1. Морфология

От гръцки: morphe (форма, вид) + logos (знание, слово)
Дял от граматиката, в който се изучават преди всичко формите на думите, правилата за тяхното образуване и свързаните с различните форми на една дума граматически значения.

1.1. Що е дума?

Трудности за дефиниране:

- единството на думата като цялост, образувана от отделни структурни елементи, има относителен и условен характер: например: *сбогом = с + богом*, *помежду* (е било *по между*). Така че **правописният критерий** - низ между две шпации - е променлив и несигурен;
- семантичен подход: "Думата е един от най-малките напълно самостоятелни късове изолиран смисъл, към който се свежда изречението" (Е. Сепир). Проблеми: предлозите и частиците нямат самостоятелно значение.
- А. Мейе "Думата е резултат от свързване на определен смисъл с даден звуков комплекс, пригоден за определена граматическа употреба".

Думата е единица от ЕЕ (написана или произнесена) и означава нещо. Ще използваме и други, по-точни категории:

1.2. Форми, словоформи

Съществуват различни (морфологични) варианти на думите: град, града, градът, градове и т.н. Това са словоформи. Следователно от морфологична гледна точка, *изменяемата дума е комплекс от всички свои словоформи*:

Дума Д1: дърво Ф1
дървото Ф2
дървета Ф3, дървеса Ф4, дърва Ф5
дърветата Ф6 дървесата Ф7 дървата Ф8

Една от формите се приема за **основна** (но това не значи, че е по-важна). (**citation form**). Като абстрактна езикова единица, особено във връзка с речници или лексикони, основната форма се нарича **лексема**. Съвкупността от морфологичните варианти на изменяема дума се нарича **парадигма**. При неизменяемите думи се счита, че те имат само по една единствена форма (словоформа).

1.3. Граматически категории

Абстрактни граматически понятия за описание на обекти и техни свойства: род, число, падеж, част на речта и т.н.

1.4. Строеж на думата

Морфема - значеща част от думата, която не може да се разложи на по-малки значещи части.

1.5. Видове морфеми

Корен: Главната морфема в състава на думата, която се явява като основен носител на нейното лексикално значение. **Сродни** думи (с еднакъв корен) образуват **гнездо**.

(вода -> воденица -> воденичар -> воденичарка)

Афикси (наставки и представки, съотв. префикси и суфикси): Морфеми, които се поставят преди корена (**представки**) или след корена и евентуално други морфеми (**наставки**), за да образуват нови думи.

пия -> препия -> изпия -> напия -> отпия -> допия -> опия
-> попрепия -> понапия ->

Окончание: Морфема, с която от дадена форма се образува друга.

град -> град-ове; пия - пиеш - пие - пием - пиете - пият

Определителен член: не е наставка, окончание или частица.

комунизъм -> комунизмът

Съединителни морфеми: особен вид структурни единици, изпълняват свързваща роля в сложни думи, без да имат самостоятелно значение (-е-, -о-, -и-): вод-о-пад, грозд-о-бер, рус-о-кос, зем-е-делец, пет-и-летка.

1.6. Основа на думата

Онази част на думата, която остава след отстраняването на заднопоставения член (определителен или неопределителен) и окончанието, ако има такова: написа-хме, чете-ш, преподавател-и-те.

От *семантична* гледна точка основата е носител на лексикалното значение. От *структурна* гледна точка основа = корен + афикси. От *словоизменителна* гледна точка, основа е онази част на думата, от която е образувана всяка нейна форма (преподава-м, преподава-ш, преподава-ме, преподава-те, ...).

1.7. Допълнителни понятия

Синтетизъм - аналитизъм: *начин, по който се изразяват синтактичните отношения на имена и местоимения към други думи в изречението или словосъчетанието (падежи)*. Говорим за синтетичен/ аналитичен език или

граматически строй. Повечето славянски езици са синтетични (вкл. старобългарският); новобългарският е предимно аналитичен:

- липсват падежи (само остатъци) и се използват предлози;
- аналитично се изразяват инфинитивите;
- аналитични форми са и степенуваните (по-хубав, най-хубав);
- много глаголни форми са аналитични.

Флексия - начин за образуване на (много) нови форми чрез промяна на окончанието на основната форма. *Флективен език*. Българският е *по-флективен* от английския.

2. Компютърна морфология

• Проблеми по структуриране на данните - да се реши как ще се организират. Словоизменителна или словообразувателна морфология? Тоест, първо трябва да дефинираме понятията. Обаче е важно речникът да се организира "по думи", т.е. всяко значение да е записано отделно. Например лесно можем да си спестим записването на наречията в речника, защото те съвпадат с форми на прилагателни в среден род. Но така губим възможност да дефинираме съответно значение (а както видяхме при "Разбиране на ЕЕ", значението е предикат от определен вид; така че в лексикона на системата трябва да имаме отделно място да го асоцииране към лексикалната му форма).

• Проблеми по дефиниране на операциите *анализ* и *синтез*. **Анализ:** разпознаване на всяка форма от парадигмата като морфологичен вариант на основната форма с дадени стойности на съответните граматически характеристики; **Синтез:** по зададена основна форма и допустим набор граматически характеристики, произвеждане на нужната форма.

• Имплементационни проблеми с обема данни.

Практически подходи при дефиниране на морфологични данни

Пример за Съществителни имена

Съществителните имена на български език притежават следните граматически категории: *род* (свойство на основата), *число* и *определеност* (членуване). **Основна форма:** ед. число, нечленувано. Лингвистичните данни са както следва:

РОД: мъжки, женски, среден

Завършват на:

- мъжки род: съгласен звук (ден, град, мъж, народ, език); гласен звук - а, я, о, е, когато означават лица от мъжки род: баща, съдия, татко, аташе.

- женски род: гласен звук - а, я - (жена, земя); съгласен звук - вечер, младост, полет.

- среден род: гласен звук: о, е, при чужди думи и, у, ю: село, дете, такси, бижу, меню.

Очевидно родът не може да се разпознава автоматично по последната буква, дори и да знаем, че дадена дума е съществително. Трябва да се намери друг начин за организация на данните - *а именно, дефиниция в лексикон.*

ЧИСЛО: единствено, множествено

Образуване на множествено число за мъжки род става с окончанията:

•• -И: *народ - народи, гост - гости, зъб - зъби*.
Възможни промени при прибавяне на окончанието:
- палатализация (редуване) на звуковете К, Г, Х пред окончанието -И в Ц, З, С: *език - езици, подлог - подлози, кожух - кожуси*.
- изпадане на -Е: *чужденец - чужденци*.
- изпадане на -Ъ: *театър - театри*.
- изпадане на -Е и вмъкване на -Ъ пред Л или Р: *беглец - бегълци*.
- изпадане на -ИН: *българин - българи*.
- промяна на -Е след гласна в Й: *боец - бойци*.
- комбинация от горните ефекти: например изпадане на -Ъ и палатализация: *момък - момци*.

•• -ОВЕ: *град - градове, стол - столове, блок - блокове, вятър - ветрове*.
Възможни промени при прибавяне на окончанието:
- промяна на мястото на ударението: *град - градове, стол - столове*.
- преглас на -Я в -Е: *бряг - брегове*.
- промяна на мястото на -Ъ в групата -РЪ: *връх - върхове*.
- смекчаване на предходната съгласна: *зет - зетьове*.
- комбинация от горните ефекти: например изпадане на -Ъ и преглас Я-Е или смекчаване на предходната съгласна: *вятър - ветрове, огън - огньове.*

•• -ЕВЕ: *брой - броеве.*

•• -Е: *мъж - мъже, кон, цар, крал, княз.*

•• -А: *крак - крака, рог, лист.* (подобно е *господин - господа* с изпадане на *-ин)*.

•• -Я: само едно съществително *брат - братя.*

•• -ИЩА: *път - пътища, сън, край - краища.*

•• -ОВЦИ: *дядо - дядовци, чичовци, мързеланковци.*

•• -ОВЦЕ: *градец - градовце.*

Внимателно разглеждане на горната схема показва колко е несъвършенна. Тази класификация има нужда от допълнително прецизиране, ако искаме да я приложим в програма. Обаче има и други лингвистични данни:

•• бройна форма за неодушевени *(стол - два стола - столове).* При лица се предпочита множественото число *(двама студенти).* Пак е налице изпадане: театър - театри - два театъра; прозорец - прозорци - два прозореца. Наблюдаваме и омонимия:

метър - метри - два метра/два метъра в друг смисъл на "метър",
литър - литри - два литра/два литъра в друг смисъл на "литър".
път - пътища - два пътя
път - пъти (колко пъти ти казвам ... два пъти)

Стигаме до първа приблизителна структура на **флективни типове** за съществителни от мъжки род (флективен тип - набор окончания):

No. тип	Основна форма	Членуване	Мн. число	Членуване	Бройна форма
1	град	града, -ът	градове	градовете	гра'да
2	бряг	брега', -ът	брегове	браговете	бря'га
3	връх	върха', -ът	върхове	върховете	вър'ха
4	вятър	вятъра,-ът	ветрове	ветровете	вятъра
5	клон	клона, -ът	клонове (клони)	клоновете (клоните)	клона
6	лист	листа, -ът	листове, листа, листи, листя	листовете, листата, листите, листята	листа
7	народ	народа, -ът	народи	народите	народа
8	чужденец	-а, -ът	чужденци	чужденците	- - - - - -
9	театър	театъра, -ът	театри	театрите	театъра
10	метър	-а, -ът	метри	метрите	метра
11	организъм	организм-а,-ът	организми	-измите	органи-зъма
12	грък	гърка', -ът	гърци	гърците	- - - - - -
13	боец	боеца, -ът	бойци	бойците	- - - - - -
14	език	езика, -ът	езици	езиците	езика
15	подлог	подлога, -ът	подлози	подлозите	подлога
16	кожух	кожуха, -ът	кожуси	кожусите	кожуха
17	камък	камъка, -ът	камъни	камъните	камъка

18	българин	българина, -ът	българи	българите	- - - - - - -
19	момък	момъка, -ът	момци	момците	- - - - - - -
20	беглец	беглеца, -ът	бегълци	бегълците	- - - - - - -
21	турчин	турчина, -ът	турци	турците	- - - - - - -
22	крак	крака, -ът	крака	краката	крака
23	господин	господина, -ът	господа	господата	- - - - - -
24	мъж	мъжа, -ът	мъже	мъжете	- - - - - -
25	брат	брата, -ът	братя	братята	брате
26	съд	съда, -ът	съдилища	съдилищата	съда
27	градец	градеца, -ът	градовце	градовцете	градеца
28	брой	броя, -ят	броеве	броевете	броя
29	зет	зетя, -ят	зетьове	зетьовете	----------
30	огън	огъня, -ят	огньове	огньовете	огъня
31	другар	другаря, -ят	другари	другарите	----------
32	герой	героя, -ят	герои	героите	----------
33	ден	деня, -ят	дни	дните	дена, дни
34	нокът	нокътя, -ят	нокти	ноктите	нокътя
35	кон	коня, -ят	коне	конете	коня
36	край	края, -ят	краища	краищата	края
37	път	пътя, -ят	пътища	пътищата	пътя
38	баща	бащата	бащи	бащите	----------
39	съдия	съдията	съдии	съдиите	----------
40	дядо	дядото	дядовци	дядовците	----------

Звателни форми:

N o . тип	Основна форма	Звателна форма
7	народ	народе
8	чужденец	чужденецо
12	грък	гър'ко
13	боец	боецо
18	българин	българино
19	момък	момко
20	беглец	беглецо
21	турчин	турчино
23	господин	господине
24	мъж	мъжо

25	брат	брате
29	зет	зетко, зетьо
31	другар	другарю
32	герой	герою
35	кон	коньо

Една идея за имплементацията на компютърна морфология е да организираме данните както следва. Тук се дава пример за съществителни от мъжки род на български език:

основа (до 3 основи, плюс номер на флек-тивен тип)	оконч. осн. форма	оконч. кратък член	оконч. пълен член	оконч. мн. ч. нечл.	оконч. мн. ч. член.	оконч. бройна форма	оконч. зват. форма
1а. геро	-й	-я	-ят	-и	-ите	няма	-ю
2а. бряг 2б. брег	0 -	- -а	- -ът	- -ове	- -овете	-а -	- -о??
3а. момък 3б. момк 3в. момц	0 - -	-а - -	-ът - -	- - -и	- - -ите	няма	- -о -
4а. грък 4б. гърк 4в. гърц	0 - -	- -а -	- -ът -	- - -и	- - -ите	няма	- -о -
5. народ	0	-а	-ът	-и	-ите	-а	-е

И по подобен начин за всички съществителни, прилагателни, глаголи и числителни.

Литература:

1. Академична граматика, том 2 Морфология.
2. Боримир Кръстев, Българския език в таблици и склонения. София, Наука и изкуство, 1984.
3. Елена Паскалева - многобройни публикации.

2. ЛЕКСИКОНИ

Terminological Banks (termabnks)
Machine-Readable Dictionary
Computer Dictionary
Lexicon

Четири термина, с които означаваме четири вида езикови ресурси - с различна степен на детайлност и ниво на вътрешна организация.

Термбанки:
- Eurodicautom (10 млн. термини на 5 езика),
- Банката на Канадското правителство

Компютърни речници: Webster

Разлики: във функциите, класа потребители, труда за подготовка и цената.

За сведение можем да посочим цени от рекламен материал за немско-английската лексика в системата за машинен превод METAL. Лексиконът във вътрешно представяне съдържа прецизно дефинирани признакови структури за всяка основна дума, както и информация за съчетаемостта на думите в изреченията и семантични филтри. Поради това такива лексикони са извънредно скъпи.

Лексиконите са разделени по предметни области, като са посочени двадесет тематики: селско стопанство, медицина, металообработка, оптика, социални науки, естествени науки, администрация, икономика, транспорт, спорт и други по-долу. Според броя единици на терминологичните речници, можем да подредим офертата както следва:

Тематична област:	Брой единици:	Цена в долари:
Data Processing	45 000	35 000
Telecommunications	26 000	27 000
Technology	21 000	22 000
Electronics	17 000	20 000
Natural Sciences	16 000	18 600
Chemistry	8 000	10 000
Mechanical Engineering	5 800	7 500
Textile Industry	5 300	7 000

Лексиконите за всяка от останалите 12 предметни области съдържат под 5 000 единици. Сумарно, 176 400 терминологични единици се предлагат за 188 800 долара.

За сравнение, Оксфордският речник на английския език върху CD струва под 1000 лири. Този речник може да се определи като пренесено върху магнител носител книжно тяло от текстове, допълнено с интелигентни средства за търсене и бърз достъп. Но лексиконите на системите за МП съдържат много повече информация към речниковите единици, която показва морфологични, синтактични и семантични характеристики в рамките на избрания модел за представяне на лингвистичното знание.

Разликата в цената на двата вида речници е очевидна. Но при цените на труда в развитите страни, е по-евтино да се закупи готов речник, тъй като за две човекогодини в никакъв случай не могат да се съберат и организират 176 400 терминологични единици. Така лингвистичните ресурси са също така стока, както и данните за други видове системи, и са обект на пазарни продажби.