Крушков, Хр., Хр. Христов. "Многоплатформена система за автоматичен морфологичен синтез и анализ на естествен език", Научни трудове на СУБ, Серия В – Техника и технологии, т. VII, Пловдив, 2008 г., стр.279-284, ISSN 1311-9419

# МНОГОПЛАТФОРМЕНА СИСТЕМА ЗА АВТОМАТИЧЕН МОРФОЛОГИЧЕН СИНТЕЗ И АНАЛИЗ НА ЕСТЕСТВЕН ЕЗИК

Христо Димитров Крушков
катедра "Компютърна Информатика"
ФМИ при ПУ "Паисий Хилендарски"

Христо Тошков Христов

# MULTIPLATFORM SYSTEM FOR AUTOMATIC MORPHOLOGICAL SYNTHESIS AND ANALYSIS OF NATURAL LANGUAGE

Hristo Dimitrov Krushkov
Department of Computer Science
FMI, University of Plovdiv

Hristo Toshkov Hristov

**Abstract**

*In this paper some theoretical and algorithmic problems of the automatic morphological synthesis and analysis are discussed and generalized. A model using a new morphological classification is described. A multiplatform system for automatic morphological synthesis and analysis of natural language has been developed based on this model.*

***Keywords****: Natural Language Processing, Computational Linguistic, Morphological Analysis.*

## 1. Въведение.

Текстът и речта са основни обекти, които са предмет на изследване от страна на компютърната лингвистика. Обособени направления в дисциплината са: автоматичен анализ на текст; генериране на текст; анализ и синтез на реч. Мястото на компютърната лингвистика в пространството на съвременната наука е трудно да се определи, предвид, че в него общуват, от една страна, фундаменти от областта на математиката и информатиката, а от друга страна, науки като фонетика, морфология, синтаксис, семантика, семиотика и др. В настоящата разработка се разглеждат въпроси, които се отнасят към първите две направления.

За да се осъществява диалог между естествен език и машина (съвременен компютър), е необходим формален модел, върху който да се построи комуникативен интерфейс. Най-ниското ниво на такъв модел е морфологично, т.е. морфологичен процесор, който да извърши автоматичен морфологичен синтез и анализ. Ключови обекти, с които процесорът работи, са: парадигма от словоформи; шаблон на парадигма; правила и морфеми на словоизменение; типове словоизменения; морфологични категории и др. В катедрата по Компютърна Информатика при ФМИ на ПУ са решени в голяма степен засяганите проблеми. Изграден е морфологичен речник от 70 000 основни форми, произвеждащи над 1 400 000 словоформи с техните граматични характеристики. Част от постигнатите резултати при автоматичния морфологичен анализ са достъпни чрез http://www.uni-plovdiv.bg/dcs/morph.htm.

В статията се цели да се обобщят основни проблеми при морфологичния синтез и анализ, да се създадат методи и средства за преодоляването им и да се представи интегрирано реализацията им в многоплатформена система. Новост в разработката е предложеният различен вариант на морфологична класификация от направените до момента.

## 2. Проблеми и постановка.

Проблемите пред които се изправя реализацията на многоплатформена система за автоматичен морфологичен синтез и анализ, условно може да се разделят на два вида – теоретични(формални), такива които са фундаментални за изграждане на системата, и алгоритмични – такива, които са необходими за работа на системата.

- Теоретични:
  - Проектиране на формален модел на морфологичен процесор;
  - Построение на формален модел на морфологична класификация (необходима за анализа);
  - Изграждане на формални модели на автоматичен морфологичен синтези и анализ
- Алгоритмични:
  - Образуване на шаблон от парадигма;
  - Извличане на словоизменителни морфеми, съставяне на словоизменителни правила;
  - Съставяне на словоизменителни типове;
  - Разпознаване на шаблон от словоформа;
  - Генериране на парадигма посредством шаблон и тип словоизменение (морфологичен синтез);
  - Свързване на словоформа с морфологични категории (морфологичен анализ).

2.1. Образуване на шаблон от парадигма и извличане на словоизменителни морфеми.

В граматиката под парадигма се разбира съвкупността от словоформи на лексемата, придружени със съответните правила на словоизменение. При наличие на цялата парадигма на дадена лексема (всички нейни форми), шаблонът може да се получи автоматично, т.е. добиването може да се алгоритмизира, да се интерпретира от машина. Общият алгоритъм за образуване на шаблон е разгледан в [1], като при софтуерна реализация се оказва подходящо граматичните правила да се категоризират до няколко вида, което улеснява съставянето и работата в двете посоки - образуване на шаблон и синтезиране на парадигма.

2.2. Структури и класификация.

А) Хеш таблица от типове словоизменения.

Една от структурите данни, която е базисна за работа на системата, е хеш таблица с типове словоизменения, въз основа на която се извършва автоматичен морфологичен синтез. За синтеза е достатъчно да имаме два обекта – шаблон на парадигма и морфемите, от които се генерират словоформите. Става ясно, че за да се извърши автоматичен морфологичен синтез, е нужно да се съхраняват в база от данни посочените обекти. Удобно е морфемите да се запазят в хеш таблица, съставена от две колони, едната колона да е ключ – целочислен идентификатор, а другата колона да е символен низ от морфеми. В символния низ влизат и някои метазнакове, които допълнително улесняват обработката. Тези знакове служат за извличане на допълнително информация и за разпознаване на отделните морфеми в низа. Освен това, ако две или повече парадигми имат еднакъв символен низ от морфеми, то те спадат към един и същи *тип словоизменение*. По този начин се получават неголям брой типове, които се отнасят за цялата лексика на езика.

Б) Морфологична класификация.

По форма думите се делят на изменяеми и неизменяеми. Изменяеми са тези части на речта, които менят формата си, чрез замяна на морфеми посредством някое установено граматично правило, така, че да се измени поне една граматична категория. Такива части са съществителните имена, прилагателните имена, местоименията, числителните имена и глаголите. Неизменяемите части на речта не менят формата си. Такива части са наречията, предлозите, съюзите, междуметията и частиците. В отделни части на речта се разглеждат граматични категории като род, число, определеност при имената; лице, време, вид, залог, наклонение при глаголите и т.н.

**Категории граматичен клас, граматичен подклас и граматична група**

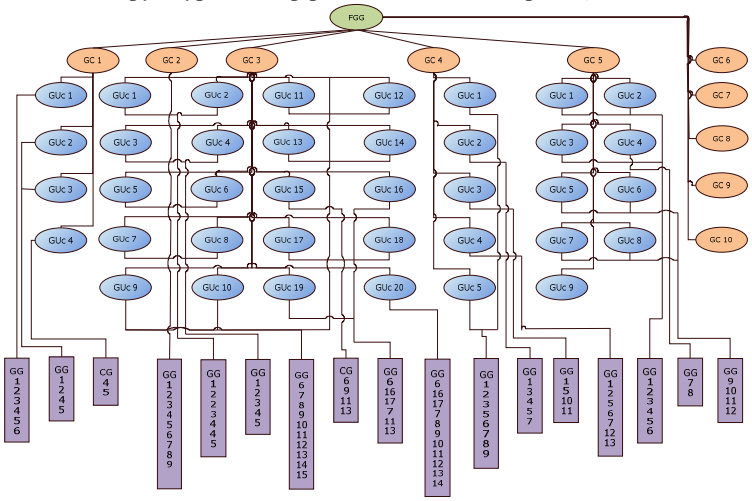
Граматичните класове представят главните части на речта – съществително име, прилагателно име, числително име, местоимение, глагол, наречие, предлог, съюз, междуметие и частица. Словоформите в парадигмите, притежаващи общи за главната част граматични категории, са наречени *граматични групи в парадигмата.* Цялата морфологична категоризация се разделя на три части: *граматичен клас* – характеризира думи, които се категоризират като главни части на речта, *граматичен подклас* – характеризира думи, които спадат към класа, но притежават поне една различна категория от останалите подкласове в рамките на класа, и *граматични групи в парадигма* – характеризира отделните словофрми и носят общ характер за класа, независимо от граматичния подклас.

Подробно описание на категориите е представено в [2].

Категориите от езика се систематизират в силно разклонена графовидна структура от четири, условно наричани, нива. Първо и второ ниво (оранжевият и синият цвят) представляват съответно граматичен клас - GC и граматичен подклас - GUc. Ниво три се

състои от граматични групи - GG. Всяка морфологична форма, клас, подклас и група се разпознава чрез идентификационен номер. Конкретните граматични групи зависят от две числа – номера на класа и номера на групата. Идентификационните номера на граматичните групи и поредният номер на словоформите в парадигмата в някои от случаите съвпадат. Граматичните подкласове също са зависими от две числа – номера на класа и номера на подкласа. Граматичните класове са зависими единствено от своя идентификационен номер.

*Фиг.№1 Структура на морфологичната класификация.*




И така, морфологичните категории могат да се представят под формата на структура така, че характеристиките на думите да се получават по формулата *FGG = GC + GUc + GG*. При това, за да се определи *FGG,* е достатъчно да се запаметят *GC* и *GUc*, докато информация за *GG* ще се даде от поредния номер в парадигмата. Структурата е статична и от гледна точка на съвременните програмни езици съществуват различни подходи за реализирането ѝ.

В) Колекция данни

Колекцията данни е третата структура, която прави връзка между хеш таблицата на типове словоизменения, морфологичната класификация и шаблоните на парадигми. Тя се състои от морфологични елементи, всеки от които съдържа четири полета - шаблон на парадигма, номер на тип словоизменение, номер на граматичен клас и номер на граматичен подклас. Организацията на данните може да се разглежда като шаблонен речник на парадигмите от гледна точка на информатиката и като морфологичен речник от граматично гледище. Морфологичните елементите в колекцията се свързват с типа словоизменение от хеш таблицата, чрез номера на типа и с морфологичната структура чрез номерата на граматичният клас и граматичният подклас.

## 3. Решения.

Типично информатично представяне на решение е спазване на подход, при който информацията се разглежда от абстрактен апарат в три основни състояния – входно, състояние в момент на обработка и изходно състояние. Формалните модели, които са

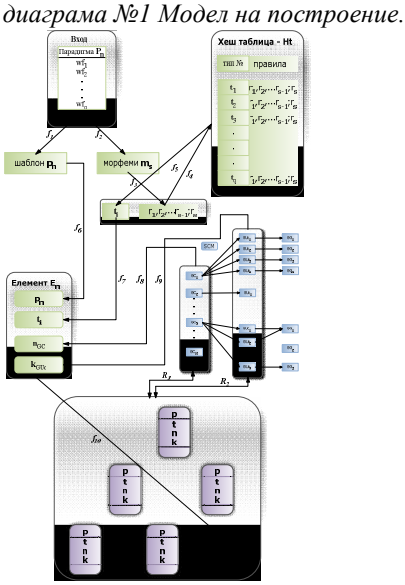
показани, спазват такъв подход и се разглеждат интегрирано като система, която е изградена от три модула:

- модел за построение на колекция данни(морфологичен процесор и речник);
- модел на автоматичен морфологичен синтез;
- модел на автоматичен морфологичен анализ.

В архитектурата на системата има три базови структури данни – колекция от морфологични елементи, структура на морфологична класификация и таблица на типове словоизменения. Морфологичната структура е изградена предварително, отделно от колекцията. Тя се използва, за да се категоризират словоформите в парадигмата, и участва в процеса на анализ. Таблицата с типове словоизменения се изгражда успоредно с построяването на колекцията. Всеки елемент в колекцията е свързан с таблицата и със структурата на морфологична класификация чрез целочислени идентификатори – номер на тип(връзка към таблицата с типовете словоизменения), номер на граматичен клас и номер на граматичен подклас(връзки с морфологичната класификация). На практика проблемите, стоящи при модела на построение (морфологичен процесор и речник), са успешно решени в[3], като в настоящата разработка се предлага конкретен вариант с нововъведение – реализация на платформена независимост на модела. Системата се разработва с език за програмиране Java (jdk 1.6.1.) и среда на работа NetBeans IDE 6.1

### 3.1. Построение на колекция данни.

*диаграма №1 Модел на построение.*




**Тълкование на обработката:**

На входа на системата се получават словоформите на парадигма, образува се шаблонът и се извличат словоизменителните морфеми. Лексикограф избира граматичния клас и граматичния подклас. Парадигмите, които са на входа, се вземат от файл. И така, на входа е парадигмата $P_n$ от словоформи $wf_1$ , $wf_2$ , . . . , $wf_q$. Функцията $f_1$ използва алгоритъма за автоматично получаване на шаблон, чрез който от парадигмата $P_n$ се добива шаблонът $p_n$.

Функцията $f_2$ извлича морфемите $m_s$, които са променливи, и ги записва в символен низ. От морфемите, чрез функцията $f_3$ се образуват правилата на словоизменение $r_1$, $r_2$, . . . $r_{s-1}$ ; $r_s$. Образуваните правила съставляват един тип словоизменение. Чрез $f_4$ се добавя типът в хеш таблицата и се генерира негов номер. При добавянето съществуват две възможности: да има такъв тип в хеш таблицата или да няма такъв тип. Ако типът го има в таблицата, чрез $f_5$ се записва неговият номер в $t_i$, ако типът го няма, то се генерира следващ по ред номер на тип, добавя се типът и отново, чрез $f_5$ се записва номерът в $t_i$. Посредством функциите $f_6$ , $f_7$ , $f_8$ , $f_9$ и участието на лексикограф следва добавянето на шаблона $p_n$, номера на типа $t_i$, номера на граматичния клас $n_{GC}$ и номера на граматичен подклас - $k_{GU}$. Ролята на лексикографа е да свърже граматичния клас $GC$ и граматичния подклас $GUc$ с шаблона $p_n$ посредством техните номера. Накрая остава така съставеният морфологичен елемент да се добави в колекцията.

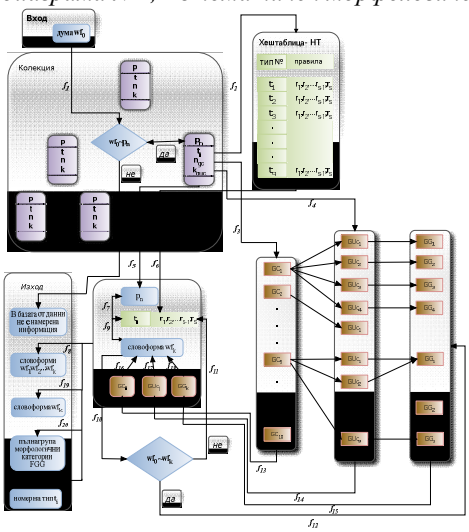
Ако два шаблона съвпадат, т.е. има шаблонна омонимия, новият шаблон се добавя като следващ по ред. В последствие при синтеза и анализа следва да се изведе информация за омонимия, като се изпишат случаите на еднаквост.

### 3.2. Автоматичен морфологичен синтез и анализ.

В тълковните речници синтезът се определя като метод на изследване, който свързва отделните елементи в едно цяло и по този начин изучава същността на явленията. В настоящата разработка под синтез се разбира свързването на шаблона и типа словоизменение и генерирането на парадигмата на произволна дума от колекцията данни. Резултат на синтеза са словоформите на парадигмата.

Общото понятие анализ се дефинира като метод на научно изследване, чрез който разглежданият предмет се разчленява на съставящите го части. В работата под морфологичен анализ се разбира категоризацията на думите от българския език, т.е. според формата на думата ще се извеждат нейните морфологични категории.

*диаграма №2, Автоматичен морфологичен синтез и анализ.*



**Тълкование на обработката при синтеза:**

На входа на системата постъпва произволна словоформа $wf_0$. Функцията $f_1$ претърсва колекцията, за да открие подходящия шаблон, който отговаря на входната словоформа. Ако не е открит шаблон, то се извежда съобщение, че в базата от данни не е намерена информация за въведената словоформа. В противен случай, ако е намерен търсеният шаблон, то от морфологичния елемент се извеждат нужните данни, за да се извърши синтезът. Следват следните действия: Чрез номера на типа $t_i$ функцията $f_2$ свързва морфологичния елемент с хеш таблицата $HT$. Чрез $f_5$ и $f_6$ се извлича отделно от структурите съответно шаблонът $p_n$ и словоизменителните морфеми, които се преобразуват в правила $r_1$, $r_2$, . . . $r_{s-1}$ ; $r_s$ на типа $t_i$. Правилата с индекси от $1$ до $s-1$ са правила на замяна, а правило с индекс $s$ е, за да се добави окончание. Функцията $f_7$ свързва шаблона $p_n$ с типа на словоизменение $t_i$ и се прилагат последователно правилата $r_1$, $r_2$, . . . $r_{s-1}$ ; $r_s$., . По този начин се генерират последователно словоформите на парадигмата. Накрая чрез функциите $f_8$ и $f_{20}$ се извеждат словоформите на парадигмата и номерът на типа словоизменение $t_i$ .

**Тълкование на обработката при анализа:**

При морфологичния анализ се повтарят част от операциите от морфологичния синтез, като заедно с това се извършват и следните действия: Чрез $f_3$ и $f_4$ номерата $n_{GC}$ и $k_{CUc}$ на морфологичния елемент от колекцията данни се свързват с граматичния клас $GC_i$ и неговия подклас $GUc_j$ от морфологичната структура. Следва изпълнението на функцията $f_9$, която служи да генерира следващата по ред словоформа. Генерира се първата по ред(основната форма), след което чрез $f_{10}$ се сравняват генерираната словоформа и словоформата от входа. Ако не се получи съвпадение, $f_{11}$ връща отново за изпълнение $f_7$ и $f_9$, докато се стигне до съвпадение. В цикъла, който се описва, има брояч, който пази текущия номер на словоформа и на всяко повторение го увеличава с единица. По този начин се следи за номера на словоформата в парадигмата. При получаване на съвпадение, функцията $f_{12}$ чрез брояча на словоформи свързва морфологичния елемент с група от морфологични категории: $GG_k$ в парадигмата на класа $GC_i$ и неговия подклас $GUc_j$. Словоформа с пореден номер $k$ ще се свърже с граматични категории, които са $k$-ти по номер в парадигмата, т.е. словоформа с пореден номер $k$ за морфологичния елемент, който е свързан с конкретни $GC_i$ и $GUc_j$ ще притежава пълна група граматични категории $FGG = GC_i + GUc_j + GG_k$.

Следва изпълнение на $f_{13}$, $f_{14}$ и $f_{15}$, чрез които се извличат от структурата на морфологичната класификация съответният клас, подклас и група. Посредством функциите $f_{16}$, $f_{17}$ и $f_{18}$ съответно $GC_i$, $GUc_j$ и $GG_k$ се свързват със словоформата $wf_k$. Накрая се извеждат словоформата $wf_k$, пълната група морфологични категории $FGG = GC_i + GUc_j + GG_k$ и номерът на типа словоизменение $t_i$.

Литература:


1. Крушков Хр., Танев Хр., Крушкова М., "Автоматично извличане на шаблони и правила за формообразуване от парадигми", VII Национална конференция "Съвременни тенденции в развитието на фундаменталните и приложни науки", 6-7 юни 1996, Стара Загора стр. 167-171.
2. Христов Хр., „Автоматичен морфологичен синтез и анализ”, ВТУ „Св. Св. Кирил и Методий”, Дипломна работа за придобиване на образователна степен „Магистър”, 2008
3. Крушков Хр., “Моделиране и изграждане на машинни речници и морфологични процесори”, Пловдив, Дисертация за присъждане на образователна и научна степен “Доктор”, 1997