите формират по-еднородна група, което ще позволи по-концентрирано изследване върху особеносите на съставните лексикални единици.

### 4.3.1. Прилагане на синтактичен филтър

Прилагането на синтактичен филтър за подбиране на релевантните за изследването синтактични конструкции е изключително прост и ефективен метод, който значително намалява възможните кандидати, с което от една страна се намалява времето, за което се извличат автоматично несвободните фрази, а от друга страна се подобрява значително и точността на резултатите. Методът е използван от Джъстесон и Катц (1995), Смаджа (1993), и др.

За български език методът е използван от Коева (2007b) и Стоянова (2009) и е демонстрирано подобрението в резултатите, което е постигнато чрез приложението на синтактичния филтър.

По същество лингвистичен, този метод се използва като допълнителен, заедно с

честотен анализ или мерки за асоциация, за да отсее подходящите от неподходящите кандидати, затова тук е разгледан в категорията на хибридните методи.

### 4.3.2. Векторно представяне на значението

През последното десетилетие анализът на несвободните фрази често се обвързва с анализа на тяхната идиоматичност в семантично отношение. Пиърс (2001) предлага метод, използващ WordNet, подобен е и методът на Оравеч и др. (2005), които обаче ползват синонимен речник. В тези изследвания се разчита на това, че несвободните фрази често блокират възможни синонимни замени на конституенти, а когато допускат такива, те са силно ограничени, т. е. авторите измерват идиоматичността чрез (не)възможността за семантични замени.

Започват да се прилагат и методи, при които значението се представя като вектор и се анализира математически. Болдуин и др. (2003) предлагат метода на латентния семантичен анализ, а Гизбрехт (2009) описва друг подобен метод, който използва пространствените модели на думите за геометрично представяне на значението.

Векторнопространствените модели са основани на хипотезата, че значението се формира от контекста, която е издигната първоначално от Фирт (1957). Значението се представя като вектор в $n$-мерно векторно пространство, където измеренията са представени от пълнозначни думи. При латентния семантичен анализ се избира голяма стойност за $n$, след което се прилага Singular Value Decomposition (SVD, Дируестър и др., 1990) за намаляване на броя на измеренията.

Гизбрехт (2009), следвайки Уидоус (2008), предлага модел за композиционалността, като описва четири възможни векторни операции, чрез които да се изрази композирането на несвободни фрази $w\,v$ от векторите $w$ и $v$.

1. $(w\,v)_i = w_i + v_i$ (Събиране на вектори)
2. $(w\,v)_i = w_i \cdot v_i$ (Скаларно произведение)
3. $w\,v = w \otimes v = A$ (матрица) и $A_{ij} = w_i \cdot v_j$ (Тензорно произведение)
4. $(w\,v)_i = \sum_j w_j \cdot v_{i-j+1}$ (Конволюция).

За изчисляване на подобието между векторите в първия, втория и последния случай се използва косинуса между тях, при което два израза се приемат за несвързани, ако косинусът е 0 и синонимни, ако е 1, т. е. ако двата израза са успоредни.

Количественото измерване на подобието между всеки от двата компонента $w$ и $v$ и фразата $(w\,v)$ ще даде информация доколко значението е композирано. Близостта на значенията на компонентите до това на фразата е показател за композираност на значението, докато отдалечеността на векторите е показател, че значението на компонента не присъства по разпознаваем начин във фразата, което пък свидетелства за по-висока степен на идиоматичност.

Този подход, наред с латентния семантичен анализ, въобще векторното представяне на значението, може да бъде изключително полезно за целите на автоматичното разпознаване на съставните лексикални единици и най-вече за разграничаване между тази и другите категории несвободни фрази – неразложимите и идиосинкретично разложимите.

Уидоус (2008) и Гизбрехт (2009) изтъкват възможностите, които този подход предлага за изследване на композиционалността, но той все още не е подробно тестван за целите на разпознаването на несвободните фрази и съставните лексикални единици.

## 4.4. Анализ на факторите, влияещи върху ефективността на методите

Съставните лексикални единици представляват широка и разнообразна група от явления, които се разграничават от другите категории несвободни фрази по това, че значението им е композирано и разложимо – то е формирано с участието на значенията на конституентите. От друга страна обаче съставните лексикални единици също се характеризират с определена степен на идиоматичност, което ги отделя от групата на свободните фрази. В 3.2.3 е представена класификация на съставните лексикални единици според степента и вида идиоматичност, която се определя от семантични и прагматични фактори.

Основните фактори, които определят успешното изпълнение на задачата за разпознаване на съставните лексикални единици, са следните:

- **Конкретната цел на изследването. Съответствие между целите и метода**.

  Важно е дали се цели разпознаването на несвободните фрази, на съставните лексикални единици като цяло или и на отделните категории съставни лексикални единици.

- **Системата на класификация**.

  Критериите за класификацията, броят на отделните категории и техните характеристики имат особено значение за ефективността на метода. При по-подробна класификация например е необходимо да бъде избран по-чувствителен метод за съответните признаци.

- **Текстовите ресурси**.

  Характеристиките на изследвания корпус (големина, тематична област и др.) също оказват влияние и затова подборът на подходящи ресурси е особено важна задача. При изследване на по-редки явения например е необходимо използването на по-голям корпус от текстове, особено при използването на статистически методи, тъй като те не дават добри резултати за явления с ниска честота.

Някои явления също така са специфични за определени тематични области или стилове (например терминологията). Ето защо е необходимо да се подбрере корпус, в който явлението е достатъчно добре представено.

Трябва да се има предвид също така, че резултатите от повечето методи ще бъдат различни при прилагането им върху специализиран корпус или общоезиков корпус. В разработката на Стоянова (2010) е даден пример с *тънък клиент* и метода на латентния семантичен анализ в общ и в специализиран корпус, като във втория случай резултатите са значително по-добри.

- **Анотацията на корпуса**.

  Лингвистичната анотация на корпуса – тагоране с част на речта и граматически характеристики, семантична информация и др. има значение за това, какви методи могат да бъдат приложени за разпознаването на съставните лексикални единици.

- **Езиково специфични характеристики**.

  Свободният словоред в българския език например допуска много кандидати за съставни лексикални единици глаголни фрази и с това значително затруднява анализа. Макар и глаголите да не влизат в настоящия анализ, те са пример за езиково специфична черта, която има силно влияние върху успешността на даден метод.

## 4.5. Възможности за разпознаване на съставните лексикални единици

Прави впечатление, че различните типове методи имат свойството на регистрират различни характеристики на несвободните фразии съставните лексикални единици. Количествените методи например, по-специално методът, основан на честотата на срещане, отчита факта, че несвободните фрази са институционализирани и поради това се появяват в постоянна форма и именно на това се дължи и по-високата им честота.

В такъв случай ще се очаква този метод да дава добри резултати при разграничаване на всички категории несвободни фрази от свободните фрази, но не е чувствителен към различията между отделните категории несвободни фрази, тъй като не отчита особеностите на семантичната им структура.

Методите, които представят значението във векторна форма, предлагат възможност за измерване на семантичната композиционалност като обусловена от контекста. Този метод е подходящ за разграничаване на разложимите несвободни фрази (съставни лексикални единици) и неразложимите и идиосинкретично разложимите (фразеологизми). Съставните лексикални единици обаче проявяват сходство със сво-

бодните фрази, тъй като значението им е композирано и се наблюдава близост между компонентите и фразата, което прави векторният метод неподходящ в този случай.

Чрез прилагане на практическата Дефиниция 4 (стр. 38) и представянето ѝ във 2.2 (стр. 38), може да се очертаят няколко последователни етапа на разпознаването на съставните лексикални единици чрез последователното им разграничаване от свободните фрази и другите категории несвободни фрази. Първият етап е свързан с разпознаването на несвободните фрази и за него може да се използва честотният метод и различните мерки за асоциация.

Вторият етап е разграничаването между съставните лексикални единици и останалите категории несвободни фрази, при което е подходящи прилагането на векторен метод за анализ за представяне на значението. Прилагането на различни мерки за асоциация също може да даде добри резултати при тази задача.

Допълнително предизвикателство е отделянето на съставните лексикални единици от свободните колкокации, които също проявяват институционализираност на формата и композиционалност на значението, поради което няма да бъдат отделени в предходните два етапа.

В този случай е подходящо изследването на възможностите за замяната със синонимни изрази, което ще покаже по-голямата свобода на свободните колокации спрямо съставните лексикални единици. Друг възможен подход за тяхното отделяне е и детайлното изследване на векторния модел и възможностите за това, да бъде развит така, че да отчита по-слабата композиционалност на свободните колокации. Не на последно място стои възможността от прилагане на лингвистичен метод, който да оцени степента и характера на композиционалността. За целта ще трябва да бъде изграден добър лингвистичен модел на композиционалността и методология за нейното изследване.

За разграничаване на отделните категории съставни лексикални единици (Класификация 6, стр. 76) са необходими лингвистични методи, които да разпознават определени семантични и синтактични модели, речници с наименования и други ресурси. Успешното отделяне на категориите може да подобри значителното качеството на компютърни приложения за машинен превод и извличане на информация.

В следващата глава 5 са представени експерименти, които демонстрират приложението на методите на няколко етапа. Избраният корпус *Уики1000+* съдържа научнопопулярни текстове. Поради това броят на неразложими и идиосинкретично разложими несвободни фрази е много малък – те представляват едва 1.25% от несвободните фрази в корпуса. Поради това се оказва недостатъчно за качественото приложение и оценка на метода за тяхното разграничаване от съставните лексикални единици.

От друга страна основните приложения на разпознаването и тагирането на съставни лексикални единици са в научните тематични области за автоматично извличане на терминология, автоматичен превод и др. Това се отразява и от характера на из-

ползваните корпуси в различни изследвания по въпросите на несвободните фрази – основно научни и административни корпуси.

Без да се пренебрегва значението на задачата за разграничаване на съставните лексикални единици и останалите категории несвободни фрази, тази задача ще бъде оставена за бъдещата работа по проблемите на съставните лексикални единици, а настоящият анализ ще се ограничи с методите за разпознаване на несвободни фрази, които в корпуса *Уики100+* преобладаващо са съставни лексикални единици, след което ще бъдат дадени някои възможности за разграничаване на отделните категории съставни лексикални единици.

# 5. Автоматично разпознаване и тагиране на съставни лексикални единици в българския език

Настоящата глава има за цел да представи серия от експерименти, прилагащи методи за автоматично разпознаване и тагиране на съставни лексикални единици в българския език. На първо място трябва да се изтъкне фактът, че използваните методи, както и описаните в предходната глава 4, са разработени за целите на обработката на несвободни фрази в различни езици. Тъй като не беше намерена цялостна система за разпознаване и анализ на съставните лексикални единици, в 4.5 беше направен опит за очертаване на етапите на такъв метод. Неговото приложение е описано в 5.3.

Методите, представени тук, имат експериментален характер и са адаптирани към конкретната задача за разпознаване и тагиране на съставни лексикални единици в българския език.

Изложението в тази глава започва с кратък преглед на някои налични компютърни програми, които включват измежду функционалностите си такива за разпознаване и определяне на несвободни фрази. След този преглед е оредставена новоразработената за целите на дисертацията компютърна програма `bgMWE`, която включва набор от инструменти за обработка на текстови корпуси, разпознаване на съставни лексикални единици по няколко различни метода и отчитане на резултатите. Експериментите по приложението на методите са проведени с помощта на `bgMWE` и са представени в 5.3.

## 5.1. Преглед на някои компютърни програми за разпознаване на несвободни фрази и съставни лексикални единици

### 5.1.1. Xtract

`Xtract` е компютърна програма за автоматично извличане на колокации от големи текстови корпуси (Смаджа, 1993). Програмата използва статистически тестове за колокации, наред с филтриращи методи, чрез които се повишава точността на резултатите.

Програмата има три фази на работа. При първата се изчисляват мерките за асоциация на всички биграми и се извличат значими биграми (англ. **significant bigrams**) с честота над определена граница.

При втората фаза се идентифицират несвободните фрази в цялост. Това изисква преминаване от биграми към $N$-грами, като се избират максималните $N$-грами, т. е. отстраняват се всички $M$-грами ($M < N$), които се съдържат в максималния $N$-грам.

Третата фаза представлява приложение на синтактичен филтър за подбор на $N$-грамите, които формират фрази. При този етап по данни на Смаджа (1993) отпадат около половината от колокациите и качеството на резултатите значително се подобрява. Докато точността след втората фаза е 40%, при третата се покачва до 80%. Отчетено е

Този подход разчита на две основни предположения:

- Значимите биграми се появяват заедно по-често, отколкото биха се появявали, ако бяха случайни съчетания.
- Заради синтактичните ограничения върху колокациите, комбинациите ще се появяват в относително постоянна форма.

Смаджа (1993) отчита, че програмата работи добре за извличане на колокации със сравнително висока честота на поява в корпуса (от порядъка на няколко десетки срещания в корпус от 10 милиона думи). Смаджа (1993) обръща специално внимание на характеристиките на корпуса, които имат значение за резултатите от работата на `Xtract` – вид (общоезиков или специален), големина (с по-голям се постигат по-добри резултати) и тематична област (много колокации са тематично обусловени).

### 5.1.2. TermeX

`TermeX` (TermeX, 2009) е програма, разработена от Факултета по инженерни и компютърни науки на Университета в Загреб, Хърватско (Делач и др., 2009). Основната фунционалност на програмата е автоматичното разпознаване на колокации и терминологични конструкции.

Работата на `TermeX` се базира на статически методи, приложени върху лематизиран корпус, като е използван и синтактичен филтър, но без цялостно тагиране на корпуса с части на речта с цел съкращаване на времето за извличане на колокации. Имплементирани са четиринадесет различни мерки за асоциация между думите. Дава се възможност за избор на мярка на асоциация, дължина на разглежданите $N$-грами (от 2 до 4), а впоследствие – за ръчно селектиране на подходящите единици.

Програмата поддържа английски и хърватски език. Тя е независима от операционната система и предлага и графичен потребителски интерфейс.

### 5.1.3. Collocation Extract

`Collocation Extract` (CollocationExtract, 2000) е програма за извличане на колокации от корпус. Колокациите могат да са с дължина от 2 до 5 думи. Използвани са три различни статистически метода – логаритмично подобие, взаимна информация и $\chi^2$.

Позволява се избор на минимална честота на поява на колокациите, с което може да се регулира точността на резултатите според големината на корпуса и честотата на определени конструкции. Могат да се търсят колокатите на дадена главна дума или да се извличат всички колокации с определена дължина и произволна главна дума (ядро), като може да се избира и посока на търсене на колокатите – в дясно, в ляво или от двете страни на ядрото.

### 5.1.4. NooJ

`NooJ` (NooJ, 2002) е компютърна система, която представлява лингвистична среда за обработка на естествен език. Тя е по-различна от представените по-горе програми, тъй като функционалностите ѝ не се ограничават до извличането и обработката на колокации и несвободни фрази, а предлага широк спектър от инструменти за разнообразна обработка на текстови корпуси на различно лингвистично ниво.

`NooJ` притежава модул за български език с редица ресурси, разработени от Секцията по компютърна лингвистика (ИБЕ, БАН). Позволява създаването и приложението на лексикални ресурси, съдържащи несвободни фрази, наред с речниците от прости думи. Освен чрез лексикални ресурси, несвободните фрази могат да се анализират и с помощта на регулярни граматики. Има възможности за извличане на синтактични конструкции (синтактичен филтър), налагане на синтактични и други ограничения и т. н.

Проведеното изследване от Лесева и Стоянова (2008), извършено с помощта на `NooJ`, се занимава с извличане на наименования, включително съставни лексикални единици – имена на организации.

`NooJ` е базирана на `.NET`, което я прави зависима от операционната система и затруднява използването ѝ под `Linux`[1].

### 5.1.5. Обобщение

Представените програми, макар и да не се занимават конкретно със задачата за разпознаване на съставни лексикални единици, включват функционалности, които

[1] Възможно е използването на системата под `Linux` през `W ine` (http://www.winehq.org/), но е затруднено, свързано е с по-бавна скорост и някои функционалности и системи за кодиране не се поддържат.

могат да бъдат полезни за извличането на несвободни фрази и тяхната обработка. Те също така предлагат интересни и полезни идеи за това, какви функционалности трябва да предлага компютърна система, занимаваща се с проблемите на съставните лексикални единици.

Въпреки наличието на редица програми с подобни функционалности, не беше открита програма, която да е съсредоточена или най-малкото да включва задачата за разграничаване на отделните категории несвободни фрази, както и за разпознаване на съставни лексикални единици като отделна категория.

Наред с това представените програми, както и други подобни, демонстрират следните слабости:

- Голяма част от програмите на са предоставени за свободен достъп, а са комерсиални. Други макар да са свободни, не са лесно достъпни за сваляне през интернет.
- Някои програми поддържат само ограничен брой езици; други – само определени енкодинги (кодировки), често не се поддържа кирилица;
- Зависимост от операционната система.
- Липса на документация документация.
- Преустановена поддръжка.

## 5.2. Разработване на компютърна програма за разпознаване и тагиране на съставни лексикални единици

Въз основа на прегледа в 5.1 беше взето решение за изработване на нова програма, която да се съсредоточи конкретно върху проблемите на разпознаването и третирането на съставни лексикални единици, която да поддържа български език (кирилица), като същевременно не е езиково зависима и позволява обработването и на текстове на други езици.

Основната цел на разработената програма към момента е да служи за експериментално приложение на разработени методи за разпознаване на съставни лексикални единици. Тя е в процес на допълване и усъвършенстване – добавяне на нови функционалности и подобряване на вече имплементирани.

В 5.2 са описани основните особености на разработената компютърна програма и някои от допълнителните ѝ функционалности. Повече внимание ще бъде обърнато на основните функционалности, свързани с приложението на методите за разпознаване и тагиране на съставните лексикални единици, които са отделно представени в 5.3.

### 5.2.1. Архитектура и някои особености на програмата

Програмата `bgMWE` се занимава с цялостната обработка и анализ на корпуси – преобразуване на текстовете между различни формати, тагирането с граматическа информация, прилагане на лексикални ресурси като речници със съставни лексикални единици, разпознаване и тагиране, анализ и оценка на резултатите. Започнато е и разработването на графичен потребителски интерфейс за наблюдение на резултатите, извличане на примери, ръчно маркиране и тагиране, валидизация и др.

Проектът е реализиран на `Java`, което го прави независим от операционната система. Той ще бъде разпространен като отворен код, за да може в бъдеще да се използва за изграждане на разнообразни приложения, свързани с анализ на естествения език.

Мащабността на задачата, необходимостта от имплементиране на различни методи и подходи към анализа, както и големият брой допълнителни задачи, свързани с преобразуване между различни формати, извличане на лексикални ресурси и други, наложиха избора на модулното проектиране като основен подход. При модулното програмиране работата на програмата се разделя на ясно обособени функционалности, които са имплементирани в отделни модули. Това улеснява разработването, тестването и бъдещото усъвършенстване.

Всеки модул на `bgMWE` може да се използва като отделна програма за изпълнението на неговата задача. Приложението на отделните модули върху корпуса става последователно. В някои случаи е възможно интегрирането на една функционалност в друга (например тагиране с част на речта и преобразуване в XML формат), което ще намали времето за обработване, но с цел простота и лесно тестване, е предпочетено това да се извършва последователно, особено в случаите, когато операциите се извършват еднократно върху корпуса.

Допълнителните приложения, които са реализирани като отделни модули, включени в рамките на програмата, са описани в Приложение Г. Те могат да бъдат използвани и самостоятелно.

Проблемите на ефективността и оптималността от гледна точка на време и ресурси не са приоритет в настоящата версия на програмата. Настоящата разработка цели да демонстрира и да позволи различни тестове при изпълнението на задачата за автоматично извличане на съставни лексикални единици. Тя е пригодна за корпус с размери до 15 милиона думи и не е тествана върху ресурси с по-голям обем.

Напълно функционална разработка със статут на програма за автоматично извличане на съставни лексикални единици, както и автоматичното им класифициране, трябва да отчита проблемите на ефективността, баланса между време за изпълнение и количество използвана памет, както и да премине през процес на детайлно тестване. Настоящата програма има демонстрационен характер за дисертацията и към момента не отговаря на тези изисквания.

### 5.2.2. Предварителна обработка на текстовете

Предварителната обработка на текстовете се извършва еднократно и обхваща няколко отделни процеса, свързани с преобразуването на текста, докато се стигне до приетия формат, описан по-нататък.

Основният корпус, използван за целите на изследването, е корпусът с текстове от Уикипедия (вж. 1.3). Текстовете първоначално са в специфичен XML формат (WikiXML), получен при изнасяне на съответната страница от Уикипедия (Wikipedia, 2011, Специални:Изнасяне; вж. 1.3).

За целите на изследването текстовете се преобразуват в краен XML формат, при който всяка дума е маркирана отделно и се характеризира с набор от атрибути: графична дума, лема, част на речта и граматически характеристики, артибут за принадлежност към несвободна фраза.

Предварителната обработка на корпуса включва дейностите, изброени по-долу. Приложение Б представя примерен файл и различните видове обработка.

1. Извличане на чистия текст от файла в WikiXML формат, свален от Уикипедия, и преобразуването му в крайния XML формат. Това включва:
   - отстраняване на маркиращите тагове, при което се извлича информация за описанието на текста във файла;
   - отстраняване на маркировка за отделни части от текста от вида [ ] или { }; едновременно с отстраняването се извличат началните лексикални ресурси (вж. 5.2.3)
   - отстраняване или заменяне на специални символи (&, <, > и др.)
   - валидизация на формата.
2. Тагиране на текста с части на речта с помощта на тагер и лематизатор за български език, разработен от СКЛ (DCL-Tagger, 2011).
3. Чънкиране

   Направено е грубо чънкиране (англ. **chunking**) на всеки текст на ниво изречение и при него са отделени поредици от думи (чънкове), в рамките на които се появяват съставните лексикални единици. Граници на чънковете са такива елементи, за които приемаме, че не могат да влизат в съставните лексикални единици – това са пунктуационните знаци. Не е задължително чънкът да съответства на фраза, тъй като чънкирането не се основава на синтактичен анализ. Съществуват и съставни лексикални единици, които съдържат пунктуационни знаци (Пример 19(а)), но беше преценено, че те са много редки и основно свързани с наименования на организации. За сметка на това разделянето на чънкове улеснява работата и ускорява извличането на кандидати, тъй като се работи с по-малки единици.

Други съставни лексикални единици, които не се включват в анализа са тези, които съдържат съюзи – синтактичните структури с еднородно свързани части не са включени в синтактичния филтър (Пример 19(б)). Причината отново е, че те са сравнително рядко срещани и ограничени до категорията на наименованията, а включването им би довело до значително увеличаване на броя разглеждани кандитати, повечето от които дори не образуват фраза (Пример 19(в)), а с това се забавя процесът на анализ и се намалява точността. Направена беше проверка на това твърдение, като бяха извадени 100 произволно избрани изречения, съдържащи конструкции от вида $(A)\,N$ *и* $(A)\,N$, като само една от конструкциите беше част от съставна лексикална единица (*Професионална гимназия по [компютърни технологии и системи]*) и една беше статистически маркирана единица (*плодове и зеленчуци*), което показва, че само приблизително 1% от тези конструкции могат да бъдат полезни, а останалите 99% ще представляват проблем за анализа. Още по-редки са случаите, при които съюзът се появява в конструкции като $A$ *и* $A\,N$ (например *Етрополска [книжовно-просветна и калиграфско-художествена школа]*).

Не се включват в анализа и каскадни наименования по терминологията на Лесева и Стоянова (2008), като в Пример 19(г). При тях дескриптивната част от наименованието и същинското име са в отделни чънкове, тъй като кавичките се приемат за пунктуационни знаци, и двете части се разглеждат като отделни единици.

**Пример 19** (Чънкиране (//) и съставни лексикални единици)**.**

(а) *Напротив, // с влизането си в* ***Министерството на образованието,*** *// **младежта и науката** // (МОМН) // тя прояви такава откритост и склонност към диалог и чуваемост, // за каквато училищните и академичните съсловия бяха направо зажаднели след годините на провокирани сблъсъци и задкулисно лансирани предпочитания при нейния предшественик.*

(б) *3 средни заплати по изчисленията на НСИ ще получават членовете на* ***Националния съвет за радио и телевизия****, // определи парламентът вчера.*

(в) *Тези коне не могат да напускат стопанството до заминаването им за местоназначението извън контролираната зона и ваксинацията се вписва в паспорта...*

(г) ***Народната библиотека*** *// "**Иван Вазов**" // и Малък театър – // Пловдив пък ще представят // "Литературна колекция" // – // образователен цикъл по български език и литература, // аудиовариванти за XI и XII клас.*

(БНК © СКЛ)

Изборът на този формат беше обусловен от следните фактори:

- XML е широко използван формат за целите на компютърното представяне на

анотирани текстове.

- Форматът е удобен за работа и има редица библиотеки за обработване на XML с `Java`.
- Програмата `Chooser`, разработена от СКЛ (Коева, Ризов, и др., 2008), която може да бъде използвана за ръчно анотиране на различни езикови нива, е основана на този формат. Това осигурява възможност за ползване на `Chooser` за ръчно анотиране или визуализиране на резултатите от проведена работа върху корпуса. Също така се улеснява и използването на корпуса за други цели, например чрез добавяне на друг вид анотация.

### 5.2.3. Подготовка на лексикалните ресурси и на полуавтоматично анотиран корпус за оценка

Корпусът *Уики1000+* е подходящ за целите на изследването, защото се отличава с няколко основни характеристики, които имат пряко отношение към работата по проблемите на съставните лексикални единици:

- Текстовете в преобладаващата си част принадлежат към научно-популярния стил. Това гарантира добро представяне на съставни лексикални единици.
- Във формата WikiXML на статиите в Уикипедия голяма част от съставните лексикални единици са маркирани от човек – съставителя на статията. Това ни осигурява списък от несвободни фрази и частично анотиран корпус.

За да бъде ползван успешно този копрус, така че да позволява достоверна оценка на резултатите от автоматичното анотиране, беше необходимо да се обработи допълнително. Една от основните цели бе да бъдат маркирани възможно най-голяма част от съставните лексикални единици. Автоматично бяха анотирани и немаркираните примери на вече маркираните единици.

Обработката на корпуса е определена като полуавтоматична, тъй като лексикалните ресурси са обработени полуавтоматично и са ръчно верифицирани. Но самото тагиране на съставните лексикални единици с помощта на лексикалните ресурси е извършено с помощта на компютърна програма, която може да се прилага и за други цели (вж. Г).

Първоначалният списък от несвободни фрази, маркирани в *Уики1000+*, включва 25,672 единици. Той беше допълнително обогатен по два начина. Единият беше добавянето на 1,150 съставни лексикални единици, извлечени от **Български тълковен речник** (1994), които бяха ръчно верифицирани и класифицирани с оглед на това, дали са съставни лексикални единици или друга категория несвободни фрази.

Също така с помощта на честотния метод са извлечени допълнително 956 съставни лексикални единици, които не са маркирани в първоначалния корпус, но се появяват с висока честота. Методът, основан върху честотен анализ и филтриране на синтак-

тични конструкции (вж. 5.3.2), може да намери приложение като начин да бъдат подготвени и подобрени ресурсите за прилагане и тестване на други методи като например статистическите методи, представени в 5.3.3. Чрез този метод може да се намали брят на разглежданите синтактични конструкции и да се подобри качеството на ресурсите за оценка на резултатите (речник със съставни лексикални единици или ръчно тагирания корпус).

Стоянова (2009) представя приложението на честотния метод върху подкорпус на Браун корпус за български език (вж. 1.3), включващ 80 научни текста от различни области, наброяващи близо 160 хиляди думи.

За да се демонстрира ролята на синтактичния филтър за повишаването на резултатите, са сравнени десетте $N$-грама с най-висока честота преди (Таблица 5.1а) и след прилагането на филтъра (Таблица 5.1б). Както се вижда, повечето примери в първата таблица са биграми, които не образуват съставна единица или фрази, а са съчетания от служебни думи, които не се включват в изследването. Вижда се, че първите съчетания, които влизат в обекта на изследване, имат значително по-малка честота.

| **Честота** | $N$**-грам** |
|---|---|
| 767 | *да се* |
| 328 | *може да* |
| 174 | *не е* |
| 171 | *да бъде* |
| 168 | *трябва да* |
| 156 | *и в* |
| 145 | *и на* |
| 136 | *това е* |
| 123 | *себе си* |
| 115 | *за да* |

(а) преди филтър

| **Честота** | $N$**-грам** |
|---|---|
| 67 | *гражданско общество* |
| 45 | *гледна точка* |
| 44 | *химически обект* |
| 40 | *комплексно число* |
| 38 | *обществени отношения* |
| 37 | *друга страна* |
| 34 | *черно море* |
| 31 | *прост химически обект* |
| 28 | *български език* |
| 28 | *диалектическа логика* |

(б) след филтър

Таблица 5.1.: Десетте $N$-грама с най-висока честота в корпуса преди и след прилагане на синтактичен филтър (Стоянова, 2009).

Приложението на синтактичен филтър е подходящо за случаите, когато се изследва определен тип синтактични конструкции и тяхната форма е предварително известна. Безспорно е, че приложението на синтактичния филтър е от особена важност за подобряване на качеството на резултатите. Също така значително се намалява броят на разглежданите кандидати, което улеснява задачата, особено при изследване на големи по обем корпуси.

Броят на кандидатите се намалява с две основни техники – чънкиране и синтактичен филтър. Чънкирането се извършва на ниво предварителна обработка (вж. 5.2.2).

Втората техника за намаляване на броя кандидати за съставни лексикални единици е прилагането на синтактичен филтър. В настоящото изследване разглежданото явление е ограничено до фрази с опора съществително име, които могат да бъдат описани с малък брой синтактични конструкции (вж. 3.2.2). Още повече, само в редки случаи се допуска вмъкване на елементи между конституентите на фразата и вмъкваните части са ограничени до кратките притежателни местоимения. Това прави лесно прилагането на синтактичен филтър и значително намалява броя на разглежданите кандидати.

Първото приложение на честотния метод е като помощна техника за подобряване на качеството на ресурсите и автоматично извличане на съставни лексикални единици. Основният използван корпус е *Уики1000+* и наброява 13.4 милиона думи. Могат да се очертаят следните етапи:

1. **Извличане на всички кандидати**.

   От корпуса бяха извлечени всички синтактични конструкции, отговарящи на модела в 3.2.2 – 2,227,026 фрази кандидати за съставни лексикални единици.

2. **Анализ на синтактичните конструкции**.

   Разпределението на фразите според синтактичните конструкции е представено в таблица 5.2. От резултатите се вижда, че конструкциите от първите три типа покриват 94.6% от всички несвободни фрази. Този резултат ще ни позволи в 5.3.3 основателно да ограничим анализа си до тези най-чести единици, които са формирани от два пълнозначни елемента. По този начин ще ограничим броя на разглежданите кандидати – в 66.7% от кандидатите се съдържат 94.6% от несвободните фрази.

   От друга страна трябва да се обърне внимание на синтактичните категории, в които се съдържат най-голям процент несвободни фрази – за съчетание от вида $A\,N$ има вероятност приблизително $\frac{1}{3}$ да бъде несвободни фрази, следвано от $A\,Pro\,N$ с вероятност приблизително $\frac{1}{5}$. Тъй като на практика става въпрос за една и съща синтактична конструкция, в единия случай с

3. **Анализ на кандидатите с честота над определена граница**.

   Бяха разгледани кандидати с честота над 100 срещания, от които само око-

| **Конструкция** | **Фрази** | **Несв. фрази** | **% несв. спрямо фрази** | **% несв. спрямо всички несв.** |
|---|---|---|---|---|
| *A N* | 720031 | 228814 | 31.78 | 70.11 |
| *N N* | 309893 | 50332 | 16.24 | 15.42 |
| *N P N* | 455027 | 29595 | 6.50 | 9.07 |
| *A A N* | 78086 | 8083 | 10.35 | 2.48 |
| *A Pro N* | 15329 | 2979 | 19.43 | 0.91 |
| *N P A N* | 150442 | 2578 | 1.71 | 0.79 |
| *A N P N* | 118373 | 2331 | 1.97 | 0.71 |
| *A A A N* | 4649 | 556 | 11.96 | 0.17 |
| *A N N* | 36399 | 532 | 1.46 | 0.16 |
| *A N P A N* | 36009 | 296 | 0.82 | 0.09 |
| *A A N P N* | 10862 | 87 | 0.80 | 0.027 |
| *N Pro P N* | 7581 | 76 | 1.00 | 0.02 |
| *B N* | 143124 | 60 | 0.04 | 0.018 |
| *N Pro N* | 7070 | 14 | 0.20 | 0.004 |
| *N P N N* | 45040 | 7 | 0.02 | 0.002 |
| *N Pro P A N* | 2506 | 1 | 0.04 | 0.000 |

Таблица 5.2.: Разпределение на фразите по синтактични конструкции. Третата колона е процентното отношение на несвободните фрази от тази категория спрямо всички фрази от тази категория; четвъртата колона представя процентното участие на несвободните фрази от тази категория във всички несвободни фрази.

ло 3% бяха свободни фрази, а останалите 956 бяха определени като съставни лексикални единици и бяха допълнително добавени към списъка.

Приложението на честотния метод в настоящия анализ демонстрира неговите качества за допълване и обогатяване на съществуващи лексикални ресурси, с което се повишава тяхното качество. Методът може да се използва и за подпомагане на ръчното анотиране на корпуси с информация за съставните лексикални единици, като най-честите единици се маркират автоматично или полуавтоматично (верифицирани от експерт).

След двете добавяния списъкът наброява общо 27,778 единици, които се представят с основните форми на всичките си компоненти и част на речта. При това представяне основната форма често представлява граматически неправилна фраза, за разлика от основната форма на цялата единица. Представянето в тази форма обаче позволява намирането на съвпадения в текста чрез лесно и бързо сравнение между лемите и затова е предпочетена. Вероятността да се получи многозначност на тази основа не е голяма. Така също могат да се проследят варианти на съставни лексикални единици, при които се допуска изменяемост по форма на компонентите на несвободната фраза.

**Пример 20** (Представяне на съставни лексикални единици в речника.)**.**

мечка N стръвница N
мешан A скара N
министерски A кресло N
министерски A съвет N
министър N без P портфейл N
мирен A договор N
миров A скръб N
миши A дупка N

Втората задача при обработката на копруса е да бъдат тагирани съставните лексикални единици с техния тип според класификацията, представена в 3.2.3. За тази цел на всички 27,778 съставни лексикални единици от списъка беше полуавтоматично приписан тип от Класификация 6 (3.2.3, стр. 76).

В случай на многозначност поради липсата на контекст (в списъка от несвободни фрази) и при колебание относно принадлежността на единицата към определена група, е предпочитено по-общото значение. В повечето случаи става въпрос за случаи, при които единицата може да бъде определена както като прагматически немаркирана (общо название, дескриптор, Пример 21(а)), така и може да се срещне като прагматически маркирана единица, т. е. название (Пример 21(б)). Многозначността се проявява на абстрактно ниво и се разрешава в контекста. В някои случаи обаче се предпочита прагматически маркираното значение, тъй като е по-често и типично от общото (Пример 21(в) и (г)). За разлика от наименованията, дескрипторите имат повече свобода – могат да образуват например форми за множествено число, както се вижда от примера.

**Пример 21.**

(а) *За голяма жалост и нещастие, богатите географически, етнографически и филологически материали, които Славейков можа да събере за България, изгоряха в Стара Загора (дето той бил учител до освобождението на България) във време на катастрофата, която този град изтегли през последната* ***руско-турска война****.*

(б) *До освобождението на България, включително и в* ***Руско-турската война****, българите от всички български земи в по-малка или в по-голяма степен и размери бяха участвували в общ, единен фронт за своето национално освобождение.*

(в) *През следващите години този проект ще мобилизира усилията на ЕСФ и неговите органи, на* ***националните [съюзи на физиците]*** *както в Европа, така и в други страни на света.*

(г) *Подкрепиха ни и държавни организации като АЕЦ "Козлодуй", КИАЕМЦ,* ***Съюзът на физиците****, разбира се, даде своята подкрепа и участва много активно, редица частни фирми също ни подкрепиха.*

(БНК © СКЛ)

### 5.2.4. Третиране на конкурентни кандидати

При обработката на корпуса анализираме всички възможни фрази. Някои думи участват в няколко различни синтактични конструкции и следователно в няколко кандидата за несвободни фрази. Като **конкурентни** определяме тези конструкции, които не могат едновременно да бъдат отделни фрази. Макар и в настоящото изследване да не се отчитат проблемите на конкурентността, е необходимо те да бъдат регистрирани и взети предвид при анализа.

Пример 22 представя различни примери за конкурентни фрази кандидати. 22(а) и (б) представляват пример за съставна лексикална единица *световна война*, която е опора и хипероним в друга – *Втора световна война*. В този случай и двата кандидата са съставни лексикални единици и не са конкурентни. Конкурентни са трите варианта на *свършен вид на глагола* в Пример 22(в), (г) и (д), тъй като не са възможни едновременно. В текста се маркират и трите варианта, но трябва да се отстрани съществуващата многозначност и да се избере максималната фраза. Пример 22(е) и (ж) представят интересно явление, при което съчетанието *китайска стена* е придобило ново значение извън цялата единица *Велика [китайска стена]*, но е спорно доколко това е самостоятелна несвободна фраза освен в ограничен контекст и дали е например хипероним на наименованието *Велика [китайска стена]*. Пример 22(з) и (и) демонстрират случай на словосъчетание, което не е несвободна фраза извън по-голямото словосъчетание.

**Пример 22** (Конкурентни кандидати)**.**

*световна война*

(а) *Но [нова класическа* ***световна война]*** *— по форма, съдържание и методи на водене на бойните действия — не се разрази само защото от класическото продължение на политиката с военни средства нямаше никаква необходимост.*

(б) *В последно време съчинителите на реакционните американо-английски издания на историята на* ***Втората [световна война]*** *се стараят въобще да премълчат забележителните победи на съветската армия, опитвайки се с това да заличат от съзнанието на народите на Европа и Америка спомена за решаващата роля на Съветския съюз в разгрома на фашистка Германия.*

*[съвършен вид] на глагола, съвършен [вид на глагола], [съвършен вид на глагола]*

(в) *Глаголите от* ***съвършен вид*** *в сегашно и минало несъвършено време се срещат в подчинени изречения, свързани с главното съюзно или безсъюзно.*

(г) *В 5. клас това е и невъзможно – поради непознаването на някои от лингвистичните понятия, имащи отношение към изключенията: причастия, спрежение,* ***вид на глагола****.*

(д) *Дублев опростява метафората, представяйки я като краен и окончателен резултат (****съвършен вид на глагола****, минало съвършено време).*

*Велика [китайска стена]*

(е) *Той бил трезв човек и гледал на* ***Великата китайска стена****, която още тогава била архитектурен паметник, само като на фортификационно съоръжение.*

(ж) *A. R. C. имаме всички възможности да построим* ***китайска стена*** *между клиенти със сходен бизнес.*

*първи принцип на ...*

(з) *[Първият принцип] е некредитиране от Българска народна банка на държавата или държавните институции.*

(и) *Подчертава се, че докато* ***първият принцип на термодинамиката*** *е свързан с изменението на вътрешната енергия, то вторият принцип на термодинамиката може в най-общ вид да се разглежда като закон за нарастване на ентропията.*

(БНК © СКЛ)

В XML файла е отразена конкурентността на единиците (Пример 23).

**Пример 23** (Представяне на XML формат при конкурентни единици.)**.**

```
<word w="Старши" l="старши" sen="974792" pos="Apo" mwe="70545:0"
mwe_type="7" />
<word w="научен" l="научен" sen="974792" pos="Asmo" mwe="70544:0;70545:1"
mwe_type="7;7" />
<word w="сътрудник" l="сътрудник" sen="974792" pos="NCMsom" mwe="70544:1;70545:2"
mwe_type="7;7" />
<word w="II" l="ii" sen="974792" pos="M" />
<word w="степен" l="степен" sen="974792" pos="NCFsof" />
<word w="от" l="от" sen="974792" pos="R" />
<word w="декември" l="декември" sen="974792" pos="NCMNsom" />
<word w="1985" l="1985" sen="974792" pos="BCAbo" />
<word w="г" l="г" sen="974792" pos="B" />
<word w="." l="." sen="974792" pos="U" />
```

### 5.2.5. Методи за оценка на резултатите

За оценяване на резултатите от приложението на методите за разпознаване и тагиране на съставни лексикални единици е използван полуавтоматично анотираният корпус. Той е представен в 5.2.3. Съставните лексикални единици, наред с малък брой несвободни фрази от другите категории, са маркирани като съставни единици и е определена категорията, към която принадлежат (според Класификация 6).

При прилагането на методите за автоматично разпознаване на несвободни фрази и съставни лексикални единици всяка фраза кандидат се определя като свободна или несвободна по различните показатели на използвания метод. След това се извършва оценка на единицата. Ако в текста комбинацията е маркирана като съставна единица в своята цялост (да включва компоненти от точно една фраза и да обхваща всичките ѝ компоненти), тя се смята за вярна. Ако фразата е неразпозната или частично разпозната, се смята за грешка.

Двата основни количествени показателя, по които се оценява успешността на всеки метод, са **точността** (англ. **precision**) и **пълнотата** (англ. **recall**). Ако общият брой съставни лексикални единици в корпуса е $N_{\text{всички}}$, броят на всички разпознати е $N_{\text{всички\_разпознати}}$, а броят на верните е $N_{\text{верни}}$, точността $Pr$ и пълнотата $R$ се изчисляват по следните формули:

$$Pr = \frac{N_{\text{верни}}}{N_{\text{всички\_разпознати}}}$$

$$R = \frac{N_{\text{верни}}}{N_{\text{всички}}}.$$

Често даден метод е прилаган неколкократно върху корпуса с различни параметри. В тези случаи сравнителният анализ между двата варианта на метода ще се

базира на сравнение на точността и пълнотата. Много често емпирично се установяват оптималните стойности за определени параметри и става въпрос за балансиране между точността и пълнотата. Подходът, който ще бъде използван за комбинирана оценка на точността и пълнотата, ще бъде основан на следната формула

$$\text{Eval} = \alpha\, Pr + \beta\, R,$$

където $\alpha$ и $\beta$ са параметри, придаващи различна тежест на двата показателя. За оценка в текущото изследване ще бъдат използвани $\alpha = 1.5$ и $\beta = 1$, което дава по-голяма тежест на точността. Тази оценка ще бъде ползвана само за сравнителен анализ при приложение на различни методи или варианти на даден метод в рамките на настоящото изследване.

Оценката за работата на даден метод не може изцяло да се основава на количествените показатели, тъй като много често върху резултатността на метода оказват влияние фактори като вида и размера на изследвания корпус и други, а понякога и методите имат различна функция и се прилагат за различни цели. Затова количественият анализ ще бъде съпроводен с качествен анализ, който ще отчита основните особености на метода, резултатите от него и възможните сфери на приложение.

В случаите, когато се прилагат тренировъчни техники и техники за учене върху корпуса, усъвършенстваният метод трябва да бъде тестван върху различен корпус. По тази причина от основния корпус, който ще бъде използван като тренировъчен корпус, беше отделен малък корпус от 20 текста, който да служи като тестов корпус.

## 5.3. Експериментално приложение на методи за разпознаване и тагиране на съставни лексикални единици в българския език

При подбора на методите, които да бъдат експериментално приложени за автоматичното разпознаване и тагиране на съставни лексикални единици в българския език, бяха взети предвид следните фактори:

- **Простота на метода** – бяха подбрани методи, които не изискват сложни математически анализи и операции, така че да бъдат достъпни за по-широката научна общност – компютърни лингвисти, наред със специалисти в областта на теоретичното езикознание и др. Допълнително са положени усилия за представянето на методите по достъпен начин.
- **Разпространение на метода** – бяха предпочетени известни, утвърдени и широко използвани методи, които са били тествани за различни езици. Това ще даде възможност в бъдеще да се извършат и сравнителни анализи.
- **Обхватност на метода** – представени са методи с по-широко приложение за

сметка на по-екзотични методи или такива, използвани за специфични случаи – специфични явления, езици или тематични области.

- **Резултатност на метода** – логично бяха предпочетени методи, които дават по-добри резултати от тези с по-слаби. Представени са и някои сравнения между определен метод и неговата подобрена версия.
- **Наличие на нужните ресурси** – подбраните методи са съобразени и с наличните ресурси, в това число разнообразни текстови корпуси със сравнително голям обем, лингвистична анотация на корпусите, лексикални ресурси.
- **Предишен опит** – не на последно място изигра роля и наличието на предишен опит в използването на метода за български език.

Основният корпус, върху който са приложени методите, е *Уики1000+* (вж. 1.3). В някои случаи са използвани други корпуси – съобразно целта на демонстрацията, при представянето на предишни изследвания и др.

Изложените експерименти имат за цел да демонстрират отделните етапи от разпознаването на съставните лексикални единици и използваните методи, които дават резултати с определено качество в зависимост от сложността си и езиковата и неезиковата информация, която ползват. Резултатите имат значение от една страна за съпоставка между методите, когато това е възможно, но най-вече са анализирани възможностите за съчетаване на различни методи за по-точното определяне на съставните лексикални единици и техните подкатегории.

Разработването на цялостна, качествена и успешна методология за разпознаване на съставните лексикални единици като цяло е трудоемка и комплексна задача, изискваща съчетаване на различни методи – лингвисти1ни и статистически. Задачата за разпознаването не е решена и в световен мащаб и която изисква експериментиране с лексикални ресурси, различни по вид корпуси, лингвистични и статистически методи.

### 5.3.1. Основни принципи при провеждане на експериментите

Методите, описани тук, имат за цел разпознаването на несвободни фрази и съставни лексикални единици от *Уики1000+*, основавайки се на една от основните им особености – това, че са статистически маркирани, т. е. проявяват склонност да се появяват заедно в текста. Количествено тази особеност се проявява в две насоки – от една страна висока честота на поява, а от друга страна – честота, която говори за наличие на определена асоциация между думите, която може да се измери количествено.

Поради спецификата на корпуса концентрацията на неразложими и идиосинкретично разложими единици е много малко (1.25%) и поради това задачата за разграничаването на тези категории от съставните лексикални единици няма да бъде приоритет на тазо разработка.

В 4.2.1.4 бяха представени няколко различни мерки за асоциация между думите –

$\chi^2$, логаритмично подобие, взаимна информация и подобрена взаимна информация, които се използват и за целите на представения тук анализ.

Експериментите включват следните дейности:

- Изчисляване на количествените показатели (честота или мярка за асоциация) за всички изследвани кандидати.
- Графично представяне и анализ на разпределението.
- Определяне на параметри за метода, например долна граница $AM_{\min}$ за стойността на мярката при несвободните фрази.
- Определяне на критериите за причисляване на дадена единица като несвободни фрази.
- Оценяване на резултатите върху тренировъчния корпуса (в случай, че е ползван такъв) и върху независимия тестов корпус.
- Сравнителен анализ.

Приложението на метода включва провеждането на няколко експеримента, които са свързани с проверката на няколко предварително дефинирани интуитивни хипотези.

Проведени са следните експерименти:

1. Приложение на честотния метод върху *Уики1000+*.

   **Хипотеза 1.**
   Честотата на поява на фразите е добър критерий за разграничаване на несвободните фрази от свободните.

2. Приложение на метода с мярка $\chi^2$, логаритмично подобие, взаимна информация и подобрена взаимна информация върху корпуса от Уикипедия.

   С този експеримент са свързани следните хипотези.

   **Хипотеза 2.**
   Методите, използващи мярка за асоциация, са по-надеждни при откриване на несвободните фрази, отколкото честотния метод.

   **Хипотеза 3.**
   Мерките за асоциация са по-надеждни при единици с по-висока честота.

3. Приложение на комбиниран метод, който съчетава честотен анализ и мярка за асоциация.

   **Хипотеза 4.**
   Двата метода комбинирано ще дадат по-добри резултати.

   Основанията да очакваме по-добри резултати при комбиниране на двата метода

се коренят в направения честотен анализ, при който се вижда, че свободните съчетания се появяват преобладаващо с малка честота. По тази причина с относително голяма сигурност можем да причислим единиците с малка честота към свободните фрази и да използваме мерки за асоциация за разграничаване на единиците с по-голяма честота.

4. Приложение на комбиниран метод, който разграничава отделните категории несвободни фрази и съставни лексикални единици.

### 5.3.2. Метод, основан на честотен анализ и филтриране на синтактични конструкции

Освен за подобряване на лексикалните ресурси и анотацията на корпуса, методът може да се приложи като независим метод за разпознаване на несвободни фрази, като се разчита на това, че единиците с честота над определена граница ще бъдат в преобладаващата си част несвободни фрази и ще покриват значителна част от всички несвободни фрази.

Емпирично може да се установи каква е оптималната граница, като най-често става въпрос за установяване на баланс между точност и пълнота на резултатите (вж. 5.2.5). От Таблица 5.3, с изключение на малкото колебание при граница $N_{\min} = 100$ може да се каже, че с увеличаването на $N_{\min}$ се покачва прецизността, но за сметка на това се откриват по-малко на брой единици и намалява пълнотата.

Един от основните проблеми на статистическите методи е, че те дават проблематични резултати за единици с ниска честота на поява в корпуса. Затова обикновено при използването на статистически методи се избира определена долна граница за честотата ($N_{\min}$), която зависи както от големината на корпуса, така и от честотата на изследването явление.

Простият честотен метод с прилагане на синтактичен филтър не е подходящ за случаите, когато се цели възможно най-пълно извличане на съставните лексикални единици, тъй като при него се вземат само единиците с най-голяма честота. Също така методът не разграничава съставните лексикални единици от другите категории несвободни фрази, така че в случаите, когато се цели разглеждането на специфична категория несвободни фрази, той не е подходящ.

Също така има слабости приложението на метода върху малък по обем корпус, където единиците ще се появяват с относително ниска честота, както и за разнороден корпус, включващ разнообразни малки по обем тематично обособени подкорпуси, тъй като отново специализираната лексика ще бъде с малка честота.

Простият честотен метод е подходящ за автоматично съставяне на лексикални ресурси от сравнително големи специализирани корпуси, в които специализираната лексика ще е еднородна и ще се появява с голяма честота. Приложението на метода в

| **Долна граница за честотата** $N_{\mathbf{min}}$ | **Точност** $Pr$**, %** | **Пълнота** $R$**, %** | $\text{Eval} = 1.5 \times Pr + R$ |
|---|---|---|---|
| 25 | 76.32 | 73.84 | 188.32 |
| 50 | 96.69 | 63.48 | 208.52 |
| 75 | 97.12 | 49.34 | 195.02 |
| 100 | 96.99 | 40.72 | 186.21 |
| 200 | 97.25 | 25.33 | 171.21 |
| 300 | 97.79 | 18.57 | 165.26 |

Таблица 5.3.: Резултати от приложение на честотния метод с различни стойности за $N_{\min}$.

настоящия анализ демонстрира неговите качества за допълване и обогатяване на съществуващи лексикални ресурси, с което се повишава тяхното качество. Методът може да се използва и за подпомагане на ръчното анотиране на корпуси с информация за съставните лексикални единици, като най-честите единици се маркират автоматично или полуавтоматично (верифицирани от експерт).

Не на последно място положителна черта на този метод е неговата простота, както и малкото количество лингвистична информация, която се изисква за неговото приложение – единствено е необходимо тагиране с част на речта на думите в текста.

Добрите резултати, които методът дава, не са случайни. Честотните анализи върху свободните и несвободните фрази показват някои особености, представени в Таблица 5.4. Вижда се, че близо половината свободни фрази са с честота 1, докато от несвободните фрази това са съвсем малък процент.

| | **Средна стойност** | **Медиана** | **Макс. честота** | **% с честота 1** | **% с честота** $< 10$ |
|---|---|---|---|---|---|
| **Своб. фрази** | 6.5 | 2 | 1578 | 46.01 | 82.60 |
| **Несвоб. фрази** | 244.48 | 75 | 3206 | 1.84 | 13.54 |

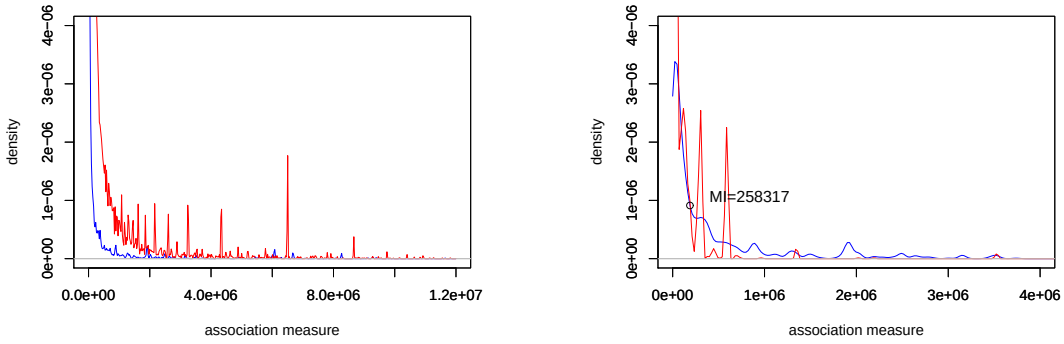
Таблица 5.4.: Честота на свободните и несвободните фрази.

Както ще се види и при сравнението с методите, използващи мерки за асоциация, това е много добър резултат, който мерките за асоциация не могат сами да постигнат.

Това потвърждава Хипотеза 1 и поставя под съмнение Хипотеза 2.

### 5.3.3. Хибриден метод, използващ мерки за асоциация

Фигура 5.1 показа разпределението на мярката за асоциация $\chi^2$. Ясно се вижда от двете графики, че $\chi^2$ не е надеждна мярка за разграничаване между свободни и несвободни фрази, тъй като не може да се намери добра граница, чрез която да се отделят.




(а) Всички фрази

(б) Фрази с честота над 100

Фигура 5.1.: Гъстота на разпределението на мярката $\chi^2$. Червено – свободни фрази; синьо – несвободни фрази.

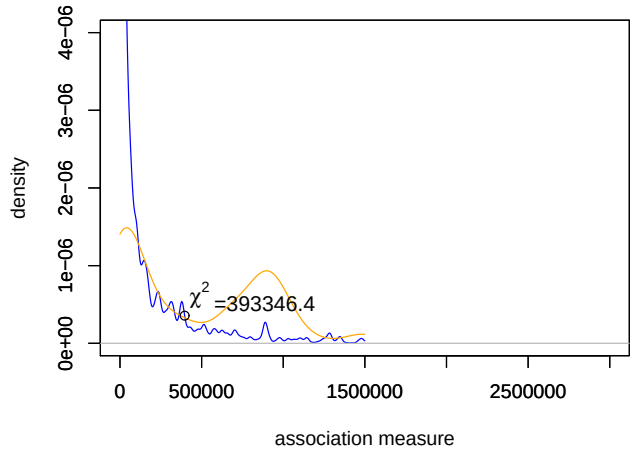
Резултатите от метода са представени в Таблица 5.5.

| $\chi^2_{\min}$ | **Разпознати** | **Точност** $Pr$, % | **Пълнота** $R$, % | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 51663.40 | 147069.00 | 57.56 | 47.58 | 133.92 |
| 103326.80 | 114057.00 | 59.61 | 36.90 | 126.32 |
| 154990.20 | 95821.00 | 60.41 | 31.00 | 121.62 |
| 206653.60 | 85522.00 | 61.53 | 27.67 | 119.97 |
| 258317.00 | 75870.00 | 61.56 | 24.55 | 116.89 |
| 309980.40 | 69525.00 | 62.06 | 22.49 | 115.58 |
| 361643.80 | 63068.00 | 61.74 | 20.40 | 113.01 |
| 413307.20 | 56512.00 | 60.86 | 18.28 | 109.57 |
| 464970.60 | 53424.00 | 61.16 | 17.28 | 109.02 |
| 516634.00 | 49786.00 | 60.79 | 16.11 | 107.30 |
| 568297.40 | 48137.00 | 61.40 | 15.57 | 107.67 |
| 619960.80 | 45592.00 | 61.49 | 14.75 | 106.99 |
| 671624.20 | 42863.00 | 61.05 | 13.87 | 105.45 |

| $\chi^2_{\min}$ | **Разпознати** | **Точност** $Pr$, % | **Пълнота** $R$, % | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 723287.60 | 40390.00 | 61.00 | 13.07 | 104.57 |
| 774951.00 | 39026.00 | 61.14 | 12.63 | 104.34 |
| 826614.40 | 37957.00 | 61.51 | 12.28 | 104.55 |
| 878277.80 | 36609.00 | 61.43 | 11.84 | 103.99 |
| 929941.20 | 33438.00 | 60.28 | 10.82 | 101.24 |
| 981604.60 | 32134.00 | 59.78 | 10.40 | 100.07 |

Таблица 5.5.: Резултати от разпознаването на несвободни фрази чрез метод с $\chi^2$ мярка за асоциация. Общият брой несвободни фрази е 309,084.

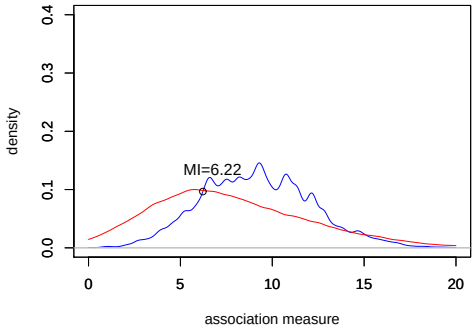
$\chi^2$ обаче се оказва сравнително добра мярка за разграничаване на неразложимите и идиосинкретично разложимите несвободни фрази от съставните лексикални единици с гранична мярка, близка до $\chi^2_{\min} = 393346.4$ – точката, от която нататък преобраладават съставни лексикални единици пред неразложимите и идиосинкретично разложимите. Точната оптимална стойност обаче може да се определи емпирично. Двете категории имат пикове на концентрацията си в различни интервали. На тази основа можем да определим единиците с мярка $\chi^2 < \chi^2_{\min}$ като съставни лексикални единици, а останалите като неразложими или идиосинкретично разложими.



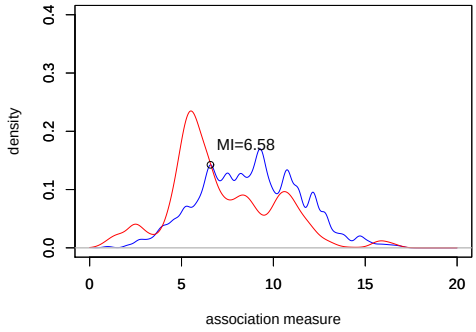

Фигура 5.2.: Гъстота на разпределението на мярката $\chi^2$. Синьо – съставни лексикални единици; оранжево – неразложими и идиосинкретично разложими несвободни фрази.

Резултатите от приложението на метода, използващ мярката за взаимна информация за измерване на асоциацията между думите, са представени на фигура 5.3. Оказва се, че взаимната информация не може да бъде успешно използвана за разграничаване на несвободните фрази от свободните, тъй като макар и да има ясно

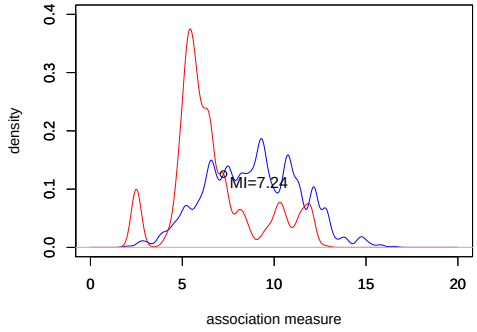
различими пикове на честотата при двете категории, попадащи в различни интервали, двете криви се движдат близо една до друга, което ще се изразява във висок брой грешки – свободни фрази, определени като несвободни, което е причина за ниската точност на метода (Фигура 5.3а и Таблица 5.6). Много близки до тези са и резултатите от прилагането на подобрената мярка за взаима информация, затова те няма да бъдат представени.




(а) Всички фрази




(б) Фрази с честота над 50




(в) Фрази с честота над 100

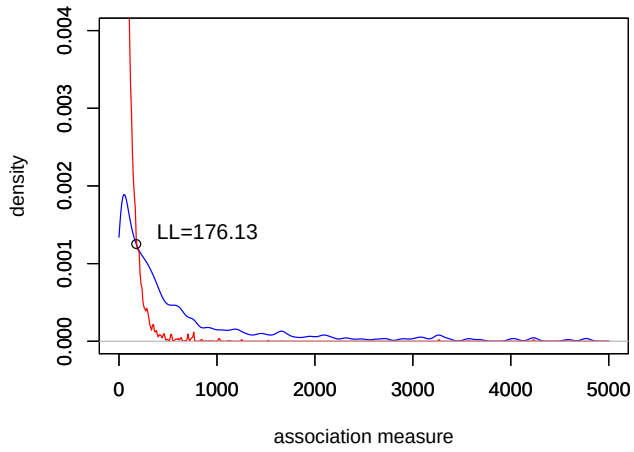
Фигура 5.3.: Гъстота на разпределението на мярката за взаимна информация. Червено – свободни фрази; синьо – несвободни фрази.

| $LL_{\min}$ | **Разпознати** | **Точност** $Pr$**, %** | **Пълнота** $R$**, %** | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 5.8 | 270169 | 26.14 | 87.41 | 126.62 |
| 6 | 264932 | 26.38 | 85.72 | 125.29 |
| 6.2 | 260562 | 26.69 | 84.3 | 124.34 |
| 6.4 | 253518 | 26.8 | 82.02 | 122.22 |
| 6.6 | 245344 | 26.84 | 79.38 | 119.64 |
| 6.8 | 238683 | 26.98 | 77.22 | 117.69 |
| 7 | 232297 | 27.17 | 75.16 | 115.92 |
| 7.2 | 225134 | 27.29 | 72.84 | 113.78 |

| $LL_{\min}$ | **Разпознати** | **Точност** $Pr$, % | **Пълнота** $R$, % | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 7.4 | 219015 | 27.48 | 70.86 | 112.08 |
| 7.6 | 210122 | 27.41 | 67.98 | 109.10 |
| 7.8 | 203338 | 27.55 | 65.79 | 107.12 |
| 8 | 198087 | 27.82 | 64.09 | 105.82 |
| 8.2 | 188875 | 27.69 | 61.11 | 102.65 |
| 8.4 | 181584 | 27.74 | 58.75 | 100.36 |

Таблица 5.6.: Резултати от разпознаването на несвободни фрази чрез метод с взаимна информация. Общият брой несвободни фрази е 309,084.

По-добри резултати дава методът, използващ лигаритмичното подобие като мярка за асоциация между думите. Резултатите са представени на Фигура 5.4 и Таблица 5.7.




Фигура 5.4.: Гъстота на разпределението на мярката за логаритмично подобие. Синьо – съставни лексикални единици; червено – свободни фрази.

| $LL_{\min}$ | **Разпознати** | **Точност** $Pr$, % | **Пълнота** $R$, % | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 35.23 | 274950 | 50.31 | 88.96 | 164.43 |
| 70.45 | 252998 | 58.34 | 81.85 | 169.36 |
| 105.68 | 234580 | 62.84 | 75.9 | 170.16 |
| 140.9 | 220278 | 65.54 | 71.27 | 169.58 |
| 176.13 | 205750 | 66.77 | 66.57 | 166.73 |
| 211.35 | 193528 | 67.35 | 62.61 | 163.64 |

| $LL_{\min}$ | **Разпознати** | **Точност** $Pr$**, %** | **Пълнота** $R$**, %** | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|
| 246.58 | 182206 | 67.32 | 58.95 | 159.93 |
| 281.8 | 169810 | 66.5 | 54.94 | 154.69 |
| 317.03 | 159759 | 65.82 | 51.69 | 150.42 |
| 352.25 | 149827 | 64.75 | 48.47 | 145.60 |
| 387.48 | 141383 | 63.69 | 45.74 | 141.28 |
| 422.7 | 134114 | 62.62 | 43.39 | 137.32 |
| 457.93 | 127947 | 61.68 | 41.4 | 133.92 |
| 493.15 | 123178 | 60.85 | 39.85 | 131.13 |
| 528.38 | 118672 | 59.98 | 38.39 | 128.36 |
| 563.6 | 113162 | 58.96 | 36.61 | 125.05 |
| 598.83 | 108763 | 58.04 | 35.19 | 122.25 |
| 634.05 | 102850 | 56.77 | 33.28 | 118.44 |
| 669.28 | 99769 | 56.04 | 32.28 | 116.34 |

Таблица 5.7.: Резултати от разпознаването на несвободни фрази чрез метод с мярка за асоциация логаритмично подобие. Общият брой несвободни фрази е 309,084.

Както се вижда, тестовете със мерки за асоциация не успяват да надминат простия статистически метод. Това отхвърля Хипотеза 2.

За да бъдат подобрени резултатите от тези статистически методи, беше приложена комбинирана техника, която отчита както честотата, така и мярката за асоциация. При този експеримент беше предпочетена мярката за логаритмично подобие, тъй като самостоятелно тя дава най-добри резултати от трите представени мерки.

Комбинираният метод изисква три параметъра:

- Долна граница за честотата $F_{\min}$ – всички единици с честота, по-малка от $F_{\min}$ се определят като свободни фрази.
- Горна граница за честотата $F_{\max}$ – всички единици с честота, по-голяма от $F_{\max}$ се определят като несвободни фрази.
- Граница за мярката за логаритмично подобие $LL_{\min}$ – всички единици с честота, по-голяма от $F_{\max}$ се определят като несвободни фрази.

Експериментът показа, че резултатите от този метод са сходни с тези при простия честотен метод (Таблица 5.8). Това частично потвърждава Хипотеза 4 – комбинираният метод отбелязва подобрение спрямо методът, ползващ само мярка за асоциация,

но отбелязва същите резултати, като честотния метод.

Положителна страна на комбинирания метод обаче е, че предлага повече възможности за вариации в съотношението между точността и пълнотата чрез избиране на различни параметри, като се запазва относително висока стойност на Eval. Резултатите потвърждават Хипотеза 3, която твърди, че резултатите са по-надеждни при единици с по-висока честота. Показателен е фактът, че когато мярката за асоциация се анализира само при единици с по-висока честота от определена стойност, резултатите се подобряват спрямо тези, при които мерките се прилагат върху всички фрази.

| $F_{\min}$ | $F_{\max}$ | $LL_{\min}$ | **Разпознати** | **Точност** $Pr$, % | **Пълнота** $R$, % | Eval $= 1.5 \times Pr + R$ |
|---|---|---|---|---|---|---|
| 30 | 2000 | 10 | 224646 | 80.4 | 72.68 | 193.28 |
| 35 | 2000 | 10 | 216900 | 84.87 | 70.18 | 197.47 |
| 40 | 2000 | 10 | 209946 | 89.17 | 67.93 | 201.68 |
| 45 | 2000 | 12 | 204468 | 92.39 | 66.15 | 204.74 |
| 50 | 2000 | 5 | 199302 | 96.04 | 64.48 | 208.55 |
| 55 | 2000 | 8 | 188525 | 96.92 | 60.99 | 206.37 |
| 60 | 2000 | 6 | 178954 | 97.06 | 57.9 | 203.49 |

Таблица 5.8.: Някои избрани резултати от разпознаването на несвободни фрази чрез комбиниран метод с честотен анализ и мярка за асоциация логаритмично подобие. Общият брой несвободни фрази е 309,084.

### 5.3.4. Приложение на лингвистична информация за разпознаване на отделните категории съставни лексикални единици

Количествените методи, представени по-горе, имат за цел разпознаването на несвободни фрази сред всички фрази кандидати, отговарящи на определена синтактична структура. Поради спецификата на корпуса преобладаващата част от определените несвободни фрази са съставни лексикални единици (неразложимите и идиосинкретично разложимите обхващат само 1.25%). Разпределението на фразите по категории е представено в Таблица 5.9.

| **Категория** | **Означение** | **Брой срещания** | **% от несв. фрази** |
|---|---|---|---|
| Неразложими | A | 700 | 0.23 |
| Идиосинкр. разложими | B | 3156 | 1.02 |
| Категория СЛЕ | 1 | 36932 | 11.95 |
| Категория СЛЕ | 2 | 11248 | 3.64 |
| Категория СЛЕ | 3 | 1461 | 0.47 |
| Категория СЛЕ | 4 | 1086 | 0.35 |
| Категория СЛЕ | 5 | 18962 | 6.13 |
| Категория СЛЕ | 6 | 27373 | 8.86 |
| Категория СЛЕ | 7 | 140394 | 45.42 |
| Категория СЛЕ | 8 | 16653 | 5.39 |
| Категория СЛЕ | 9 | 1468 | 0.47 |
| Свободни колокации | X | 49651 | 16.06 |
| Свободни фрази | Y | 1197762 | – |

Таблица 5.9.: Разпределение на фразите в корпуса по категории.

Изследването на Стоянова (2009) представя метод, използващ лингвистична информация – лексикални ресурси и синтактични модели, за разпознаване наименования на организации. В статията за третирането на наименованията с оглед на машинния превод на (Лесева и Стоянова, 2008) е предложено съчетаването на две основни техники за разпознаване на описателните наименования: речници на дескрипторите, характеризиращи различните семантични категории наименования, и синтактичен филтър във вид на регулярен автомат. В статията методът е приложен с помощта на програмата за лингвистичен анализ `NooJ` (NooJ, 2002).

Стоянова (2009) ползва речник от дескриптори, който включва 52 думи и техните форми, които най-общо са хипоними на думата *организация*. Те са извлечени от Б. WordNet (**Българският WordNet**), **Български тълковен речник** (1994) и други източници.

Зададени са допустими семантично-синтактични конструкции, като са посочени и проблемните категории и случаи, при които съществува многозначност. За структурата на съставните части на наименованията за организации могат да бъдат въведени

допълнителни семантични и синтактични ограничения: възможни предлози в предложната фраза (*на*, *за*, *по*), възможни определения към дескриптора (за място, за функция), ограничения във формите. Задачата за разпознаване на дескриптори и наименования на организации може да се сведе до разграничаване на следните случаи:

- Дескриптивни наименования, съдържащи дескриптор, които отговарят на зададената синтактична структура – влизат в категории 4, 5, 8 и 10;

  *Евробългарския културен център*

- Дескриптивни изрази, които не са наименования; съдържащи дескриптор и отговарящи на зададената синтактична структура – влизат в категории 6, 7 и 9;

- Дескриптивни наименования, съдържащи дескриптор, които не отговарят на зададената синтактична структура – не се включват в анализа

  *Министерство на образованието и науката*

- Дескриптивни наименования, несъдържащи дескриптор, които отговарят на зададената синтактична структура

  *Граждани за европейско развитие на България*

- Свободни фрази, съдържащи дескрипторна дума и отговарящи на дефинираната структура

  *В него Дан Лундгрен предупреждава, че ако не бъдат взети мерки, много скоро Калифорния ще се превърне в новия център на руската мафия.*

  (БНК © СКЛ)

- Съчетания, които не са фрази, но са линейна последователност, която отговаря на дефинираната структура

  *Михаил Чигир заяви вчера, че ще сдаде поста, след като не се състоя специалното заседание на Върховния и министерския съвет за изострената ситуация в страната и поисканата от мнозинството депутати оставка на президента.*

  (БНК © СКЛ)

Основно условие за качеството на резултатите от този метод е изчерпателността на използваните лингвистични ресурси и точността на дефинираните модели. За добре описаните семантични (по-скоро тематични) категории съставни лексикални единици разпознаваемостта е много висока, както и точността. От друга страна поради големия брой на категориите и разнообразието на възможни конструкции, пълното описание на всички от тях е непосилна задача, а също така по-изчерпателното описание на категориите ще доведе до свъхгенериране (разпознаване).

Основните резултати от проведения експеримент (Стоянова, 2009) показват, че значителен процент от разпознатите по гореописаните правила фрази са несвободни

фрази (76.18%), като в това число се включват както съставни наименования (64.8% от разпознатите), така и съставни дескриптори (11.38% от разпознатите).

По подобен начин може да се формулира система от правила, с помощта на които да се разпознават отделните категории съставни лексикални единици. Класификацията се основава на прагматичната маркираност на единицата (дали е наименование),на наличието на референции към външни обекти (има компонент наименование) и на композиционалността.

Следният списък с правила беше съставен и тестван върху част от корпуса от Уикипедия:

1. Ако се състои само от думи, определени като съществителни собствени имена, единицата е същинско наименование (категория 1).

   Тагерът (DCL-Tagger, 2011) приписва информация също и за вида на съществителното – дали е нарицателно или собствено.

2. Единица, включваща съществително собствено и други думи, на която всички думи започват с главна буква – определения към съществително собствено, това е категория 2.

3. В единица, включваща съществително собствено, и други думи, на която първата дума започва с главна буква – с най-голяма вероятност е категория 4, 5 или 8.

4. Единица, включваща съществително собствено, която не започва с главна буква – с най-голяма вероятност е категория 6.

5. Единица, която не съдържа съществително собствено, но започва с главна буква – с най-голяма вероятност е категория 8 или 1.

6. Единица, която не съдържа съществително собствено и която не започва с главна буква – с най-голяма вероятност е категория 7 или 9.

# 6. Резултати, изводи и насоки за бъдеща работа

## 6.1. Анализ на резултатите

Настоящата разработка разглежда проблемите на съставните лексикални единици в българския език с оглед на тяхното автоматично разпознаване и тагиране като част от процеса на лингвистичен анализ и анотация на български текстове. Основната цел на разработката е свързана с изграждането на теоретичен модел на изследваното явление и прилагането на този модел в практическата работа по автоматичното разпознаване на съставните лексикални единици.

Бяха поставени за решаване следните научни задачи:

1. Да бъде дефиниран обхватът на понятието за съставна лексикална единица, както и да бъдат анализирани граничните случаи, които представляват проблем за автоматичното идентифициране на съставните лексикални единици.
2. Да бъде изграден задълбочен теоретичен модел на съставните лексикални единици, който да описва техните характеристики на различни езикови нива – морфо-синтактични, семантични, синтактични, прагматични.
3. Да се изработи детайлна класификация на категорията на съставните лексикални единици, която да има практическо приложение за тяхното идентифициране, описание и анализ.
4. Да бъде разработена методология за разпознаване на съставните лексикални единици и отделните им категории.
5. Методологията да бъде приложена в практиката, като бъдат анализирани резултатите, възможностите и слабостите ѝ.

В Глава 2 са представени несвободните фрази, като се определя и мястото на съставните лексикални единици в лексикалната система на езика. Анализът се опира на традиционните виждания за съставните лексикални единици, но надгражда над това в посока към изясняване на специфичните им особености. В резултат е предложена дефиниция за понятието **съставна лексикална единица** (Дефиниция 3, стр. 37), която обособява явлението като отделна категория в рамките на несвободните фрази.

Следва и практическа характеристика на съставните лексикални единици (4, стр.

38), която позволява идентифицирането на явлението в корпуси от текстове чрез прилагане на различни тестове за проверка на условията за композиционалност на значението, регулярност и (не)рестриктивност, семантична маркираност и институционализираност на фразата. Изграждането на практически приложима дефиниция за понятието е стъпка напред към изграждането на методология за тяхното автоматично разпознаване. (Задача 1)

Проучени са и отношенията с други езикови явления, с които традиционно се свързва анализът на несвободните фрази и съставните лексикални единици в частност. Това са сложните думи, колокациите, наименованията и терминологията, които проявяват много общи черти с обекта на изследването.

В Глава 3 е представен подробен модел за лингвистичното описание на съставните лексикални единици именни фрази. В модела е включена комплексна информация на различни нива – формообразувателна, семантична, синтактична, прагматична. Особено внимание е отделено на анализа на идиоматичността като основна отличителна характеристика на съставните лексикални единици, която се проявява в различни аспекти и може да варира в широк диапазон. Дискутира се и въпросът за това, дали е необходимо изчерпателното описание на съставните лексикални единици в речниците, и са изтъкнати аргументи, подкрепящи виждането, че тяхното добавяне не е необходимо. (Задача 2)

Представени са и различни класификации на съставните лексикални единици по семантични и синтактични признаци. Изложена е и класификация според идиоматичността, която е определена като приложна класификация, тъй като отразява различията в подходите към третирането на отделните категории при обработката на езика. (Задача 3)

Глава 4 описва основните методи, които се използват за решаване на конкретните задачи по автоматичното разпознаване на съставни лексикални единици. Представени са няколко основни лингвистични, количествени (статистически) и хибридни метода, като се изтъква, че третата категория дават най-добри резултати и намират най-широко приложение.

В тази глава се анализират и факторите, които оказват влияние върху успешното прилагане на методите и качеството на резултатите. (Задача 4)

Глава 5 описва практическото приложение на методите за автоматично разпознаване и тагиране на съставните лексикални единици. Преди това е направен кратък преглед на компютърни програми, които включват функционалности за разпознаване и определяне на несвободни фрази. След този преглед е представена новоразработената за целите на дисертацията компютърна програма `bgMWE`, която включва набор от инструменти за обработка на текстови корпуси, прилагане на различни методи за разпознаване и обработка на съставни лексикални единици, за анализ и оценка на резултатите.

Очертани са основните етапи на тази комплексна дейност и са изложени аргументи за всеки метод, предложен за използване в определен етап, като се набляга на възможностите, които методът предлага, и пригодността му за разграничаване на единиците по определени характеристики. Представената серия от експерименти очертава възможна методологична рамка за разпознаване на съставните лексикални единици, която може да бъде развивана и усъвършенствана. (Задача 5)

Основният извод, който може да бъде направен от настоящата дисертация, е че разпознаването и тагирането на съставните лексикални единици в български (и други) езици е сложна и комплексна задача, изискваща комбинирането на лингвистични и статистически методи за нейното решаване. Нуждата от разработване на ефективна методология за тази цел е обусловена от все по-широкото навлизане на компютърни системи за машинен превод и автоматично извличане на информация, резултатите от които могат значително да се подобрят след решаването на тези проблеми.

## 6.2. Приноси на разработката

Приносите на изследването могат да бъдат обобщени до следното:

1. Разработката се занимава със съставните лексикални единици в българския език, което е слабо описано и изследвано явление както в теоретичната, така и в компютърната лингвистика. Явлението се разглежда за първи път самостоятелно, отделно от другите категории несвободни фрази. Изследването е важно, защото съставните лексикални единици са широко представени в езика и поставят редица проблеми пред компютърната обработка на българския език, най-вече с оглед на разработването на приложения за автоматичен превод, автоматичното извличане на информация и др.
2. Представен е задълбочен теоретичен анализ на съставните лексикални единици и техните основни характеристики на различни езикови равнища – семантично, синтактично, прагматично. Разработена е и терминологична система, която се съобразява и с другите български изследвания в областта.
3. Изработена е и класификация на съставните лексикални единици на базата на тяхната идиоматичност. Класификацията има подчертано приложен характер, защото се основава на степента и характера на тяхната композиционалност и институционализираност, което пък определя начина, по който е формирано значението им, и съответно – начина, по който да бъдат обработени в различни компютърни приложения за анализ.
4. Предложена е методология за разпознаване на съставни лексикални единици на етапи – несвободни фрази, съставни лексикални единици, отделни категории според представените класификации. Демонстрирани са възможностите за приложението на методологията и са отчетени резултатите. Методологията се

отличава с относителна простота, което я прави достъпна за по-широк кръг от специалисти.

5. Допълнителен принос е разработването на система от компютърни модули, които могат да бъдат приложени за различни задачи, свързани с обработката на езика, разпознаването на съставните лексикални единици, автоматичното извличане на ресурси и др.

## 6.3. Някои насоки и идеи за бъдеща работа

Поради мащабността на темата за цялостното ѝ разработване необходими много задълбочени изследвания. Остават редица проблеми, свързани с автоматичната обработка на езика, за които ще се търсят адекватни решения в бъдеще. Както неведнъж беше подчертано, това е проблематична област за компютърната лингвистика като цяло, което прави темата особено актуална.

На първо място е необходимо да бъде обърнато внимание на възможностите за разработване на допълнителни методи, които комбинират статистически тестове и лингвистичен анализ. Особено трябва да се наблегне на анализ, който не се базира на конкретни ресурси (речници), а проявява повече гъвкавост и универсалност, например разпознава единици с определена синтактична структура, маркирани с определени семантични признаци. За тази цел трябва да се направи подробен анализ на синтактичните конструкции на отделните семантични категории съставни лексикални единици, показателни семантични и други маркери в тези конструкции, да се отчетат също така синтактичните ограничения за това, дали и какви фрази могат да се вмъкват между елементите на дадена съставна лексикална единица.

По повод на приложението на семантичен анализ в процеса на разпознаване на съставните лексикални единици, може да се използват големите възможности, които предлагат лексикално-семантични мрежи като WordNet. Друго важно разширение на анализа е включването на контекста, тъй като в много случаи значението на съставните лексикални единици се определя от контекста.

Специално внимание трябва да се обърне на обективността на методите. Без да се засяга валдността на направените в разработката изводи, в него идиоматичността се оценява субективно (от автора) – например доколко значението на дадена единица е композирано и разложимо, което пък предопределя начина, по който единицата ще бъде класифицирана. Възприемането на идиоматичността е социолингвистично обусловено – доколко композираното значение се пази и степента, до която е придобило абстрактно значение, както и това, доколко единицата е институционализирана, което определя дали допуска или не заместване със синоними и други. По тази причина оценката на идиоматичността от човек е оправдано, но повдига въпроса за достоверността и валидността на преценка, направена от един човек.

В бъдеще е необходимо да се потърсят адекватни решения на тези проблеми, като се използва българският и чуждестранният опит в подобни изследвания.

Задача за бъдеща работа остава и комбинирането на отделните `Java` модули в цялостна система, която да предлага възможности за анализ на текстове на различно ниво, извличането на различни ресурси и други, както и разработването на добър графичен интерфейс.

# Библиография


CollocationExtract (2000): **Collocation Extract**. URL: `http://pioneer.chula.ac.th/~awirote/colloc/`. Dept. of Linguistics, Chulalongkorn University. 2000.

DCL-Tagger (2011): **DCL Tagger and Lemmatiser**. URL: `http://dcl.bas.bg/services/`. Секция по компютърна лингвистика, ИБЕ – БАН. 2011.

NooJ (2002): **NooJ**. URL: `http://www.nooj4nlp.net/`. M. Silberztein. 2002.

TermeX (2009): **TermeX**. URL: `http://ktlab.fer.hr/termex/`. Faculty of Electrical Engineering и Computing, University of Zagreb. 2009.

Wikipedia (2011): **Уикипедия**. URL: `http://bg.wikipedia.org`. март 2011.

WordNet: **WordNet: An Electronic Lexical Database**. URL: `http://wordnet.princeton.edu/`. Princeton University.

WordNet, Български: **Българският WordNet**. URL: `http://dcl.bas.bg/BulNet/`. Секция по компютърна лингвистика, ИБЕ – БАН.

Банард, С., Болдуин, Т., Лакарид, С. (2003): Bannard, C., T. Baldwin, A. Lascarides. "A Statistical Approach to the Semantics of Verb-Particles". English. In: **Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment**. 2003, pp. 65–72.

Барони, М., Еверт, С. (2007): Baroni, M., S. Evert. **Collocation Extraction with Statistical Association Measures**. English. 2007.

Бекавац, Б., Тадич, М. (2008): Bekavac, B., M. Tadić. "A Generic Method for Multi Word Extraction from Wikipedia". English. In: **Proceedings of the ITI 2008 30 International Conference on Information Technology Interfaces**. Croatia, 2008, pp. 663–667.

Болдуин, Т. (2004): Baldwin, Timothy. **Multiword Expressions, Advanced course at the Australasian Language Technology Summer School (ALTSS)**. English. 2004.

Болдуин, Т. (2006): Baldwin, Timothy. **Compositionality and Multiword Expressions: Six of One, Half a Dozen of the Other?, COLING/ACL 2006 Workshop on MWEs**. English. 2006.

Болдуин, Т., Банард, К., Танака, Т., Уидоус, Д. (2003): Baldwin, T., C. Bannard, T. Tanaka, D. Widdows. “An Empirical Model of Multiword Expression Decomposability”. English. In: **Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment**. 2003.

Бонин, Ф., Дел’ Орлета, Ф., Вентури, Г., Монтемани, С. (2010): Bonin, F., F. Dell’ Orletta, G. Venturi, S. Montemagni. “Contrastive Filtering of Domain-Specific Multi-Word Terms from Different Types of Corpora”. English. In: **Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010)**. Beijing, 2010, 77—80.

Бояджиев, Т., Куцаров, И., Пенчев, Й. (1998): Бояджиев, Т., И. Куцаров, Й. Пенчев. **Съвременен български език: Фонетика, лексикология, словообразуване, морфология, синтаксис**. Изд. къща "Петър Берон", 1998.

Брунщайн, А. (2002): Brunstein, A. **Annotation guidelines for answer types**. English. `http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html`. 2002.

Винче, В., Наги, И. Т., Беренд, Г. (2011): Vincze, V, I. T. Nagy, G. Berend. “Multiword Expressions and Named Entities in the Wiki50 Corpus”. English. In: **Proceedings of Recent Advances in Natural Language Processing**. Hissar, Bulgaria, 2011, pp. 289–295.

Вътдев, Д. (2003a): Вътдев, Д. **Записки по приложна статистика 1**. `http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/lectures/applstat1.pdf`. 2003.

Вътдев, Д. (2003b): Вътдев, Д. **Записки по приложна статистика 2**. `http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/lectures/applstat2.pdf`. 2003.

Гизбрехт, Е. (2009): Giesbrecht, E. “In Search of Semantic Compositionality in Vector Spaces”. English. In: **ICCS 2009, LNAI**. Ed. by S. Rudolph, F. Dau, F. O. Kuznetsov. Vol. 5662. Springer-Verlag Berlin Heidelberg, 2009, 173–184.

Гиржу, Р., Наков, Пр., Настасе, В., Цапакович, С., Търни, П., Юрет, Д. (2007): Girju, R., P. Nakov, V. Nastase, S. Tzpakowicz, P. Turney, D. Yuret. “Classification of Semantic Relations between Nominals”. English. In: **SemEval-2007, Task 4**. 2007.

БАН, Граматика на (1983): Стоянов, С., съст. **Граматика на съвременния български книжовен език**. Т. 1-3. Издателство на БАН, 1983.

Грийншоу, Дж. (2005): Grimshaw, J. **Words and Structure**. English. Stanford, US: CSLI Publications, 2005.

Грос, М. (1986): Gross, M. “Lexicon - grammar: the representation of compound words”. English. In: **Proceedings of the 11th coference on Computational linguistics**.

COLING '86. Bonn, Germany: Association for Computational Linguistics, 1986, pp. 1–6. DOI: http://dx.doi.org/10.3115/991365.991367.

Делач, Д., Крлежа, З., Шнайдер, Я., Башич, Б., Шарич, Ф. (2009): Delač, D., Z. Krleža, J. Šnajder, B. Bašić, F. Šarić. "TermeX : A Tool for Collocation Extraction". English. In: **CICLing 2009**. Ed. by A. Gelbukh. Vol. 5449. LNCS. Springer-Verlag Berlin Heidelberg, 2009, pp. 149–157.

Джакендоф, Р. (1995): Jackendoff, R. "The boundaries of the lexicon". English. In: **Idioms: Structural and Psychological Perspectives**. Ed. by M. Everaert, E.-J. van der Linden, A. Schenk, R. Schreuder. Hillsdale, NJ, 1995, pp. 133–166.

Джакендоф, Р. (1997): Jackendoff, R. **The architecture of the language faculty**. English. MIT Press, 1997.

Джъстесон, Дж., Катц, С. (1995): Justeson, J., S. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text". English. In: **Natural Language Engineering** (1995), pp. 9–27.

Дируестър, С., Дюма, С., Ландо, Т., Фурна, Г., Харшман, Р. (1990): Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman. "Indexing by Latent Semantic Analysis". English. In: **Journal of the American Society of Information Science** 41.6 (1990), pp. 391–407.

Еверт, С. (2005): Evert, S. "The statistics of word cooccurrences: word pairs and collocations". English. PhD thesis. Holzgartenstr. 16, 70174 Stuttgart: Universität Stuttgart, 2005.

Еверт, С. (2007): Evert, S. **Corpora and collocations**. English. 2007.

Ербах, Г. (1992): Erbach, G. "Head-Driven Lexical Representation of Idioms in HPSG". English. In: **Proceedings of International Conference on Idioms, Tilburg (NL)**. 1992.

Жанг, У., Йошида, Т., Танг, Кс., Хо, Т.-Б. (2009): Znang, W., T. Yoshida, X. Tang, Ho T.-B. "Improving effectiveness of mutual information for substantival multiword expression extraction". English. In: **Expert Systems with Applications** 36 (2009), pp. 10919–10930.

Катц, С. (1973): Katz, S. "Compositionality, idiomacity and lexical substitution". English. In: **A Festschrift for Morris Halle**. Ed. by S. Anderson, P. Kiparsky. New York : Holt, Rinehart and Wiston, 1973, pp. 357–376.

Ким, С. Н. (2008): Kim, S. N. "Statistical Modeling of Multiword Expressions". English. PhD thesis. CSSE Dept., University of Melbourne, 2008.

Ким, С. Н., Болдуин, Т. (2005): Kim, S. N., T. Baldwin. “Automatic Interpretation of Noun Compounds Using WordNet Similarity”. English. In: **Natural Language Processing – IJCNLP 2005**. Ed. by R. Dale, K.-F. Wong, J. Su, O.Y. Kwong. Vol. 3651. Lecture Notes in Computer Science. 2005, pp. 945–956. ISBN: 978-3-540-29172-5.

Ким, С. Н., Болдуин, Т. (2007a): Kim, S. N., T. Baldwin. “Interpreting Noun Compound Using Bootstrapping and Sense Collocation”. English. In: **In Proceedings of the Pacific Association for Computational Linguistics (PACLING)**. 2007, pp. 129–136.

Ким, С. Н., Болдуин, Т. (2007b): Kim, S. N., T. Baldwin. “MELB-KB: Nominal Classification as Noun Compound Interpretation”. English. In: **Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)**. 2007, pp. 231–236.

Ким, С. Н., Болдуин, Т. (2008): Kim, S. N., T. Baldwin. “Benchmarking Noun Compound Interpretation”. English. In: **Architecture** 1998 (2008), pp. 569–576.

Коева, С. (1998): Коева, С. “Граматичен речник на българския език. Описание на концепцията за организацията на лингвистичните данни”. В: **Български език** 6 (1998), стр. 49–58.

Коева, С. (2006a): Koeva, S. “Inflection Morphology of Bulgarian Multiword Expressions”. English. In: **Proceedings of 9th NOOJ Conference**. 2006.

Коева, С. (2006b): Коева, С. “Синтактични трансформации”. В: **Аргументна структура. Проблеми на простото и сложното изречение**. Съст. С. Коева. София: СемаРШ, 2006, стр. 106–138. ISBN: 978-954-8021-67-8.

Коева, С. (2007a): Коева, С. “БулНет (лексикално-семантична мрежа на българския език) – част от световната лексикално-семантична мрежа”. В: **Български език** 1 (2007), стр. 34–50.

Коева, С. (2008): Коева, С. “Българският ФреймНет. Семантико-синтактичен речник на българския език – концептуален модел”. В: **Българският ФреймНет. Семантико-синтактичен речник на българския език**. София: Институт за български език, 2008, стр. 5–57.

Коева, С. (2010): Коева, С. “Българският семантично анотиран корпус – теоретични постановки”. В: **Българският семантично анотиран корпус**. Съст. С. Коева. София: Институт за български език, 2010, стр. 7–42.

Коева, С. (2007b): Koeva, S. “Multi-word Term Extraction for Bulgarian”. English. In: **Balto-Slavonic Natural Language Processing 2007**. Prague, 2007, pp. 59–66.

Коева, С., Ризов, Б., Лесева, С. (2008): Koeva, S., B. Rizov, S. Leseva. “Chooser - A Multi-Task Annotation Tool”. English. In: **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)**. Ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani. European Language Resources Association (ELRA), 2008, pp. 728–734.

Коева, С., Стоянова, И. (2009): Коева, С., И. Стоянова. “Български национален корпус”. В: **сп. “Български език”** кн. 3 (2009).

Коева, С., Стоянова, И., Лесева, С., Търпоманова, Е., Тодорова, М. (2006): Koeva, S., S. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova. “Bulgarian Tagged Corpora”. English. In: **Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages**. Sofia, Bulgaria, 2006, pp. 78–86.

Коева, С., Благоева, Д., Колковска, С. (2011): Коева, С., Д. Благоева, С. Колковска. “Проектът Български национален корпус: резултати и перспективи”. В: **Български език** 58.3 (2011), стр. 34–53.

Крипке, С. (1980): Kripke, S. **Naming and Necessity**. English. Cambridge, Mass.: Harvard University Press, 1980.

Круз, Д. (1986): Cruse, D. A. **Lexical Semantics**. English. Cambridge University Press, 1986.

Круз, Д. (2000): Cruse, D. A. **Meaning in Language: An Introduction to Semantics and Pragmatics**. English. Oxford University Press, 2000.

Де Круиз, Т. В., Моирон, Б. В. (2007): Cruys, T. V. de, B. V. Moirón. “Lexico-Semantic Multiword Expression Extraction”. English. In: **Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands**. Ed. by Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde. 2007, pp. 175–190.

Леви, Дж. Н. (1978): Levi, J. N. **The Syntax and Semantics of Complex Nominals**. English. New York: Academic Press, 1978. ISBN: 0-12-445150-0.

Лесева, С., Стоянова, И. (2008): Leseva, S., I. Stoyanova. “Treatment of Named Entities in Machine Translation”. English. In: **Proceedings of the 2007 International Nooj Conference**. Ed. by Xavier Blanko, Max Silberztein. Barcelona, Spain: Cambridge Scholars Publishing, 2008, pp. 254–272.

Лионс, Дж. (1966): Lyons, J. “Firth’s Theory of ’Meaning’”. English. In: **In Memory of J. R. Firth**. Ed. by C.E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins. Longman : London, 1966, pp. 288–302.

Манинг, К., Шутце, Х. (1999): Manning, C., H. Schutze. **Foundations of Statistical NLP**. English. MIT Press, 1999.

Мелчук, И. (1998): Melčuk, I. "Collocations and Lexical Functions". English. In: **Phraseology. Theory, Analysis, and Applications**. Ed. by A. P. Cowie. Oxford: Clarendon Press, 1998, pp. 23–53.

Мелчук, И. (1995): Melčuk, I. "Phrasemes in language and phraseology in linguistics". English. In: **Idioms: Structural and Psychological Perspectives**. Ed. by M. Everaert, E.-J. van der Linden, A. Schenk, R. Schreuder. Hillsdale, NJ, 1995, pp. 167–232.

Мичоу, А., Серетан, В. (2009): Michou, A., V. Seretan. "A Tool for Multi-Word Expression Extraction in Modern Greek Using Syntactic Parsing". English. In: **Proceedings of the EACL 2009 Demonstrations Session, Athens, Greece, 3 April 2009**. 2009, 45–48.

Мурдаров, В. (2008): Мурдаров, В. "За правописа на новите сложни думи и съставни наименования". В: **Електронно списание LiterNet** 101.4 (2008).

Муун, Р. (1998): Moon, R. **Fixed Expressions and idioms in English: A corpus Based Approach**. English. Oxford University Press, 1998.

Накаяма, К., Хара, Т., Нишио, С. (2007): Nakayama, K., T. Hara, S. Nishio. "Wikipedia mining for an association web thesaurus construction". English. In: **Proceedings of the 8th international conference on Web information systems engineering**. WISE'07. Springer-Verlag, 2007, pp. 322–334.

Наков, Пр. (2008): **Noun Compound Interpretation Using Paraphrasing Verbs : Feasibility Study**. English. 2008, pp. 103–117.

Ничева, К. (1987): Ничева, К. **Българска фразеология**. София, 1987.

Нунавут (2000): **Nunavut, Canada's Third Territory 'North Of 60'**. English. 2000.

Нунберг, Г., Саг, И., Уасоу, Т. (1994): Nunberg, G., I. A. Sag, T. Wasow. "Idioms". English. In: **Language** 70 (1994), pp. 491–538.

Оравеч, К., Варасди, К., Наги, В. (2005): Oravecz, C., K. Varasdi, V. Nagy. "Lexical idiosyncrasy in MWE extraction". English. In: **Proceedings of the Corpus Linguistics Conference 2005**. Birmingham, 2005.

Осенова, П. (2009): Осенова, П. **Именните фрази в българския език**. София: Изд. "ЕТО", 2009.

Печина, П. (2008): Pecina, P. "A Machine Learning Approach to Multiword Expression Extraction". English. In: **Proceedings of the LREC Workshop Towards a**

**Shared Task for Multiword Expressions (MWE 2008)**. Marocco, 2008, pp. 54–57.

Пиао, С., Рейсън, П., Арчър, Д., Макенъри, Т. (2005): Piao, S., P. Rayson, D. Archer, T. McEnery. "Comparing and combining a semantic tagger and a statistical tool for MWE extraction". English. In: **Computer Speech and Language (Special issue on Multiword expressions)** 19.4 (2005), pp. 378–397.

Пиърс, Д. (2001): Pearce, Darren. "Synonymy in collocation extraction". English. In: **Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics**. Pittsburgh, 2001.

Попов, Д. (1994): Попов, Д. **Български тълковен речник**. Изд. "Наука и изкуство", 1994.

Рамиш, К., Шрайнер, П., Идиарт, М., Вилавиченцо, А. (2008): Ramisch, C., P. Schreiner, M. Idiart, A. Villavicencio. "An Evaluation of Methods for the Extraction of Multiword Expressions". English. In: **Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)**. Marocco, 2008, pp. 50–53.

Реди, С., МакКарти, Д., Манандахар, С. (2011): Reddy, S., D. McCarthy, S. Manandahar. "An Empirical Study of Compositionality in Compound Nouns". English. In: **Proceedings of the 5th International Joint Conference on Natural Language Processing**. 2011, pp. 210–218.

(1977–2008): Чолакова, К., съст. **Речник на съвременния български език**. Т. 1–13. София: Издателство на БАН, 1977–2008, А–ПРЕЛЕСТНО.

Саг, И., Балдуин, Т., Бонд, Ф., Коупстейк, А., Фликинджър, Д. (2002): Sag, I., T. Baldwin, F. Bond, A. Copestake, D. Flickinger. "Multiword Expressions: A Pain in the Neck for NLP". English. In: **Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)**. Mexico, 2002.

Секине, С. (2003): Sekine, S. **Definition of Sekine's Extended Named Entity Hierarchy**. English. 2003.

Секине, С., Судо, К., Нобата, Ч. (2002): Sekine, S., K. Sudo, Ch. Nobata. "Extended Named Entity Hierarchy". English. In: **The Third International Conference on Language Resources and Evaluation**. Canary Island, Spain, 2002.

Серетан, В., Верли, Е. (2006): Seretan, V., E. Wehrli. "Multilingual Collocation Extraction: Issues and Solutions". English. In: **Proceedings of the Workshop on Multilingual Language Resources and Interoperability**. 2006, pp. 40–49.

Серетан, В., Нерима, Л., Верли, Е. (2004): Seretan, V., L. Nerima, E. Wehrli. “A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora”. English. In: **Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004**. 2004, pp. 755–766.

Да Силва, Дж., Лопез, Дж. (1999): Silva, J. F. da, J. G. P. Lopes. “Extracting Multiword Terms from Document Collections”. English. In: **Proceedings of the VExTAL: Venezia per il Trattamento Automatico delle Lingue**. 1999, pp. 22–24.

Синклер, Дж. (1966): Sinclair, J. “Beginning the Study of Lexis”. English. In: **In Memory of J. R. Firth**. Ed. by C.E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins. Longman : London, 1966, pp. 410–430.

Синклер, Дж. (1991): Sinclair, J. **Corpus, Concordance, Collocation**. English. Oxford University Press, 1991.

Синклер, Дж. (1996): Sinclair, J. “The search for the units of meaning”. English. In: **Textus** 9 (1996), pp. 75–106.

Синклер, Дж. (1998): Sinclair, J. “The lexical item”. English. In: **Contrastive Lexical Semantics**. Ed. by E. Weigand. Amsterdam: Benjamins, 1998, pp. 1–24.

(2011): **Система за разширено търсене в Българския национален корпус**. URL: `http://search.dcl.bas.bg/`. Секция по компютърна лингвистика, ИБЕ – БАН. 2011.

Смаджа, Ф. (1993): Smadja, F. “Retrieving collocations from text: Xtract”. English. In: **Computational Linguistics** 19 (1993), pp. 143–177.

Станкович, Р., Обрадович, И., Кръстев, Цв., Витас, Д. (2011): Stanković, R., I. Obradović, Cv. Krstev, D. Vitas. “Production of morphological dictionaries of multiword units using a multipurpose tool”. English. In: **Proceedings of the Computational Linguistics - Applications Conference**. 2011, 77–84. ISBN: 978-83-60810-47-7.

Стоянова, И. (2010): Stoyanova, I. “Factors influencing the performance of some methods for automatic identification of multiword expressions in Bulgarian”. English. In: **Proceedings of the Seventh International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL), Dubrovnic, Croatia**. Ed. by M. Tadić, M. Dimitrova-Vulchanova, S. Koeva. Croatia, 2010, pp. 103–108.

Стоянова, И. (2009): Стоянова, И. “Някои методи за автоматично разпознаване на съставни лексикални единици за български език”. В: **сп. “Български език”** кн. 3 (2009).

Стоянова, И. (2008): Стоянова, И. "Основни теоретични проблеми, свързани с разпознаването на съставни лексикални единици в българския език". В: **сп. "Български език"** кн. 4 (2008).

Стъбс, М. (2002): Stubbs, M. **Words and Phrases: Corpus Studies of Lexical Semantics**. English. Blackwell Publishing, 2002.

Тинчев, Т., Коева, С. Ризов, Б., Обрешков, Н. (2008): Тинчев, Т., Коева, Б. С. Ризов, Н. Обрешков. "Система за разширено търсене в корпуси". В: **Литературата, Писането в интернет**. Университетско издателство „Св. Климент Охридски", 2008, стр. 99–116.

Тодорова, М. (2006): Todorova, M. "On The classification of Bulgarian Non-Free Phrases". English. In: **Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages**. Ed. by M. Tadić, M. Dimitrova-Vulchanova, S. Koeva. Sofia, Bulgaria, 2006, pp. 251–256.

Тодорова, М. (2010): Тодорова. "Съставни единици в Българския семантично анотиран корпус". В: **Българският семантично анотиран корпус**. Съст. С. Коева. София: Институт за български език, 2010, стр. 166–185.

Тодорова, М. (2007): Тодорова, М. "Семантико-синтактични особености на глаголни фразеологизми". В: **сп. "Български език"** кн. 4 (2007), стр. 78–91.

Тодорова, М., Обрешков, Н. (2008): Todorova, M., N. Obreshkov. "Compilation of Inflectional Dictionaries Using WordEditor". English. In: **Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages**. Ed. by M. Tadić, S. Koeva, M. Dimitrova-Vulchanova. 2008, pp. 131–138.

Уидоус, Д. (2008): Widdows, D. "Semantic vector products: some initial investigations". English. In: **Proceedings of the Second AAAI Symposium on Quantum Interaction**. 2008.

Уин, М. (2005): Wynne, M., ed. **Developing Linguistic Corpora: a Guide to Good Practice**. English. 2005.

Фелбаум, К. (1998): Fellbaum, C. **WordNet: An Electronic Lexical Database**. English. Cambridge, MA: MIT Press, 1998.

Фирт, Дж. Р. (1957): Firth, J. R. "Modes of Meaning". English. In: **Papers in linguistics 1934-1951**. Ed. by J. R. Firth. Oxford University Press, 1957, pp. 190–215.

Флайшман, М. (2001): Fleischman, M. "Automated Subcategorization of Named Entities". English. In: **39th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop**. Toulouse, France, 2001.

Франтци, К., Ананиаду, С. (1999): Frantzi, K., S. Ananiadou. “The C-value / NC value domain independent method for multi-word term extraction”. English. In: **Journal of Natural Language Processing** 6 (1999), pp. 145–179.

Халидей, М. А. К. (1966): Halliday, M. A. K. “Lexis as a Linguistic Level”. English. In: **In Memory of J. R. Firth**. Ed. by C. E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins. Longman : London, 1966, pp. 148–162.

Ханкс, П. (2000): Hanks, P. “Do word meanings exist?” English. In: **Computers and the Humanities** 34.1-2 (2000), pp. 205–215.

Хиршман, Л., Чинчор, Н. (1997): Hirschman, L., N. Chinchor. “Muc-7 named entity task definition”. English. In: **Proceedings of the 7th Message Understanding Conference (MUC-7)**. 1997.

Чатфийлд, Ч., Колинс, А. (1980): Chatfield, Ch., A. Collins. **Introduction to multivariate analysis**. English. Chapman & Hall, 1980.

Чимиано, Ф., Волкер, Дж. (2005): Cimiano, Ph., J. Volker. “Towards large-scale, open-domain and ontology-based named entity classification.” English. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)**. Ed. by G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov. Borovetz, Bulgaria, 2005, pp. 166–172.

Чомски, Н. (1980): Chomsky, N. **Rules and Representations**. English. New York: Columbia University Press, 1980.

Чуека, Я. (1988): Choueka, Yaacov. “Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases”. English. In: **RIAO**. 1988, pp. 609–624.

Чърч, К., Ханкс, П. (1990): Church, K., P. Hanks. “Word association norms, mutual information and lexicography”. English. In: **Computational Linguistics** 16 (1990), pp. 22–29.

Шен, Д., Жанг, Дж., Жу, Г., Су, Дж., Тан, С. (2003): Shen, D., J. Zhang, G. Zhou, J. Su, C. Tan. “Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain”. English. In: **41st Annual Meeting of the Association for Computational Linguistics**. Japan, 2003, pp. 49–56.

Шон, П., Джурафски, Д. (2001): Schone, P., D. Jurafsky. “Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?” English. In: **Proceedings of Empirical Methods in Natural Language Processing**. Pittsburgh, PA, 2001.

# Приложения

Приложенията включват примери, които демонстрират части от текста, но обемът им не позволява да бъдат включени непосредствено в него. Тук е поместен списъкът с използваните в текста примери за съставни лексикални единици. Също така са предоставени списък с електронни ресурси и кратко описание за всеки от тях, които са приложени към дисертацията на електронен носител, и списък с допълнителени програми и инструменти, които могат да намерят приложение при обработването на текстови корпуси.

# А. Списък на примерите, използвани в дисертацията

## А.1. Съставни лексикални единици

## А.2. Изречения

# Б. Предварителна обработка на файл

Тук са показани етапите на предварителна обработка на примерния файл 00077227 (примерът включва само част от файла) от изходното състояние на файла при свалянето му от интернет страницата на Уикипедия до окончателния XML формат, който е използван от програмата за автоматично разпознаване и тагиране.

## Б.1. Изходен файл

```
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.5/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.mediawiki.org/xml/export-0.5/
  http://www.mediawiki.org/xml/export-0.5.xsd"version="0.5"xml:lang="bg≫
  <siteinfo>
  <sitename>Уикипедия</sitename>
  <base>http://bg.wikipedia.org/wiki/</base>
  <generator>MediaWiki 1.17wmf1</generator>
  <case>first-letter</case>
  <namespaces>
  <namespace key=2"case="first-letter≫Медия</namespace>
  <namespace key=1"case="first-letter≫Специални</namespace>
  <namespace key="0"case="first-letter"/>
  <namespace key="1"case="first-letter≫Беседа</namespace>
  <namespace key="2"case="first-letter≫Потребител</namespace>
  <namespace key="3"case="first-letter≫Потребител беседа</namespace>
  <namespace key="4"case="first-letter≫Уикипедия</namespace>
  <namespace key="5"case="first-letter≫Уикипедия беседа</namespace>
  <namespace key="6"case="first-letter≫Файл</namespace>
  <namespace key="7"case="first-letter≫Файл беседа</namespace>
```

<namespace key="8"case="first-letter≫МедияУики</namespace>

<namespace key="9"case="first-letter≫МедияУики беседа</namespace>

<namespace key="10"case="first-letter≫Шаблон</namespace>

<namespace key="11"case="first-letter≫Шаблон беседа</namespace>

<namespace key="12"case="first-letter≫Помощ</namespace>

<namespace key="13"case="first-letter≫Помощ беседа</namespace>

<namespace key="14"case="first-letter≫Категория</namespace>

<namespace key="15"case="first-letter≫Категория беседа</namespace>

<namespace key="100"case="first-letter≫Портал</namespace>

<namespace key="101"case="first-letter≫Портал беседа</namespace>

</namespaces>

</siteinfo>

<page>

<title>Източно-балканска свиня</title>

<id>170305</id>

<revision>

<id>3630408</id>

<timestamp>2010-11-26T14:14:25Z</timestamp>

<contributor>

<username>Nadina</username>

<id>13270</id>

</contributor>

<comment>начало на статията</comment>

<text xml:space="preserve"bytes="17366≫[[Картинка:EBS TsvA 2.JPG|мини|300px|

Стадо Източно-балкански свине в района на село [[Ново Янково]]]]

[[Картинка:S4020798.JPG|мини|300px|Източно-балкански свине ровят в търсене на храна]]

'''Източно-балканска свиня''' е българска порода [[свиня]].

== История ==

Това е българска стара местна порода [[свиня|свине]] с древен произход. Тя възниква в резултат на кръстосването на [[дива свиня|средноевропейската дива свиня]] и архаични породи свине. Смята се, че първоначално породата е възник-

нала в [[Мала Азия]] и островите на [[Егейско море]]. По-късно е пренесена на [[Балканския полуостров]] от гръцки колонизатори от град [[Мегара]] които заселили стадата по крайбрежието на [[Черно море]]. Породата е генетично близка със средиземноморската торфена свиня (одомашнена дива свиня от подвид ''Sus scrofa scrofa''). След пренасянето и по западното крайбрежие на Черно море, тя е кръстосвана допълнително с представители на тракийската клепоуха свиня. В новото си местообитание породата е отглеждана от [[тракийски племена]] и селектирана в продължение на векове. През 1952 г. представителите от породата са представлявали 64,56

Много оскъдна е наличната информация за разпространението и характерните особености на Източнобалканската свиня. Първото съобщение за нея е от П. Германов, (1901 г.). То е в официалното издание на Министерството на търговията и земеделието наречено „Домашните животни в разните части на света и България“. В него се посочва, че в горските местности на България и най-вече по течението на река [[Камчия]] се срещат свине, които напълно приличат на „дивата свинска порода“. При проучване на акад. Хлебаров извършено в периода 1919 - 1920 г. установява, че освен по поречието на река [[Голяма Камчия]] и по долното течение на река [[Луда Камчия]] свине с подобни черти се отглеждат и в планинските райони на Варненска, Анхиалска и някои села на Бургаска околия и в почти целия Източен Балкан, поради което дава името на породата под което е известна и днес. При районирането на селскостопанските животни от 1955 г. на Източнобалканската свиня е определен развъден район основно в планинските и горски части на бившите околии: Котленска, Преславска, Шуменска, Провадийска, Варненска, Бургаска, Мичуринска, Малкотърновска, Грудовска и от части в планинските райони на околиите Поморийска, Сливенска, Еленска и Омуртагска.<ref>[http://arsis-sh.com/razv.programa%20%20.doc Стойков А., Марчев Й., Иванова-Пенева С., Кулев К., Развъдна програма за съхранение, поддържане и използване на Източнобалканската свиня]

[[Категория:Бозайници в България]]

[[Категория:Породи свине]]

[[Категория:Български породи домашни животни]]

[[Категория:Бозайници в Европа]]</text>

</revision>

</page>

</mediawiki>

## Б.2. Тагиран текст

```
,,, ,,, M
  Източно-балканска източно-балкански Asfo
  свиня свиня NCFsof
  ,,, ,,, M
  е съм VLINr3s
  българска български Asfo
  порода порода NCFsof
  [ [ U
  [ [ U
  свиня свиня NCFsof
  ] ] U
  ] ] U
  . . U
  == == M
  История история NCFsof
  == == M
  Това този PDOsn
  е съм VLINr3s
  българска български Asfo
  стара стар Asfo
  местна местен Asfo
  порода порода NCFsof
  [ [ U
  [ [ U
  свиня свиня NCFsof
  | | M
  свине свиня NCFpof
  ] ] U
  ] ] U
  c c R
```

```
древен древен Asmo
произход произход NCMsom
. . U
Тя аз PHi3sf
възниква възниквам VLINe2s
в в R
резултат резултат NCMsom
на на R
кръстосването кръстосване NCNsdn
на на R
[ [ U
[ [ U
дива див Asfo
свиня свиня NCFsof
| | M
средноевропейската средноевропейски Asfd
дива див Asfo
свиня свиня NCFsof
] ] U
] ] U
и и C
архаични архаичен Apo
породи порода NCFpof
свине свиня NCFpof
. . U
```

## Б.3. Краен XML формат

```
<?xml version="1.0"?><text current="0≫
  <word w="Източно-балканска"l="източно-балкански"sen="11437"pos="Asfo"mwe="1051:0"mwe_type
  <word w="свиня"l="свиня"sen="11437"pos="NCFsof"mwe="1051:1"mwe_type="6"/>
  <word w="е"l="съм"sen="11437"pos="VLINr3s"/>
```

<word w="българска"l="български"sen="11437"pos="Asfo"/>
<word w="порода"l="порода"sen="11437"pos="NCFsof"/>
<word w="свиня"l="свиня"sen="11437"pos="NCFsof"/>
<word w="."l="."sen="11437"pos="U"/>
<word w=-="l=-="sen="11438"pos="M"/>
<word w="История"l="история"sen="11438"pos="NCFsof"/>
<word w=-="l=-="sen="11438"pos="M"/>
<word w="Това"l="този"sen="11438"pos="PDOsn"/>
<word w="е"l="съм"sen="11438"pos="VLINr3s"/>
<word w="българска"l="български"sen="11438"pos="Asfo"/>
<word w="стара"l="стар"sen="11438"pos="Asfo"/>
<word w="местна"l="местен"sen="11438"pos="Asfo"/>
<word w="порода"l="порода"sen="11438"pos="NCFsof"/>
<word w="свине"l="свиня"sen="11438"pos="NCFpof"/>
<word w="с"l="с"sen="11438"pos="R"/>
<word w="древен"l="древен"sen="11438"pos="Asmo"/>
<word w="произход"l="произход"sen="11438"pos="NCMsom"/>
<word w="."l="."sen="11438"pos="U"/>
<word w="Тя"l="аз"sen="11439"pos="PHi3sf"/>
<word w="възниква"l="възниквам"sen="11439"pos="VLINe2s"/>
<word w="в"l="в"sen="11439"pos="R"/>
<word w="резултат"l="резултат"sen="11439"pos="NCMsom"/>
<word w="на"l="на"sen="11439"pos="R"/>
<word w="кръстосването"l="кръстосване"sen="11439"pos="NCNsdn"/>
<word w="на"l="на"sen="11439"pos="R"/>
<word w="средноевропейската"l="средноевропейски"sen="11439"pos="Asfd"/>
<word w="дива"l="див"sen="11439"pos="Asfo"mwe="1052:0"mwe_type="B"/>
<word w="свиня"l="свиня"sen="11439"pos="NCFsof"mwe="1052:1"mwe_type="B"/>
<word w="и"l="и"sen="11439"pos="C"/>
<word w="архаични"l="архаичен"sen="11439"pos="Apo"/>
<word w="породи"l="порода"sen="11439"pos="NCFpof"/>
<word w="свине"l="свиня"sen="11439"pos="NCFpof"/>

```
<word w="."l="."sen="11439"pos="U"/>
</text>
```

# В. Списък на електронните ресурси, приложени към дисертацията

## В.1. Списък на съставните лексикални единици именни фрази, използвани в изследването по категории

около 27,000 единици

Този списък включва съставни лексикални единици от два основни източника:

- Съставни лексикални единици, извлечени полуавтоматично от **Български тълковен речник** (1994) и верифицирани ръчно.
- Съставни лексикални единици, извлечени полуавтоматично от Уикипедия.

Единиците са разпределени по категории според Класификация 6, стр. 76.

## В.2. Списък на съставните лексикални единици именни фрази по синтактична структура

## В.3. Честотен списък на съставните лексикални единици, използвани в изследването

# Г. Списък на допълнителните компютърни инструменти

По-долу са описани модули и приложения, които могат да бъдат ползвани за изпълнението на конкретни задачи, свързани с автоматичната обработка на български език. Голяма част от приложенията са езиково независими.

1. Създаване на корпус от статии от Уикипедия
   - Запазване на текстове от Уикипедия в XML формат
   - Селекция на текстовете по определени критерии (например обем)
   - Извличане на информация за описанието на корпусните единици
2. Извличане на лексикална информация
   - Автоматично извличане на списък със съставни лексикални единици от корпуса от Уикипедия
   - Автоматично извличане на списък с всички срещани (слово)форми на съставните лексикални единици
   - Извличане на преводен речник
3. Модул за преобразуване на формата на текстовете
   - Преобразуване Уикипедия XML → чист текст
   - Преобразуване чист текст → XML
   - Преобразуване XML → чист текст
4. Система от `Java` класове за описание на текст
5. Модул за предварителна обработка
   - Чънкиране
   - Тагиране на текстовете – за български език, през интернет
   - Честота на несвободни фрази – леми и форми
6. Модул за извличане на честотна информация от корпус
   - Честота на прости леми и прости словоформи
   - Честота на синтактични конструкции
   - Честота на несвободни фрази – леми и форми

- Честота на думи (леми и словоформи) и несвободни фрази в БНК – за български език, през интернет

7. Модул за разпознаване на съставни лексикални единици
   - Изчисляване на мерките за асоциация
   - Оценка на резултатите
8. Модул за визуализация
9. Модул за статистически анализи с `R` през `Java` с визуализация