Концептуално и компютърно моделиране на езикови структури и обекти (с приложения за българския език)

доц. Г. Тотков

I. **Въведение**

Морфологични анализатори

II. **Идеята на** BulMorph **2.0**

Морфологичният анализ

III. **Системата BulMorph 2.0**

Анализатор

Синтезатор

Лематизатор

IV. **Компютърно моделиране на българското словообразуване**

BulMorph 2.0 – Морфологичен анализатор, Синтезатор, Лематизатор и Експерт

За морфологичните анализатори

Морфологични анализатори за естествени езици:

- KIMO – морфологични анализатори за английски език
- MULTEX – лингвистични средства за 6 европейски езика
- MORPHY - анализатор за немски.

Първите морфологични анализатори за български текстове:

- BulMorph 1.0 (Пловдивски Университет, 1988-1990) използва 187 флективни типа
- MorphoAsistant (БАН)

Първите морфологични анализатори са доста бавни:

- BulMorph 1.0 анализира около 900 думи в секунда (на 133 MHz процесор)
- MORPHY анализира 300 немски думи в секунда на бързо PC

Подходи за моделиране на граматични речници

Базирани на ациклични крайни автомати с етикети на крайните състояния и ациклични преобразуватели с крайни състояния. Някои проблеми и недостатъци:

- от словоформа на дадена дума не е възможно да се синтезира друга (произволна) нейна словоформа;
- използваните структури не съответстват по "естествен" начин на моделираните граматични явления;
- представянията не са удобни за получаване на граматична информация за "непозната" словоформа (или това въобще не е възможно);
- необходимо е да се създават различни структури (автомати с крайни състояния), ако се налага моделиране на друг вид анализ или речници (например деривационен речник) и др .

Концепцията на BulMorph 2.0

- Решава посочените проблеми и недостатъци
- Използва модификации на посочените модели
- Представя двустепенен преобразувател с крайни състояния (bPFST)

bPFST е подобен на стандартен FST, но преходите се управляват от два FSTs.

bPFST са преобразуватели с крайни състояния със свойството да преминават по време на прехода от един FST към друг

Запазвайки простотата на FST има по-голяма мощност

Пълната морфологична информация за началните 1,500,000 словоформи заема по-малко от 1 MB дисково пространство в BulMorph 2.0. За сравнение, текстовото представяне на същия речник изисква около 106 MB

Нови решения

- Разпределяне на лингвистичната информация между два FSTs интегрирани в един bPFST:
  - 1-ви FST наречен Морфологично Ядро (МЯ) остава непроменен за всички словоформи от парадигмата на една дума. Включва само един шаблон за всяка парадигма (инвариантна част)
  - 2-ри FST, наречен Морфологична Обвивка (МО) представя флективните явления и словообразувателните характеристики на думите извлечени от изходния корпус (вариантна част)
- Конструиране на FST с постепенно нарастване осигуряващо едновременното създаване и минимизиране
- Използване на bPFST като анализатор, синтезатор, лематизатор и за разпознаване на "непозната" дума

Парадигмите на думите *пера* и *вятър* като входни данни в МЯ и МО

The word paradigm	Entries to	
	Morphological kernel (patterns)	Morphological shell (paradigms)
пера (wash) 1	п *е- р 170[1]	е , а 1
пере ш 2		е , еш 2
пере 3		е , е 3
перем 4		е , ем 4
перете 5		е , ете 5
перат 6		е , ат 6
прах 7		ε[2] , а 7
...		...
вятър (wind) 1	в *я-е т *ъ- р 4[3]	я , ъ , ε 1
вятъра 2		я , ъ , а 2
вятърът 3		я , ъ , ът 3
ветрове 4		е , ε , ове 4
ветровете 5		е , ε , овете 5
вятъра 6		я , ъ , а 6

1. Номерът на морфологичния клас е 170
2. ε-символ (низ с нулева дължина)
3. Номерът на морфологичния клас е 4

Флективни явления в българския език и тяхното представяне

№	Inflectional phenomena	Wild characters	Examples
1.	Shift between 2 characters 'я'–'е', 'е'–'й', 'ц'–'ч'	*я-е *е-й *ц-ч	вятър-ветрове
2.	Shift between 4 characters 'ъ'–'а'–'о'–'и' 'я'–'е'–'и'–'й'	*я-е-и-й *ъ-я-е-и	такъв-така-това-такива
3.	Disappearing 'е', 'ъ', 'р', 'и', 'о'	*е- *ъ- *р- *и- *о-	пера-п|рах вятър-вет|рове
4.	Shift between 3 characters and ε 'ε'–'а'–'о'–'и'	*а-о-и-ε	какъв|то-каквато-какво-каквито
5.	Sequence without inflection	,	пер|а-пер|е

Изграждане на BulMorph 2.0

Системата е базирана на Български Изходен Корпус (БИК) съдържащ повече от 1,750,000 български словоформи, производни на повече от 80,000 основни форми от 231 морфологични класа

- част от корпуса (около 69,500 основни форми и съответстващите им 1,500,000 словоформи) са използвани за конструирането на прототип на BulMorph 2.0
- останалите (повече от 11,000 основни форми и 230,000 словоформи)– за тестване на методологията и прототипа
- BulMorph 2.0 е конструиран на базата на пълния корпус БИК

BulMorph 2.0 съдържа три модула: морфологичен *Анализатор* (включващ *Експерт*), *Синтезатор* (модул генериращ възможните форми на думите) и *Лематизатор* (определя основната форма)

Изграждане на морфологичното ядро и обвивка

Около 1,500,000 словоформи са обработени на 5 етапа:

Етап 1. Морфологична класификация на българските словоформи в 231 флективни типа (парадигми на думите)

Етап 2. Образуване на част от БИК който за всеки вид парадигми включва всички словоформи на дадената парадигма

Етап 3. Генериране на множество от шаблони за избраните парадигми (алгоритъм базиран на класическия метод за представяне чрез съседни матрици)

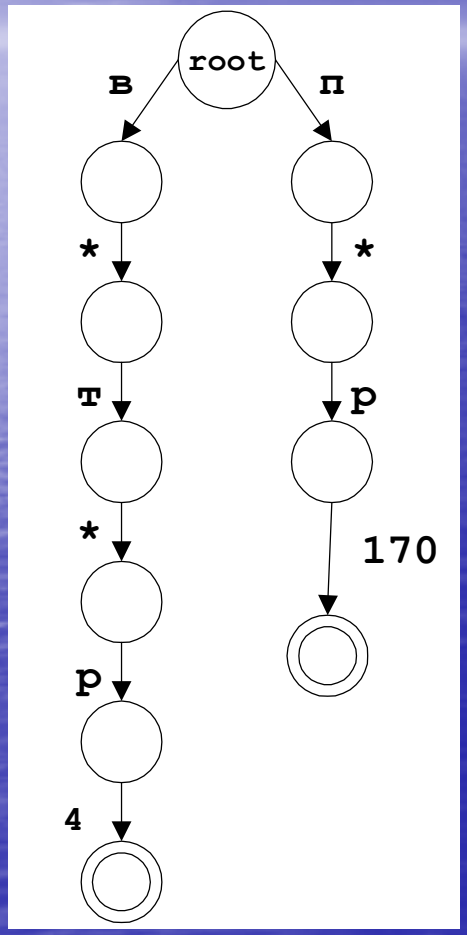
Етап 4. Генериране на морфологичната обвивка използвайки избраната парадигма и кореспондиращите шаблони

Етап 5. Изграждане на морфологичното ядро на базата на морфологичната обвивка

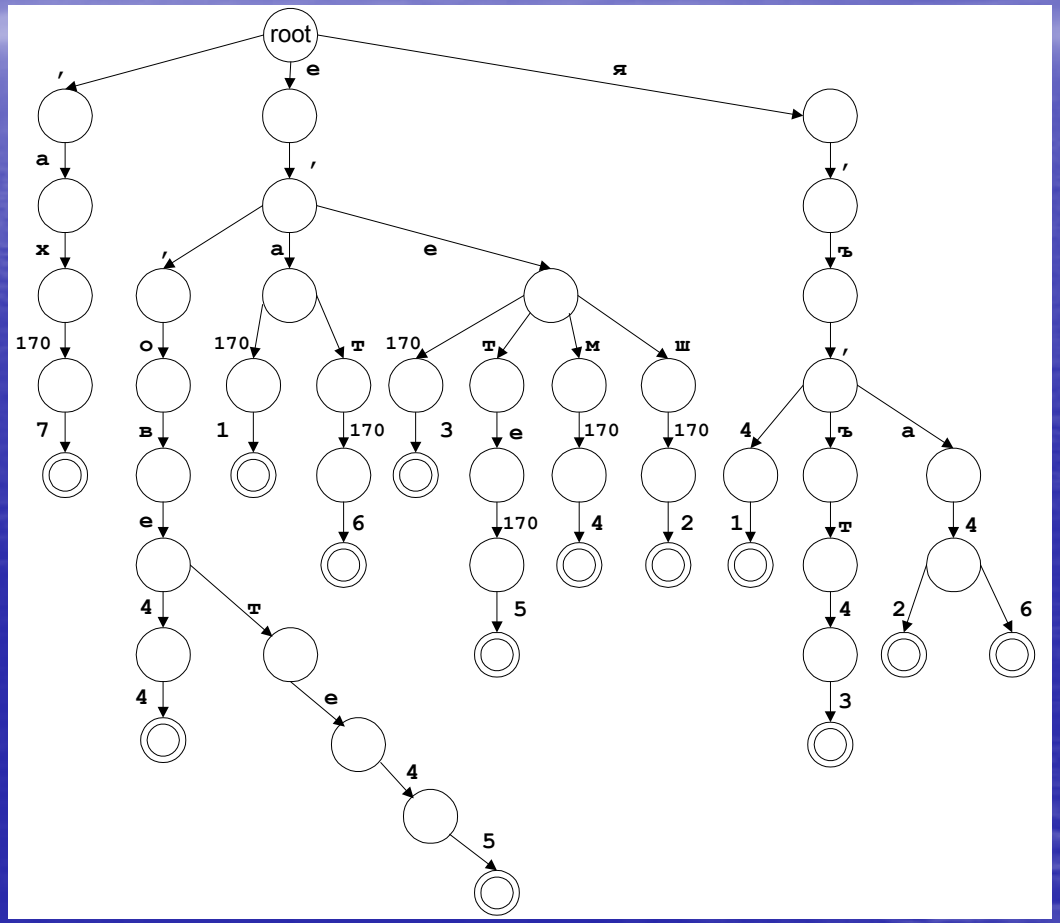
МЯ и МО представени в прав ред за парадигмите

‘вятър-вятъра-вятърът-ветрове-…’

‘пера-переш-пере-перем-…’



А. Морфологично ядро

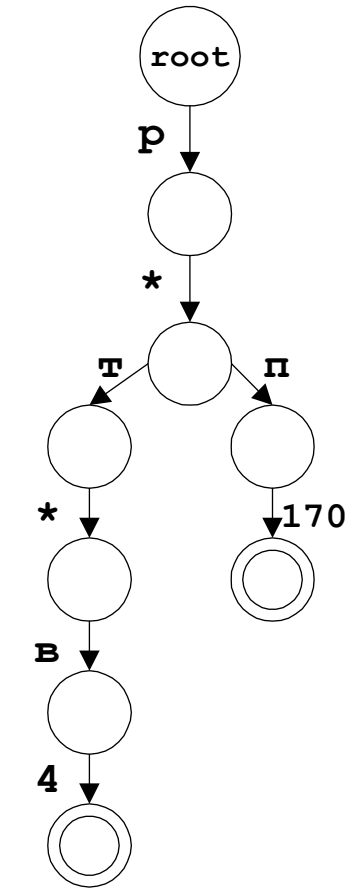


В. Морфологична обвивка

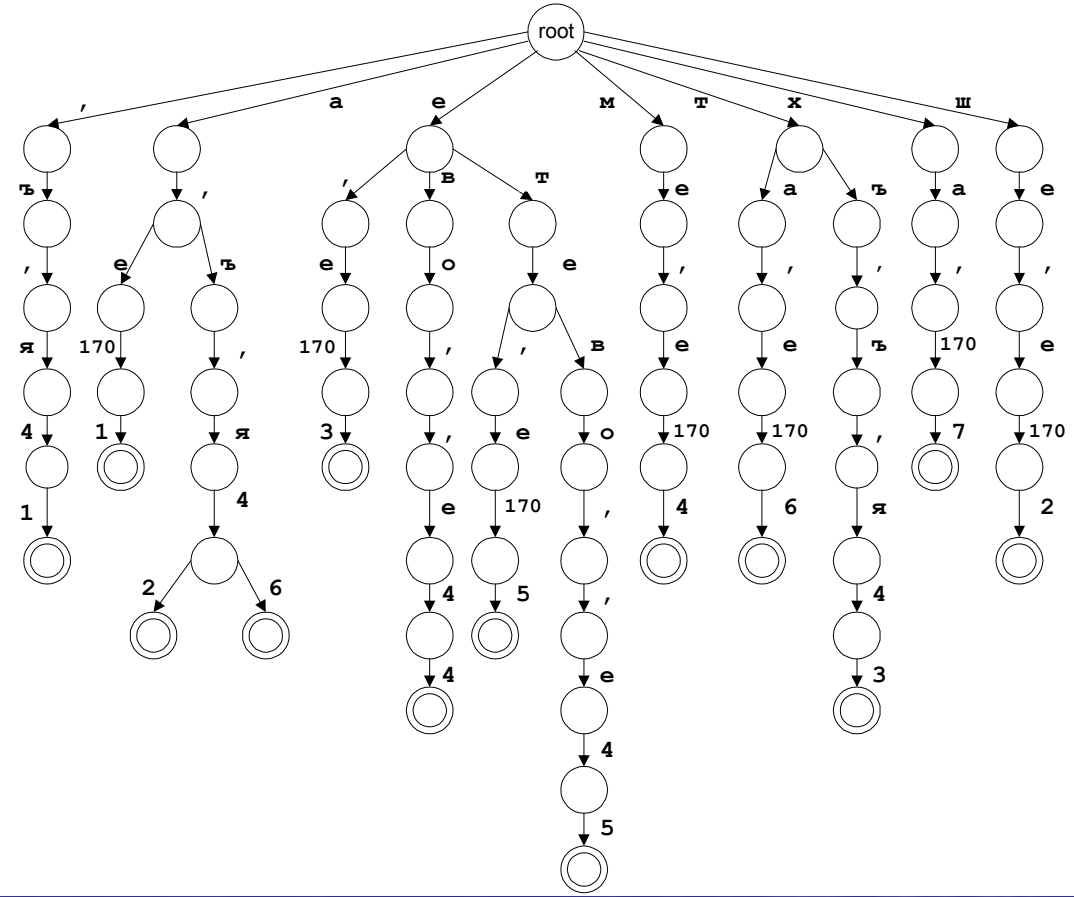
МЯ и МО представени в обратен ред за

‘вятър-вятъра-вятърът-ветрове-…’

‘пера-переш-пере-перем-…’



А. Морфологично ядро



В. Морфологична обвивка

Алгоритъм на конструуиране

- Редуцира изискванията за памет и времето за обработка:
  - чрез конструиране на на минимален trie с постепенно нарастване (шаблон по шаблон, запазвайки минималността)
  - чрез предотвратяване на това да има дърво в паметта
- Минимизацията е постигната чрез отрязване на ненужните данни от trie по следния начин:
  - ако за което и да е състояние всички преходи водят до една и съща анотация, тогава всички състояния и свързаните преходи от тази част не представляват полезна информация и могат да бъдат изтрити (*изтриване*);
  - ако във верига от състояния всичките са свързани едно с друго чрез единствен преход, тогава те всички могат да бъдат представени от едно-единствено състояние (*намаляване*)

Морфологичен анализатор

Процеса на търсена за съвпадение е цикличен. Той се контролира от дължината на възможния суфикс на входната дума започвайки от най-късата:

Стъпка 1. Опитва за съвпадение на възможния суфикс с дадена дължина и създава шаблон използвайки дървото на обвивката

Стъпка 2. Преминаване между дървото на ядрото и на обвивката опитвайки за съвпадение със шаблона (рекурсивно търсене)

Стъпка 3. Ако 2 е успешна връща шаблона и парадигмата

В резултат алгоритъма генерира всички възможни морфологични характеристики на анализираната словоформа.

Морфологична многозначност се среща в 34% от словоформите. Скоростта е висока - повече от *150,000 словоформи в секунда (на 1.6 GHz процесор)*

Експерт и лематизатор

Експерта за анализ (*Guesser*) извършва морфологична класификация (присвоява морфологичен клас) на "непозната" словоформа. Дава и някои предполагаеми граматични характеристики и каква част на речта е

Пример. За "непозната" словоформа 'бера' със същия флективен тип като 'пера' (парадигма 170):

- множеството на морфологичните правила извлечени от дървото на обвивката е {8:2, 2:2, 170:1, 13:2, 57:4, 33:6}
- съответстващото множество от дървото на ядрото е {170}
- резултатното сечение е {170:1}, {глагол, сег. вр., ед. ч., 1 л}

Приближен анализ

Скоростта е сравнима с тази на точния анализ

За оценката, на около 11,600 нови основни форми и съответстващите им 226,730 словоформи (останалата част от БИК) използвахме:

- 4,964 думи – съвпадат със словоформите от речника на прототипа на BulMorph 2.0 и са погрешно идентифицирани при анализа
- за други 33,724 думи – анализът е неуспешен (при действителна реализация приближения анализ се счита за надежден ако съответната форма на думата суфикс с дължина >3, приета от bPFSA прототипа)
- от останалите 188,042 словоформи – грешно интерпретирани са граматичните характеристики на 12,599 думи
- за 69,438 думи – са дадени алтернативни предложения включително и верните
- за всяка от останалите 106,015 словоформи са определени точно граматичните характеристики

Приближен анализ (точност и пълнота)

$$P = \frac{4{,}964 + 69{,}438 + 106{,}015}{226{,}730 - 33{,}724} 100\% = 93.5\%$$

$$R = \frac{4{,}964 + 69{,}438 + 106{,}015}{226{,}730} 100\% = 79.6\%$$

Компютърно моделиране на българското словообразуване

- BulMorph 2.0 предоставя средства за компютърно моделиране на словообразуването
- позволява извличането на трансформационни правила
- от 6,000 произвеждащи основи и над 30,000 техни производни са получени 9,400 трансформационни правила. От тях само 10 правила се срещат повече от 100 пъти, не повече от 300 – между 10 и 100 пъти, и над 6,800 правила – само по един път
- получените правила са приложени върху компютърен речник на БЕ от 70,000 основни форми

Трансформационни правила

Всяко трансформационно правило се представя от няколко елемента:

- тип на базовата основа (ядрото на словообразувателната парадигма)
- тип на резултатната основа
- функционално преобразуване на базовата до произвеждащата основа
- префикс и суфикс, добавяни към произвеждащата основа

Пример за трансформационно правило

Пример 1. От двойката топъл – позатопли последователно се получават:

а) двойки (шаблон, номер на морфологичен клас): (топ*л,80) – (позатопл,173) и

б) трансформация <80,–*л,+поза,+л,173>, т.е шаблон позатопл с номер 173 се получава от шаблон топ*л с номер 80 след „отсичане" на суфикс *л и „добавяне" на префикс поза и суфикс л към шаблон топ*л.

Практическо приложение

Системата от правила е приложена върху речник на БЕ от над 70,000 основни форми

За всяка основна форма се търси и прилага трансформационно правило, което "води" до друга дума от речника с дължина не по-голяма от изходната.

При търсенето броят на възможните правила се редуцира от факта, че приложимите правила са <Nom,–X,+Y,+Z,Morph>, където Nom е морфологичния клас на които принадлежи думата а X е суфикс на думата

Пример 2. След обработка на изходния речник и откриване на съответните трансформационни правила, за думите абен, абаджийски, абаджийство, абаджия се „открива" основа аба, а за абаджийски и абаджийство – абаджия, т.е. на практика е синтезирано гнездото на аба. Случва се непроизвеждащата основа да бъде и грешно определена (баница – банка).

Трансформационни правила

Предимства:

- създаване на нови (в редица случаи и нестандартни) лексикални ресурси за БЕ и софтуерни средства за анализ на български компютърни текстове
- възможност за анализ на „непознати“ думи (не включени в изходния речник) и разпознаване със сравнително голям процент на точност и пълнота на техните морфологични характеристики и семантика
- възможност за оптимизиране на всеки компютърен речник на БЕ чрез определяне на неговото “лексикално ядро” (непроизводни основи) и на съответните трансформационни правила

Заключение

Системата BulMorph 2.0:

- съдържа речник с повече от 80,000 основни форми и съответните им 1,730,000 словоформи
- речникът е компактен, което е важно за PC базирани системи
- скоростта на анализ е много голяма (повече от 150,000 думи в секунда на 1.6 GHz процесор)
- трансформационните правила дават възможност за анализ на "непознати" думи