ЮБИЛЕЙНА НАУЧНА СЕСИЯ – 30 години ФМИ
ПУ “Паисий Хилендарски”, Пловдив, 3-4.11.2000

АВТОМАТИЧЕН АНАЛИЗ И СИНТЕЗ НА СЛОЖНИ ГЛАГОЛНИ ФОРМИ

Христо Димитров Крушков, Диан Димитров Ганчев, Мариана Иванова Крушкова

Системата на глагола в българския книжовен език е твърде развита и богата откъм форми и значения. Автоматичната им обработка е необходима предпоставка за извършване на синтактичен анализ на изречението. До момента не е реализиран алгоритъм за такава обработка. Прави се само анализ на отделните лексеми. В настоящата статия е представен формален модел на сложните глаголни форми и алгоритъм за автоматичен анализ на такива форми. Моделът и алгоритъмът са в основата на програмна реализация - модули към Интегрираната лингвистична среда [5]. Модулте съдържат редактор за въвеждане и актуализация на модела на сложните глаголни форми и функции за анализ и синтез на такива форми.

AMS Subject Classification: 68T50

1. Увод

Автоматичната обработка на естествен език в последните години се превърна от чисто научна в научно-приложна област с нарастващ брой комерсиални приложения. Извличането на подходяща информация от огромни текстови корпуси (особено в Интернет), автоматичното резюмиране на документи, граматичната проверка и други лингвистични "екстри" в текстовите редактори и дори автоматичният машинен превод започват да стават ежедневие. Автоматичният лингвистичен анализ неотменно присъствува във всички тези дейности. Обикновено той се извършва на няколко нива, наредени във възходяща последователност - лексическо, морфологично, синтактично, семантично и т.н. Всяко по-високо ниво черпи информация от по-ниското и увеличава прецизността на посочените по-горе дейности. Реализирането на всяко по-горно ниво, обаче изисква повече лингвистични ресурси и средства. Създаването на модел на сложните глаголни форми и средства за техния анализ и синтез на морфологично ниво са една от предпоставките за осъществяване на автоматичен синтактичен анализ.

Лингвистичният анализ на морфологично ниво се осъществява от морфологичен процесор (МП). МП на българския език (БЕ) се построява на базата на общите закономерности, на които се подчиняват морфологиите на флективните езици [1, 2, 6]. В тези МП се анализират отделните лексеми без да се прави анализ на сложните глаголни форми.

Например при анализ на израза ***Щяхме да сме учили***, МП дава следния резултат:

щяхме: Глагол, мин.св/несв.вр.,мн.ч.,1л., осн.ф-ма:ща

да : Частица

сме : Глагол, сег.вр.,мн.ч.,1л., осн.ф-ма:съм

учили : Глагол, мин.св.деят.прич.,мн.ч.,осн.ф-ма:уча

Изследванията представени в този доклад, както и програмната реализация на тяхна база целят преодоляването на този недостатък и получаването на резултата:

Щяхме да сме учили: Глагол, бъд.предв.в миналото,мн.ч.,1л., осн.ф-ма:уча

2. Класификация на сложните глаголни форми.

Най-характерната особеност на новобългарския глагол е това, че той притежава много добре развита система за изразяване на категорията време - форми за девет различни глаголни времена [3, 4]. Сложните глаголни форми могат да бъдат класифицирани в три основни групи :

- Сложни глаголни времена.
- Залог на глагола.
- Наклонение на глагола.

2.1. Сложни глаголни времена

Към сложните глаголни времена спадат бъдеще, бъдеще предварително, бъдеще време в миналото, бъдеще предварително време в миналото, минало неопределено време, минало предварително време. Тези от тях, които се образуват с помощта на причастия имат по три представителя за всяко лице в единствено число. Ако в глаголното време няма причастия (например бъдеще време), то броят на глаголните форми е 6, както при простите глаголни форми:

Ед.ч.	Мн.ч.
Ще работя	Ще работим
Ще работиш	Ще работите
Ще работи	Ще работят

Табл.1 Сложни глаголни форми без причастия.

От друга страна, ако в глаголното време има причастия (например бъдеще предварително време), то броят на глаголните форми става 12:

Ед.ч.			Мн.ч.
Ще съм работил	Ще съм работила	Ще съм работило	Ще сме работили
Ще си работил	Ще си работила	Ще си работило	Ще сте работили
Ще е работил	Ще е работила	Ще е работило	Ще са работили

Табл.2 Сложни глаголни форми с причастия.

За всяко лице и число са възможни още възвратни, отрицателни и въпросителни форми, както и всички техни комбинации (ще мия, ще се мия, няма да мия, ще мия ли, ще се мия ли, няма да се мия, няма ли да се мия , няма ли да мия).

2.2. Сложни форми породени от залога на глагола

В българския език глаголите имат форми за деятелен залог (всички описани по горе форми на глаголни времена) и страдателен залог. Формите за страдателен залог се образуват по два начина:

с възвратната частица ***се*** и съответния глагол (***се пише***). Такива форми се наричат възвратно- страдателни форми.

със спомагателния глагол ***съм*** и минало страдателно причастие на глагола (***е писан***). Наричат се причастно-страдателни форми.

Страдателните форми в единствено число имат варианти за род (се е писала, бе писано, ще бъде писана). Отрицателни форми се образуват с частицата не (не се пише, не е писан). Въпросителни форми се образуват с частицата ли (пише ли се, писан ли е).

2.3. Сложни форми породени от наклонението на глагола

В българския език има 4 наклонения: изявително, преизказно, повелително и условно.

Изявителното наклонение е основно. Всички форми на глагола посочени по-горе, са в изявително наклонение.

Преизказното наклонение е специфично за БЕ. Всички глаголни времена в изявително наклонение имат и форми за преизказно наклонение. Тези форми се образуват с минали деятелни причастия и спомагателния глагол съм (работел съм, били сме работили, щяла съм да работя, щели да са работили).

Повелително наклонение. В общия случай формите за повелително наклонение са прости. Освен тях има и сложни (описателни) форми. Това са формите за сегашно време за всички лица и числа и частиците ***да***, ***нека***, ***нека да***, ***хайде да***, (например: да работим, нека работим, нека да работим, хайде да работим).

Условно наклонение. Формите за условно наклонение се образуват от особени форми на спомагателния глагол съм (бих, би, би, бихме, бихте, биха) и минало свършено деятелно причастие (бих работил, би работила, биха работили).

3. Програмна реализация

Реализирани са модули на DELPHI с които е разширена съществуващата Интегрирана лингвистична среда [5]. Първият от тях служи за въвеждане и редактиране на сложните глаголни времена. Вторият осъществява точен и приближен анализ на такива времена. Третият е добавка към модула за морфологичен синтез, като позволява синтез на сложни глаголни форми. Четвъртият прави морфологичен анализ на текстов файл.

3.1. Въвеждане и редактиране на сложните глаголни времена

Модулът осигурява удобен начин за представянето на сложните глаголни времена, така че те да могат лесно да се съхраняват, редактират и използват. Всяка сложна глаголна форма си има шаблон и идентификатор на времето. В шаблона частиците и спомагателните глаголи се съхраняват като низове, а на мястото на основния глагол стои номера на словоформата му т.е. число в интервала от 1 до 35 [2, 3], което представлява номер на проста глаголна форма в парадигмата от глаголни словоформи. Така напр. шаблонът ***щяхме да сме 22*** показва, че на мястото на цифрата 22 трябва да се постави словоформа номер 22

(минало свършено деятелно причастие - мн.ч.) от парадигмата на съответния глагол, чието бъдеще предварително време в миналото, първо лице, множествено число искаме да формираме. Дефиницията на сложната глаголна форма е следната:

```
Ttabl = record
    pattern: array[0..6]of string[10];
    vreme: string[50];
end;
```

като в pattern стоят отделните лексеми и номерът на словоформата на основния глагол (не-повече от 7). Във vreme се съхранява идентификаторът на това време както и характеристиките на конкретната форма т.е. лице число и род ако има такъв както и дали е възвратна отрицателна или каквато и да е друга форма. Цялата база за сложни глаголни времена представлява файл от такива записи.

При инициализиране на базата е достатъчно да се въведат само основната форма, възвратната, отрицателната и въпросителната форма на това време, като и четирите трябва да са в първо лице единствено число. Останалите форми в първо лице единствено число се генерират автоматично. От възвратната и отрицателната форма се образува възвратно-отрицателна, от възвратна и въпросителна - възвратно-въпросителна, от въпросителната и отрицателната - въпросително-отрицателна, от възвратна и въпросително-отрицателна се образува възвратна-въпросително-отрицателна форма. Самият алгоритъм е следният:

Обхожда се първата форма и се извлича лексемата, непосредствено разположена след частицата ch, която може да бъде "се" или "ли". Резултатната форма се получава от втората форма, като частицата ch се вмъква непосредствено пред извлечената от първата форма лексема. Ако ch е последна в първата форма тя остава последна и в резултатната форма.

Например за бъдеще предварително време възвратната форма е *Ще съм се* ***къпал***, а отрицателната *Няма да съм* ***къпал***. Съответно отрицателно-възвратната е *Няма да съм* ***се къпал***

След като вече са открити осемте форми в първо лице единствено число, останалите форми се получават от тях чрез спрягане на основния и спомагателните глаголи (ако има), като се взима предвид, че ако основният глагол е в причастна форма, то формите са 12, а иначе са 6 (табл. 2 и 1).

3.2. Точен и приближен анализ на сложни глаголни времена

Модулът осигурява средства за извършване на анализ от системата на израз въведен от потребителя въз основа на съществуващите в базата данни глаголни форми. Анализът може да бъде точен или приближен, като при точният анализ се търси точно съвпадение на израза с глаголна форма от базата данни, а при приближения анализ се търсят глаголни форми, състоящи се от лексеми, които се съдържат и то в същата последователност в израза въведен от потребителя, но в този случаи не е необходимо всички лексеми от израза да участват в глаголната форма.

При точния анализ за анализирания израз се търси подходящ шаблон от базата данни. При намирането на такъв шаблон, се анализира лексемата от израза, която съответства на число от шаблона. Ако тя е проста глаголна форма (номер на таблица от 142 до 187) и

номерът на словоформата й е равен на числото от шаблона [2], от базата се извлича идентификаторът на време за този шаблон и се присвоява на анализирания израз.

При приближения анализ от израза се отделят по ред всички лексеми, които са глаголи или някоя от частиците: *да*, *не*, *няма*, *ли*, *се*, *нека* и *хайде*. След това се прилага точен анализ.

4. Резултати и перспективи

При точния анализ се разпознават правилно всички сложни глаголни форми, за които основният глагол е в речника. При приближения анализ проблеми възникват единствено при анализа на текстов файл при определяне на границите на сложните форми, ако липсват знакове за разделяне (словесни или пунктуационни). Перспективите са свързани с разработването на краен автомат за ускоряване на анализа на сложните глаголни форми. Предстои включването на обработката на такива форми в синтактичен анализатор.

ЛИТЕРАТУРА

[1] Кичович М., "Методи за компютърна морфологическа обработка на лексиката на българския език и за поддържане на автоматични речници".,Автореферат, канд. дисертация, С., 1988.

[2] Крушков Хр. "Моделиране и изграждане на машинни речници и морфологични процесори", Пловдив, Дисертация за присъждане на образователна и научна степен "Доктор", 1997

[3] Кръстев Б. "Морфологията на българския език в 187 типови таблици", София, Наука и изкуство, 1984

[4] Граматика на съвременния български книжовен език, т. II. Морфология, София, БАН, 1983

[5] Krushkov Hr. "Development of Integrated Linguistic Environment" 20-th International Conference ITP'95:Interaction between Inteligent Entities, 16-21 June1995 г., Plovdiv, p.131

[6] Simov K., **Angelova** G., **Paskaleva** E. MORPHO-ASSISTANT: The Proper Treatment of Morphological Knowledge. COLING'90 - 13th Int. Conf. on Computational Linguistics, Helsinki, 1990, vol 3:pp.455-457.

Христо Димитров Крушков e-mail: hdk@pu.acad.bg
Диан Димитров Ганчев e-mail: dian_bg@gotmail.com
Мариана Иванова Крушкова e-mail: mik@pu.acad.bg
ПУ "Паисий Хилендарски", ФМИ, катедра "Компютърна Информатика"
ул. "Цар Асен" 24, 4000 гр. Пловдив
http://www.pu.acad.bg/hdk/hdk1.htm

AUTOMATIC ANALYSIS AND SYNTHESIS OF COMPLEX VERB FORMS

Hristo Dimitrov Krushkov, Dian Dimitrov Ganchev, Mariana Ivanova Krushkova

The Bulgarian verb system is very rich of forms and meanings. The automatic processing of such forms is a very important prerequisite for the syntactic parsing. An algorithm has not been realized yet for such a processing. Only separated lexemes are analyzed. In this paper a formal model of complex verb forms as well as an algorithm for automatic analysis and synthesis of these forms are presented. The model and the algorithm are in the basis of software modules added to the Integrated Linguistic Environment [5]. The modules contains an editor for input and updating the model of complex verb forms as well as functions for analysis and synthesis of such forms.

AMS Subject Classification: 68T50