
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

Автореферат

---

Автор:
Ивелина Стоянова

Научен ръководител:
Проф. Светла Коева

София
19 април 2012 г.

# Съдържание


**1 Увод** **1**

1.1 Актуалност на проблематиката . . . . . 1

1.2 Цел, задачи и принципи на изследването . . . . . 3

1.2.1 Цел на изследването . . . . . 3

1.2.2 Задачи на изследването . . . . . 3

1.2.3 Основни принципи на изследването . . . . . 4

1.3 Използвани ресурси . . . . . 5

**2 Теоретично представяне и анализ на СЛЕ** **7**

2.1 Преглед на основната терминология . . . . . 7

2.2 Основни възгледи за несвободните фрази . . . . . 11

2.3 Класификация на несвободните фрази . . . . . 13

2.4 Дефиниция и обхват на понятието за СЛЕ . . . . . 17

**3 Лингвистично описание и класификация на СЛЕ** **21**

3.1 Лингвистично описание на СЛЕ . . . . . 21

3.1.1 Семантични особености на СЛЕ . . . . . 21

3.1.2 Синтактични особености на СЛЕ . . . . . 24

3.1.3 Прагматични особености на СЛЕ . . . . . 25

3.2 Необходимо ли е включването на СЛЕ в речника? . . 26

3.3 Приложна класификация . . . . . 28

**4 Автоматично разпознаване и тагиране на СЛЕ** **32**

4.1 Обща характеристика на методите . . . . . 32

4.2 Анализ на факторите, влияещи върху ефективността 34

4.3 Възможности за разпознаване на СЛЕ . . . . . 36

4.4 Разработване на компютърна програма . . . . . 39

4.5 Подготовка на ресурси и корпус за оценка . . . . . 40

4.6 Методи за оценка на резултатите . . . . . 41

4.7 Експериментално приложение на методи . . . . . 43

4.7.1 Метод с честотен анализ и синтактичен филтър 44

4.7.2 Метод с лингвистични правила . . . . . 46

**5 Резултати** **49**

5.1 Приноси на разработката . . . . . 49

5.2 Някои насоки и идеи за бъдеща работа . . . . . 50

**Библиография** **52**

# 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 (2006) 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 още *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)**.**

- 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 разпозната/преведена несвободна фраза

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

  *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****.*

(Google Translate, URL: `http://translate.google.com/`)

В българската теоретична 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 отделните категории несвободни фрази.

Дисертацията 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.

## 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 – качествена методология и нейното успешно приложение.

Необходимо е да се подчертае, че разработката има приложна насоченост и представеният теоретичен модел е съобразен с практическата цел, която се изразява в разработване на методология за автоматично разпознаване на съставните лексикални единици.

### 1.2.2 Задачи на изследването

Във връзка с реализирането на поставената цел се очертават 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 за усъвършенстване.

### 1.2.3 Основни принципи на изследването

Изследването се основава на ясно определени 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 да 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 конкретните практически експерименти 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, 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.

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 обектът е ограничен до съставните лексикални единици именни фрази.

## 1.3 Използвани ресурси

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 намерила приложение в различни

изследвания в областта на компютърната лингвистика (Накаяма и др., 2007; Бекавац и Тадич, 2008; Винче и др., 2011).

Корпусът е наречен *Уики1000+* и обхваща текстовете, съдържащи над 1000 думи. Това значително ограничи броя файлове – *Уики1000+* се състои от 6311 текстови единици, наброяващи общо 13.4 милиона думи (за сравнение – в Уикипедия има над 170,000 статии на български език).

В изследването намериха приложение лексикални ресурси, някои от които са предварително разработени от Секцията за компютърна лингвистика, а други – специално създадени за целите на дисертацията с помощта на автоматични и полуавтоматични методи. Използват се списъци с наименования от различни области: лични имена, топоними, дескриптори и наименования на организации и др.

Също така беше извлечен списък с несвободни фрази, които са включени като такива към съответните речникови статии в **Български тълковен речник** (Попов, 1994) – общо 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.

Автоматично бяха извлечени и списъци и речници със съставни лексикални единици от *Уики1000+*. Този процес е описан по-нататък.

# 2 Теоретично представяне и анализ 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. Необходимо е да бъде 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.

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.

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. 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 колокациите като лексикализирани идиосинкретични комплекси от думи (за подробности вж. Еверт, 2007).

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 се дава 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.

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**). Българският превод на терминологията е заимстван от Тодорова (2007, 2010).

Обект на изследване в дисертацията 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 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.

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. Новоконструираният термин **свободна колокация** изтъква характера им на колокации, като същевременно насочва и към това, че са свободни фрази.

Таблица 1 представя в обобщен вид терминологията, използвана в дисертацията.

| **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 |

| **Основен термин** | **Английски** | **Синоними** |
|---|---|---|
| несвободна фраза | non-free phrase, phraseme, set phrase, multiword expression, MWE, multiword unit | лексикална единица от повече от една дума, фразема |
| неразложима фраза | non-decomposable phrase, idiomatic phrase, frozen expression | идиоматична фраза, некомпозирана фраза, фразеологизъм, фразеологично съчетание, устойчиво съчетание |
| идиосинкретично разложима фраза | idiosyncratically decomposable phrase, idiomatically combining expressions | идиоматично комбиниран израз, идиосинкретично композирана фраза |
| съставна лексикална единица | decomposable phrase, compound, extended lexical item, pragmateme, pragmatic phraseme | разложима несвободна фраза, съставна дума, композирана фраза, прагматема, прагматична фразема |

| **Основен термин** | **Английски** | **Синоними** |
|---|---|---|
| свободна колокация | free collocation<br>purely statistically marked phrase | |

Таблица 1: Основна терминология, използвана в дисертацията.

## 2.2 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. 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. Разграничават се 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 (изцяло) обусловени от логическите отношения. Утвърждаването на фразеологичните съчетания става чрез традицията и широката употреба.

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.

Нунберг, Саг и Уасоу (1994) 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.

Мелчук (1995) въвежда понятието **несвободна фраза**, като използва също и названията **устойчива фраза** и **фразема**. Свободните фрази се отличават с **нерестриктивност** и **регулярност** (англ. **unrestrictedness** и англ. **regularity**) във формата и значението.

За разлика от свободните фрази, при несвободните значението не се формира нерестриктивно и регулярно, а се наблюдават посложни 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 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).

## 2.3 Класификация на несвободните фрази

Болдуин, Банард, Танака и Уидоус (2003) разделят несвободните фрази на няколко основни типа: неразложими, идиосинкретично разложими и разложими. Авторите описват първата категория като лексикални единици, които не предлагат възможност за декомпозиционен анализ и в повечето случаи не допускат вътрешни модификации. 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 вариа-

ции, 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).

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 *ритна 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** – определена степен на (не)композираност на значението, участие в семантични релации с прости лексикални единици (например *пощенска станция – поща*);
- **Прагматична маркираност** – прагматическите характеристики на частите не съвпадат с тези на цялата единица или изразът се асоциира с определена прагматична отправна точка (например изразът *Пушенето забранено!*, който се свързва с определена комуникативна ситуация и се характеризира с ла-

пидарност, който в друга ситуация би бил неучтив, например употребен от сервитьор към клиент);

- **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*).

Направеният в дисертацията преглед на възгледите относно несвободните фрази е необходим, за да се очертае мястото на съставните лексикални единици в системата на езика. В настоящата разработка се приема 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).

Към тях са добавени две допълнителни категории. 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. Втората добавена категория в класификацията е тази на свободните колокации, които са единствено статистически маркирани, без да са маркирани по другите (езикови) показатели.

**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*).

При съставните лексикални единици 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 2).

**Пример 2** (Свободни колокации)**.**

*пресен хляб*

| **Пример** | **Честота в БНК** |
|---|---|
| *пресен хляб* | 127 |
| *пресен бял хляб* | 5 |

| | |
|---|---|
| ?*нов хляб* | 7 |
| ?*свеж хляб* | 1 |

## 2.4 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.
- Значението им се формира от значението на конституентите и е разбираемо. По тази характеристика се доближават до сложните думи и често се разглеждат във връзка с тях (Болдуин, 2004; Джакендоф, 1997).
- Съставните лексикални единици се разграничават от свободните словосъчетания по това, че се появяват със значителна честота в езика, като се противопоставят на синонимни/антонимни съчетания, появяващи се със значително по-ниска (или дори нулева) честота (Болдуин, 2004, Пиърс, 2001).
- Макар като цяло да се твърди, че проявяват относителна синтактична/структурна свобода, в много случаи те се характеризират със синтактични ограничения.

Може да се изведе следната дефиниция.

**Дефиниция 1** (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.

Предложената дефиниция има за цел да обособи категорията на съставните лексикални единици в системата на несвободните фрази, като очертае макар и условни граници с останалите категории, както и със свободните фрази (Фигура 1).

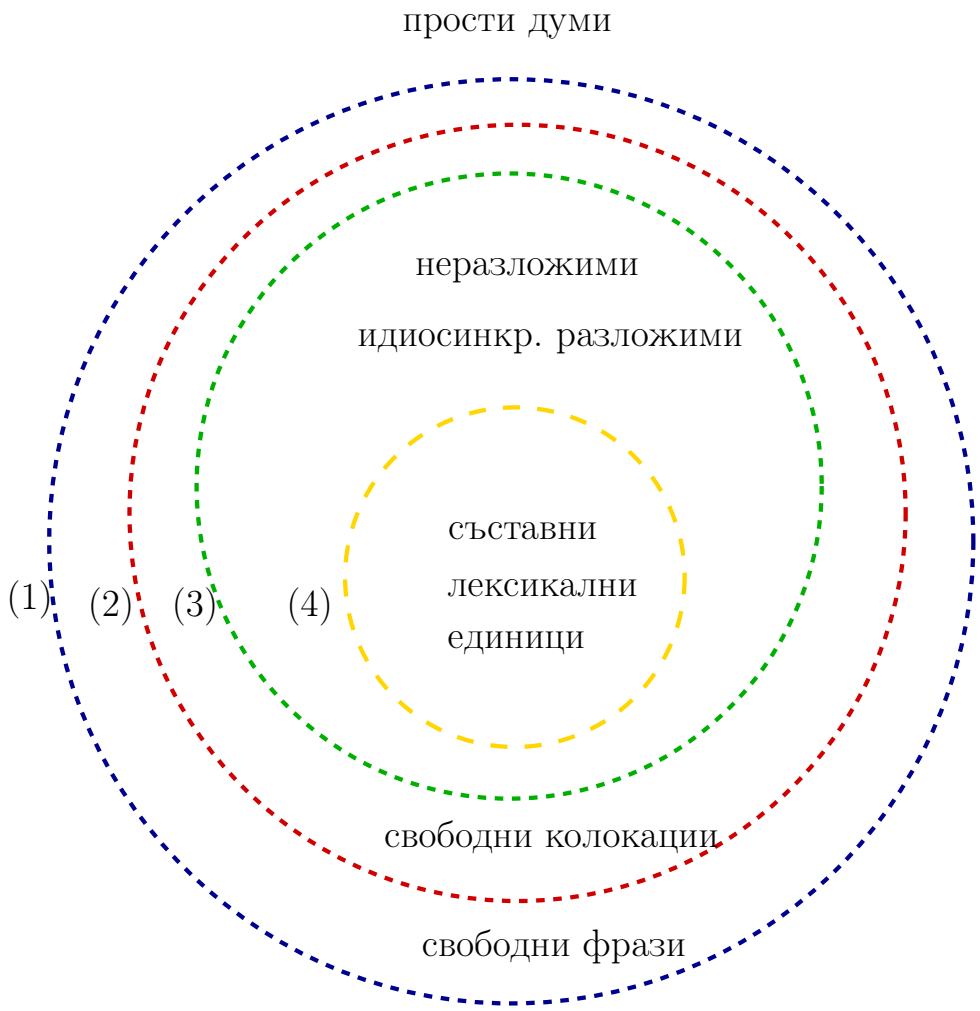
Посочената дефиниция има теоретичен характер и произлиза от наблюденията върху езиковите характеристики на несвободните фрази. Следвайки примера на Саг и др. (2002), Банард и др. (2003), Еверт (2007) и др., е направен опит за изграждане и на практическа (емпирична) характеристика на съставните лексикални 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 2** (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 конструира-




Фигура 1: Приложение на условията от Дефиниция 1 за обособяване на категорията на съставните лексикални единици от другите езикови единици.

на (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 Дефиниции 1 и 2 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. 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.

# 3 Лингвистично описание и класификация на съставните лексикални единици

## 3.1 Лингвистично описание на съставните лексикални единици

### 3.1.1 Семантични особености на съставните лексикални единици

Подходът на Мелчук (1998) към описанието на несвободните фрази се опира на това, че те са лингвистични знаци. В лингвистичното описание на съставните лексикални единици като част от несвободните фрази ключова роля играят понятията за **нерестриктивност** и **регулярност** при формиране на означаваното и означаващото.

В случая със съставните лексикални единици много често единицата съдържа значенията на конституентите $A$ и $B$, но и допълнително значение $C$, което не се предава нито с $A$, нито с $B$ (дори имплицитно), например *свидетелство за съдимост* съдържа неизразен компонент *липса на [осъдителни присъди]*.

Нерегулярността и 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 *сълзотворен газ*.

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).

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, 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 3 представя фрази с различна опора и подчинен конституент *морски*, които се характеризират с различна степен на композиционалност.

**Пример 3.**

- *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, но впос-

ледствие фразата придобива и значение на електронна пощенска кутия.

### 3.1.2 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

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

(*къща музей*, *бизнес център*)

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*).

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*).

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.

### 3.1.3 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 уста-

новяват 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 (еднозначни означаващи). Случайните означаващи от друга страна не притежават тази способност. Добавянето на описания или дескриптори към имената се използва за фиксиране на референцията, при което дескрипторите, които могат да бъдат изразени и от фраза, стават неделима част от наименованието.

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

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

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

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 гьола*), 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 степенния и класовия подход, описани от Болдуин (2006).

## 3.3 Приложна класификация според характеристиките на идиоматичността

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.

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 $C$ от значението. 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 2** (Според идиоматичността)**.**

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

ретично 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 – *Челопечѐнско 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; *бъ̀лгарски 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 категория формално съвпада с ... (например *пещерен лъ̀в*).

(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; *Съ̀юз 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 изис-

кват специфичен подход при обработката, особено с оглед на машинния превод.

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).

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

## 4.1 Обща характеристика на методите

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.

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, тук са разгледани методи за разпознаване на несвободните фрази, но е обърнато внимание на приложимостта им с оглед на разпознаването и класификацията на съставните лексикални единици.

Лингвистичните методи се основават на прецизно разработени лингвистични ресурси и модели, които описват лексикалните и гра-

матическите характеристики на езиковите единици – думи, фрази, изречения. Лингвистичните модели са създадени от специалисти езиковеди и имат за цел да опишат възможно най-точно и изчерпателно изследваното явление и да осигурят успешното и точното му разпознаване в текста.

Като най-обща оценка може да се изтъкне, че те се отличават с висока точност, но от една страна са ограничени върху специфично езиково явление или отделна характеристика, а от друга – са неикономични за разработване, тъй като изискват голямо количество ресурси под формата на време и човешки труд. 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.

Количествените методи се появяват като алтернатива на лингвистичните. Те разчитат на това, че при наличие на значително количество данни, езиковите единици, демонстриращи сходни явления, ще проявяват и сходно поведение. Поведението може да се изразява в честота на поява 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 тър-

сят 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 да бъдат формулирани научно обосновани изводи, които да не се базират на субективна преценка, произволно тълкувание или манипулация на статистическите данни.

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.

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.

## 4.2 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.

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 явления 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.

- **Анотацията на корпуса**.

  Лингвистичната анотация на корпуса – тагиране с част на реч-та и граматически характеристики, семантична информация и други, 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.

## 4.3 Възможности за разпознаване на съставните лексикални единици

Прави впечатление, че различните типове методи имат свойството на регистрират различни характеристики на несвободните фрази и 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 да 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 свободните 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 Дефиниция 1 (стр. 17) и представянето ѝ на Фигура 1 (стр. 19), 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 използват честотният метод и различните мерки за асоциация.

Вторият етап е разграничаването между съставните лексикални единици и останалите категории несвободни фрази, при което е подходящо прилагането на векторен метод за представяне на значението. Използването 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 парафразиране, 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 2, стр. 28) 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 компютърните приложения за машинен превод и извличане на информация.

Чрез няколко експеримента се демонстрира приложението на методите на няколко етапа. Избраният корпус *Уики1000+* съдържа научно-популярни текстове. Поради това броят на неразложими и идиосинкретично разложими несвободни фрази е много малък – те представляват едва 1.25% от несвободните фрази в корпуса. Техният брой се оказва недостатъчен за качественото приложение и оценка на метода за тяхното разграничаване от съставните лексикални единици.

От друга страна основното приложение на задачата за разпознаването и тагирането на съставни лексикални единици са в научните и административните области – за автоматично извличане на терминология, автоматичен превод и др. Това се отразява и от характера на използваните 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.

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

Беше разработена компютърна 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.

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 програмиране 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 конкретна 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.

## 4.5 Подготовка на лексикалните ресурси и на полуавтоматично анотиран корпус за оценка

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.

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 допълни-

телно 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 синтактични конструкции, може да намери приложение като начин да бъдат подготвени и подобрени ресурсите за прилагане и тестване на други методи като например статистическите методи.

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.

## 4.6 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.

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.

Ван Рижсберген (1979) описва шест основни количествено измерими показателя за ефективност на програмите за автоматично извличане на информация:

1. **Покритие на корпуса** – степента, до която текстовият материал е показателен за изследваните явления.
2. **Времетраене на обработката** – времето, което отнема изследването от пускането на програмата до извеждането на крайния резултат.
3. **Форма на представяне на резултатите**.
4. **Усилие, необходимо за извършване на изследването**.
5. **Пълнота** (англ. **recall**).
6. **Точност** (англ. **precision**).

Най-често прилаганите показатели, по които се оценява успешността на даден метод, са **точността** (англ. **precision**) и **пълнотата** (англ. **recall**)[1]. 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 $P$ и пълнотата $R$ се изчисляват по следните формули:

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

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

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, като се цели балансиране между точността и пълнотата.

Като най-подходяща се приема мярката $F_\beta$ (Ван Рижсберген, 1979), която се използва за изчисляване на средната стойност на

[1] Терминът 'пълнота' е използван от Колковска (2005). Друг възможен превод на термина 'recall' е 'покритие'.

обратни величини. Формулата за $F_\beta$ е следната:

$$F_\beta \;=\; (1+\beta^2)\,\frac{P\cdot R}{\beta^2\cdot P+R},$$

където $P$ и $R$ са съответно точността и пълнотата, а $\beta > 0$ отразява колко пъти по-голям приоритет (тегло) е даден на пълнотата спрямо точността.

За оценка в настоящото изследване ще бъде използвана стойността $\beta = 0.5$, чрез която се дава двойно по-голяма тежест на точността спрямо пълнотата.

$$F_{0.5} \;=\; \left(1+\left(\frac{1}{2}\right)^2\right)\,\frac{P\cdot R}{\left(\frac{1}{2}\right)^2\cdot P+R} \;=\; \frac{5\cdot P\cdot R}{P+4\cdot R}.$$

Тази оценка ще бъде ползвана за сравнителен анализ при приложение на различни методи или варианти на даден метод в рамките на настоящото изследване.

## 4.7 Експериментално приложение на методи за разпознаване и тагиране на съставни лексикални единици в българския 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 – лингвистични и статистически. Задачата за разпознаването не е решена и в световен мащаб и изисква експериментиране с лексикални ресурси, различни по вид корпуси, лингвистични и статистически методи.

### 4.7.1 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 намери каква е оптималната граница, като най-често става въпрос за установяване на баланс между точност и пълнота на резултатите (вж. 4.6). От Таблица 3, с изключение на малкото колебание при граница $\mathrm{Freq}_{\mathrm{min}} = 100$ може да се каже, че с увеличаването на $\mathrm{Freq}_{\mathrm{min}}$ 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 ($\mathrm{Freq}_{\mathrm{min}}$), която зависи както от големината на корпуса, така и от честотата на изследваното явление.

| **Долна граница за честотата** $\text{Freq}_{\textbf{min}}$ | **Точност** $P$, % | **Пълнота** $R$, % | $F_{0.5}$ |
|---|---|---|---|
| 25 | 76.32 | 73.84 | 75.81 |
| 50 | 96.69 | 63.48 | 87.53 |
| 75 | 97.12 | 49.34 | 81.36 |
| 100 | 96.99 | 40.72 | 75.99 |
| 200 | 97.25 | 25.33 | 62.03 |
| 300 | 97.79 | 18.57 | 52.77 |

Таблица 3: Резултати от приложение на честотния метод с различни стойности за $\text{Freq}_{\text{min}}$.

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.

Проблемно е и 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 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.

### 4.7.2 Метод, основан на лингвистични правила

Следният списък с правила, съответстващи на алгоритъма, представен на Фигура 2, беше съставен и тестван върху *Уики100+*:

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.

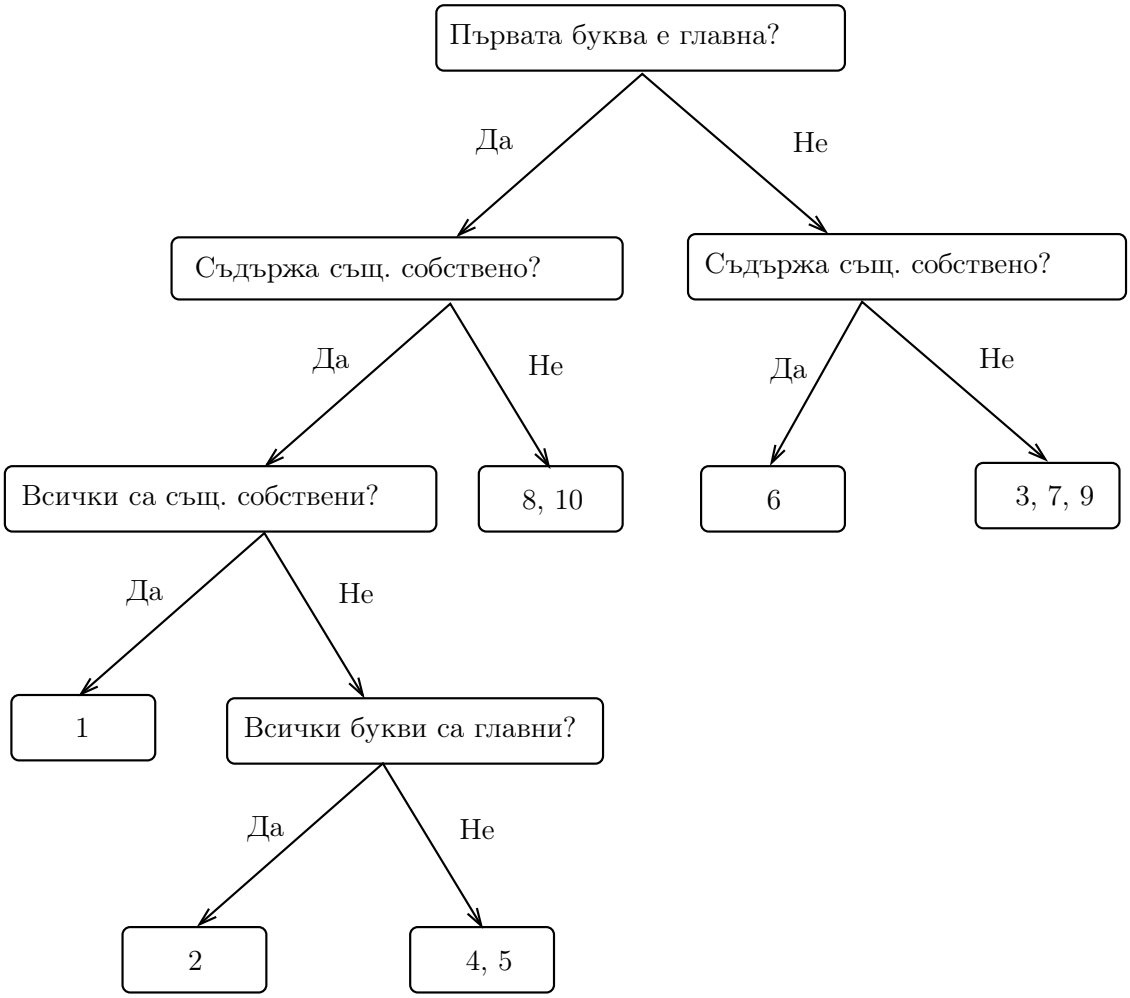
3. В единица, включваща съществително собствено, и други думи, на която първата дума започва с главна буква – с най-голяма вероятност е категория 4 или 5.

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 10.

6. Единица, която не съдържа съществително собствено и която не започва с главна буква – с най-голяма вероятност е категория 3, 7 или 9.




Фигура 2: Алгоритъм за определяне на категориите съставните лексикални единици.

Категория 4 представлява едва 0.35% от несвободните фрази, докато категория 5 – 6.13%. Това показва, че можем да приемем групата от категории 4 и 5 за категория 5. Аналогично, групата от 8 и 10 може да бъде приравнена към 8 – 10 (0%) и 8 (5.39%); а групата от 3, 7 и 9 да бъде приравнена към 7 – 3 (0.47%), 9 (0.47%) и 7 (45.42%).

При прилагане на този прост алгоритъм се постига точност от 91.51%. Грешките се получават от изкуственото приравняване на категории, както и от някои допълнителни особености – например появата на главни букви в началото на изречение. Въпреки това резултатът е сравнително добър. Също така моделът може да бъде прецизиран за постигане на по-точни резултати.

# 5 Резултати

## 5.1 Приноси на разработката

Приносите на изследването могат да бъдат обобщени до следното:

1. Разработката се занимава със съставните лексикални единици в българския език, които не са добре описани и изследвани 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.

## 5.2 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 валидността 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


DCL-Tagger (2011): **DCL Tagger and Lemmatiser**. URL: `http://dcl.bas.bg/services/`. Секция по компютърна лингвистика, ИБЕ – БАН. 2011.

Wikipedia (2011): **Уикипедия**. URL: `http://bg.wikipedia.org`. март 2011.

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.

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.

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.

Ван Рижсберген, К. (1979): Van Rijsbergen, C. J. **Information retrieval**. English. URL: `http://www.dcs.gla.ac.uk/Keith/Preface.html`. Butterworth-Heinemann, 1979.

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.

БАН, Граматика на (1983): Стоянов, С., съст. **Граматика на съвременния български книжовен език**. Т. 1-3. Издателство на БАН, 1983.

Джакендоф, Р. (1997): Jackendoff, R. **The architecture of the language faculty**. English. MIT Press, 1997.

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.

Коева, С. (2006): Koeva, S. "Inflection Morphology of Bulgarian Multiword Expressions". English. In: **Proceedings of 9th NOOJ Conference**. 2006.

Колковска, С. (2005): Колковска, С. **Модели на концептуалносемантичните отношения между термините в специален (химически) текст с оглед на автоматичното им разпоз-**

**наване**. URL: `http://liternet.bg/publish16/s_kolkovska/modeli/index.html`. LiterNet, 2005.

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.

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.

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.

Ничева, К. (1987): Ничева, К. **Българска фразеология**. София, 1987.

Нунберг, Г., Саг, И., Уасоу, Т. (1994): Nunberg, G., I. A. Sag, T. Wasow. "Idioms". English. In: **Language** 70 (1994), pp. 491–538.

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.

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.

Синклер, Дж. (1991): Sinclair, J. **Corpus, Concordance, Collocation**. English. Oxford University Press, 1991.

Синклер, Дж. (1998): Sinclair, J. "The lexical item". English. In: **Contrastive Lexical Semantics**. Ed. by E. Weigand. Amsterdam: Benjamins, 1998, pp. 1–24.

Стъбс, М. (2002): Stubbs, M. **Words and Phrases: Corpus Studies of Lexical Semantics**. English. Blackwell Publishing, 2002.

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.

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.

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.