ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 49, КН. 1, СБ. А, 2011 – ФИЛОЛОГИЯ,
PAISIY HILENDARSKY UNIVERSITY OF PLOVDIV – BULGARIA
RESEARCH PAPERS, VOL. 49, BOOK 1, PART A, 2011 – LANGUAGES AND LITERATURE

СИНТАКСИС И КОМПЮТЪРНА ОБРАБОТКА НА СЛОЖНИ ИЗРЕЧЕНИЯ

Петя Бъркалова
Пловдивски университет „Паисий Хилендарски“
Светла Коева
Институт за български език – БАН

The paper presents an approach for describing the syntactic structure. The ultimate aim is automatic parsing of Bulgarian clauses. The existing classifications of Bulgarian clauses are presented and discussed briefly. The constituent structure defined in some modern Bulgarian Syntax textbooks is also considered. In contrast, a context-dependent approach based on statistical information conning from a large amount of data is suggested.

Key words: Bulgarian syntax, clause structure parsing, context-dependent grammar

1. Въведение

При описанието на строежа на сложното изречение за целите на автоматичния синтактичен анализ се пораждат следните въпроси:

Какви са основните типове сложни изречения в съвременния български език и кои са техните индивидуални характеристики?

Какви са комбинаторните свойства на подчинените изречения по отношение на доминиращите категории NP, AP, VP, ADP и кои са средствата, осигуряващи хипотактичното свързване?

Как да се представи строежът на сложното изречение във вид, подходящ за компютърна обработка?

Какви са закономерностите при паратактично и хипотактично комбиниране на изреченията и как синтактичните факти – структури и закономерности, могат да се представят чрез правила?

Какъв е най-подходящият формат на правилата от гледна точка на компютърната им обработка?

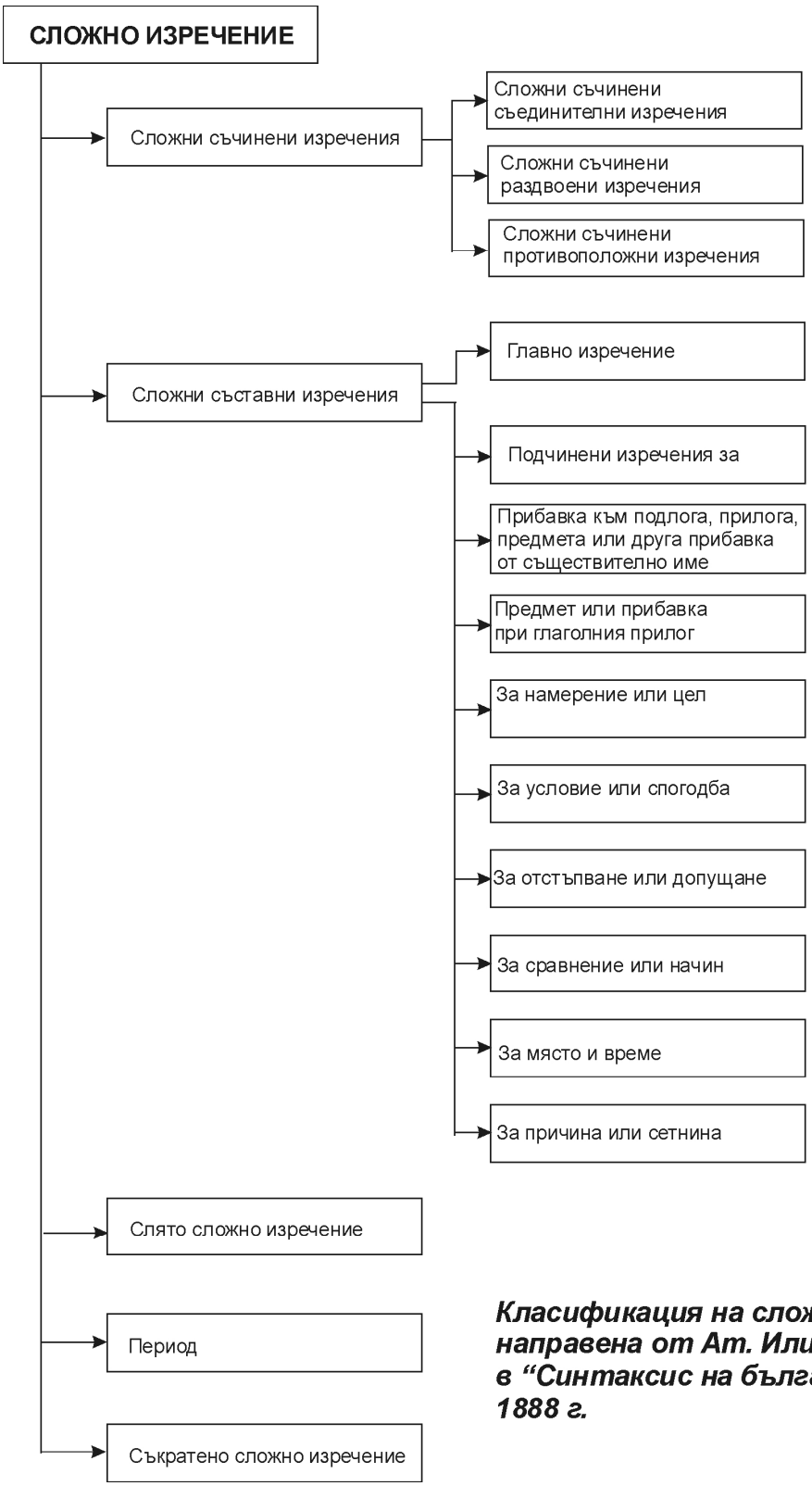
Отговорът на всеки един от тези въпроси изисква сериозни изследвания и експерименти – по тази причина настоящото

изложение се ограничава до: кратко представяне на някои класификации на сложните изречения, известни в българската граматична литература; коментар на правилата за конституентност, предлагани в някои съвременни граматики, и предложение за частичен синтактичен анализ за нуждите на компютърната обработка на езика. Ясно е, че структурата на голяма част от простите изречения, които влизат в състава на сложното, се различава съществено от структурата на самостоятелно употребените прости изречения. Следователно основна предпоставка за коректен синтактичен анализ (макар и частичен) е надеждната информация както за границите на простите изречения в състава на сложното, така и за границите на отделните словосъчетания.

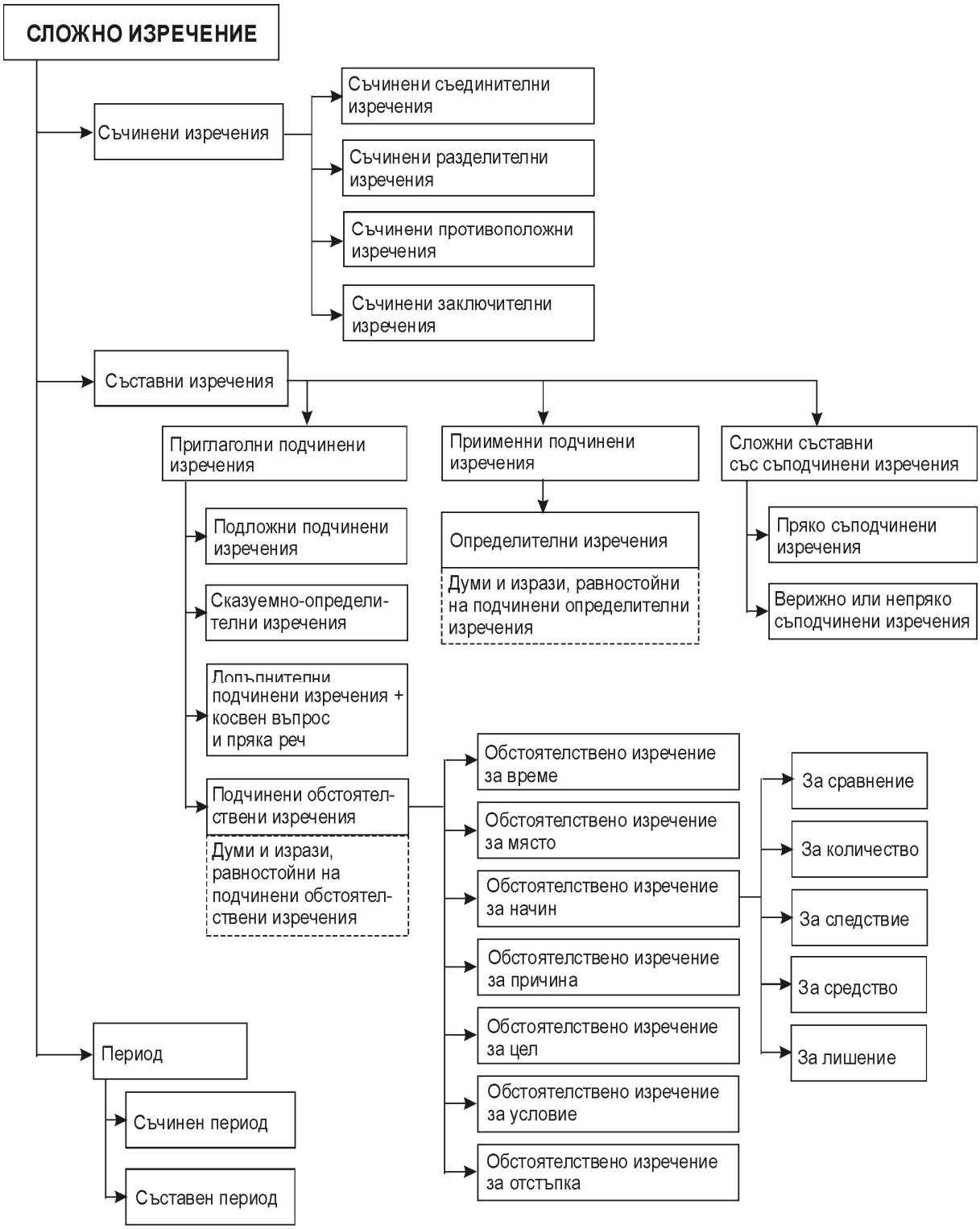
2. Класификация на сложните изречения в българската синтактична традиция в зависимост от състава им

По-долу накратко се илюстрират четири класификации на българското сложно изречение, предложени от Илиев (Илиев, 1888), Калканджиев (Калканджиев, 1936), Пенчев (Пенчев, 1993) и Пашов (Пашов, 1994)[1] в рамките на период, надвишаващ 100 години. Традиционно класификацията включва вида на връзката между простите изречения – съчинителна или подчинителна, и вида на подчинените изречения от гледна точка на синтактичната им функция. По-нови класификации (Пенчев, 1993) обръщат внимание и на начина на свързване на простите изречения – съюзно или безсъюзно (с помощта на относителни и въпросителни местоимения или местоименни наречия или словосъчетания с тях).

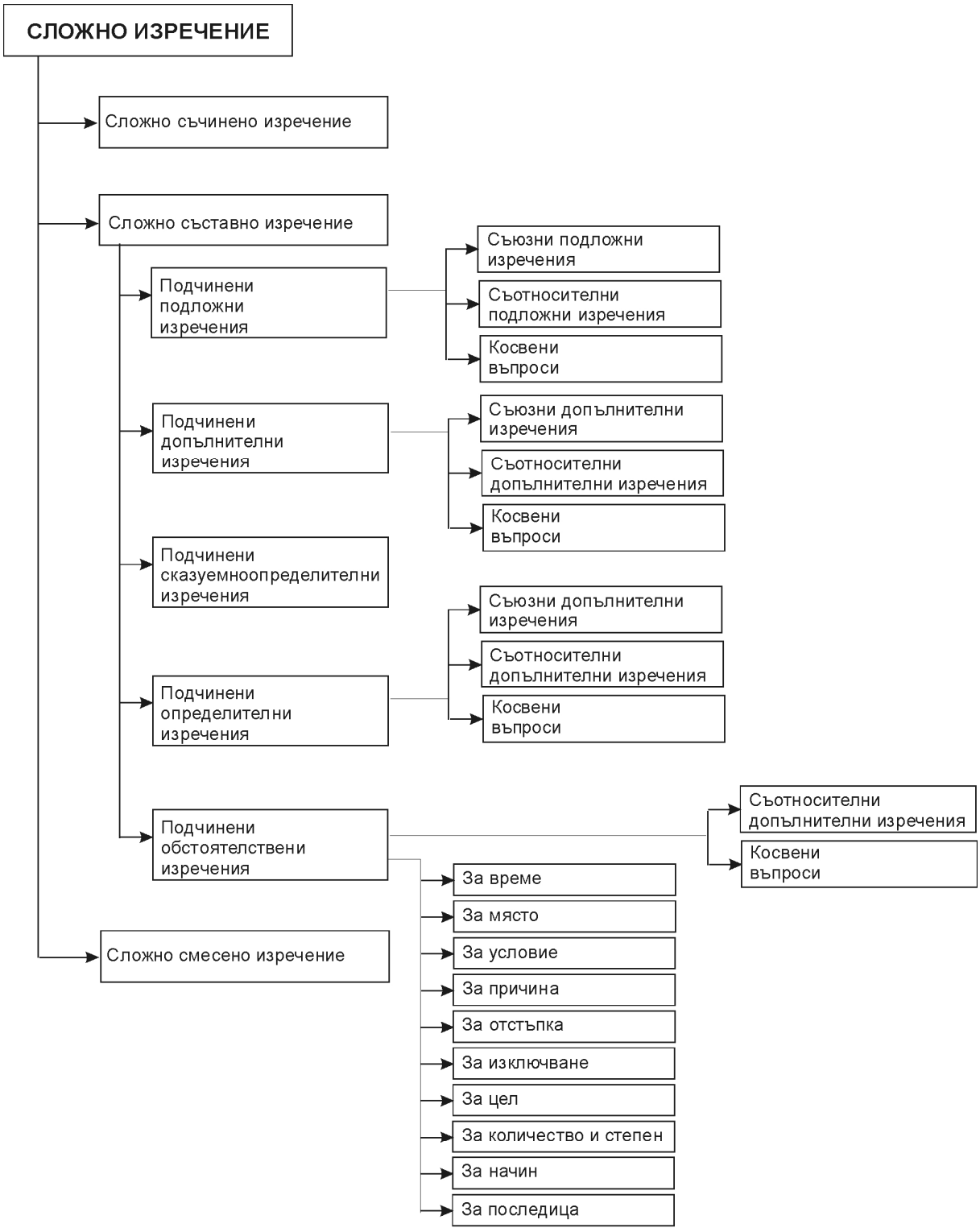
[1] Таблиците са предоставени от П. Бъркалова.



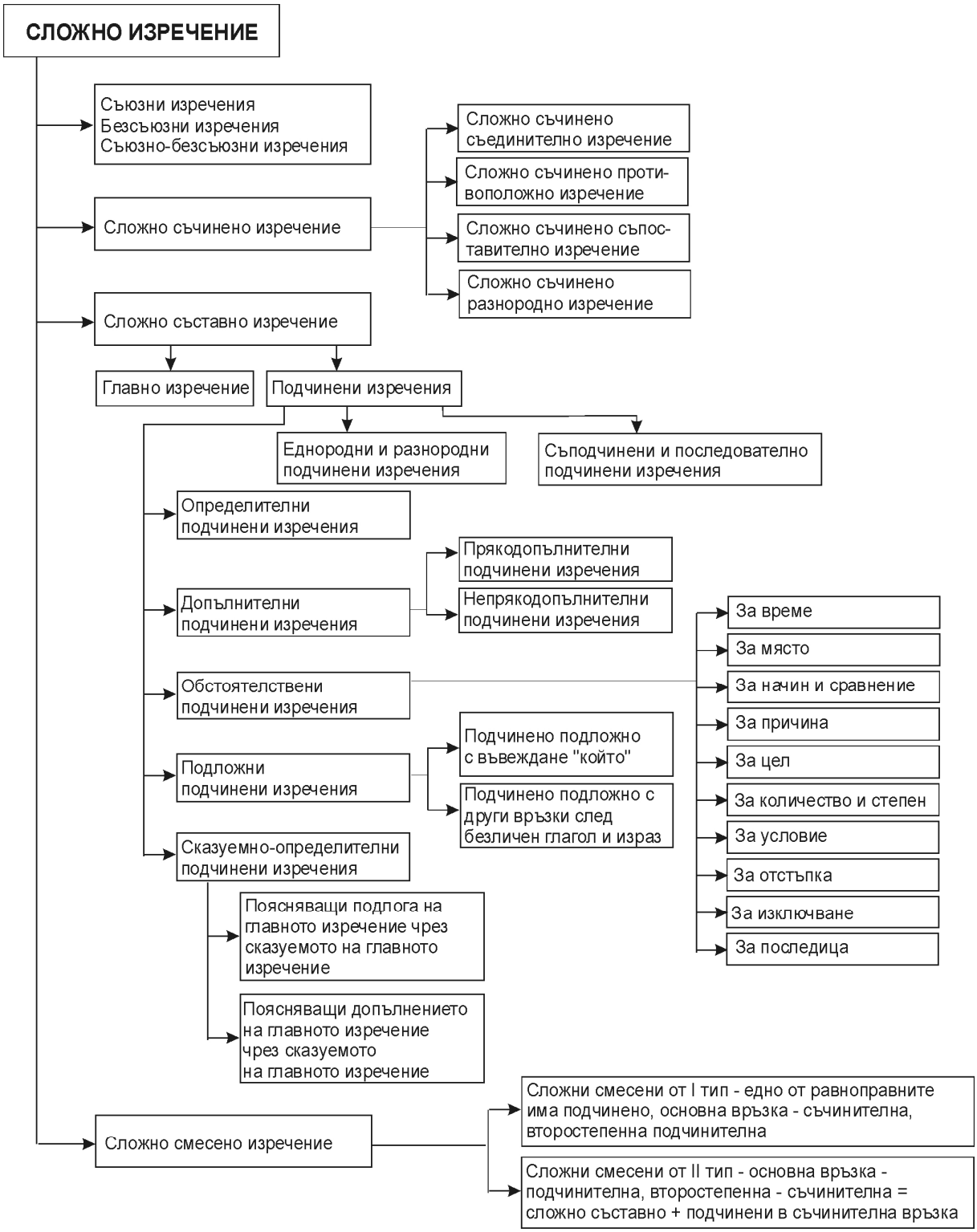
Класификация на сложните изречения, направена от Ат. Илиев в "Синтаксис на българския език" 1888 г.



Класификация на сложните изречения, направена от П. Калканджиев в "Българска граматика" 1936 г.



Класификация на сложните изречения, направена от Йордан Пенчев в "Български синтаксис. Управление и свързване" 1993 г.



Класификация на сложните изречения, дадена от П. Пашов в "Практическа граматика" 1994 г.

До каква степен съществуващите класификации могат да помогнат при решаването на задачи, свързани с автоматичен синтактичен анализ на български сложни изречения? Ако анализът се базира на лингвистични правила (а не е изцяло статистически), за формулирането на правилата е необходимо еднозначно формулиране за границите на фразите, в това число и на простите изречения в състава на сложното; за начините на изразяване на паратактични и хипотактични отношения; за категориите, които са допустими опори; за категориите и подкатегориите (в някои случаи дори групи от лексикални единици), които са допустими предпоставени и задпоставени модификатори на опорите. В този смисъл съществуващите класификации са само необходимата отправна точка, съдържаща минимално лингвистично знание, която трябва да бъде прецизирана и разширена за нуждите на автоматичния компютърен анализ.

3. Конституентни правила

Фразовоструктурните правила, които обикновено се свързват с Трансформационната граматика на Чомски (Чомски, 1957; 1965), по-късно заменени в съвременните лингвистични теории от т.нар. правила за непосредствено доминиране, задават възможните синтактични (конституентни) структури за даден език. Правилата са контекстно-свободни и задават възможните комбинации от лексикални категории (части на речта, допустими в дадена синтактична позиция) и синтактични категории (словосъчетания и изречения, допустими за дадена синтактична позиция). За български основните фразовоструктурни правила са представени от няколко автори Й. Пенчев (Пенчев 1984; 1993), П. Бъркалова (Бъркалова 1997), Св. Коева (Коева 1999) и И. Кръпова (Кръпова 2000).

Например Бъркалова (Бъркалова 1997) описва следните възможности за поява на различни видове подчинени изречения към съответна главна част, разделени в зависимост от начина на свързване и представени в обобщени конституентни правила.

Определителни изречения – в приименна позиция:

- При съществителни: съюзни и безсъюзни (въпросителни и относителни);
- При лични, показателни, неопределителни, отрицателни, обобщителни местоимения: безсъюзни (относителни);
- При прилагателни: съюзни и безсъюзни (относителни);

- При показателни и обобщителни местоимения: безсъюзни (относителни).

Допълнителни изречения – в приглаголна позиция *съюзни* и *безсъюзни* (въпросителни и относителни).

Обстоятелствени изречения:

- При глаголи: съюзни и безсъюзни (относителни);
- При наречия: съюзни и безсъюзни (относителни).

Подложни изречения:

- При пълнозначни глаголи: безсъюзни (относителни);
- При (факултативно и задължително) безлични конструкции: съюзни и безсъюзни (относителни).

Сказуемноопределителни изречения:

- Към подлога, в позицията на предикатива в прикопулна позиция: съюзни и безсъюзни (въпросителни и относителни);
- Към допълнението: съюзни и безсъюзни (относителни).

Ясно е, че съществуват два типа подчинени изречения: изречения, които са във фразовата структура на словосъчетание с главна част съществително, прилагателно, наречие или глагол, и изречения, които заемат позицията на изреченския субект или на изреченския предикат.

Представянето на конституентната структура по подобен начин е коректно, но твърде обобщено за нуждите на автоматичния синтактичен анализ. Съществуват и някои по-детайлни описания на синтактичната структура за български (Петрова, 2009), но те са насочени само към анализа на простото изречения. Макар че, както вече беше отбелязано, структурата на простите изречения в рамките на сложното се различава от структурата на самостоятелно употребените прости изречения – това не засяга отделните фрази и подобни изследвания са полезни, тъй като все още няма пълно и непротиворечиво описание на българския синтаксис.

4. Контекстно-зависими правила

За да се формулират синтактични правила, които коректно да описват фразовата структура за български, е необходимо да се установят еднозначни маркери за границите на фразите, както и всички допустими комбинации от лексикални и синтактични категории, образуващи фрази, които са валидни за български. Това предопределя формата на правилата – бинарни контекстно-зависими правила с десен контекст – граница на фраза. Бинарните фразови

структури от своя страна дефинират зависимости не само между части на речта, но и между специфични граматични характеристики – по такъв начин се предвиждат и възможните ограничения, които се наблюдават в езиковата реализация (Коева, 2010).

Правилата се формулират на базата на статистическа информация за комбинаторните характеристики на българските лексикални единици. По-точно – извлечени са триграми, съдържащи част на речта и граматични характеристики, от големи по обем текстове, представителни за съвременния български език – в случая Българския национален корпус (Коева и Стоянова, 2011). На всяка единица от текста – включително и на пунктуацията – автоматично е приписана съответната граматична информация. Триграмите са унифицирани, като се пази информация за броя на срещанията им и са приложени произволно избрани примери – първо се анализират тези с най-голяма честота на употреба. След унификацията триграмите са сортирани по третия си конституент, а след това, в рамките на получените групи, по втория си конституент. Голяма част от извлечените триграми не илюстрират валидни за българския език фрази. Тези от тях обаче, които съвпадат с действителни фрази, дават надеждна информация за това: кои комбинации от част на речта и граматични характеристики могат да се приемат за еднозначен десен (ляв) край на фраза – десен (ляв) контекст на правилата; и кои комбинации от две части на речта и прилежащите им граматични характеристики могат да се приемат за граматични обобщения на фрази. В някои случаи е уместно да се използват и четириграми, в които се дефинира както десен, така и ляв контекст – маркиращи границите на бинарна фраза. Например в изречението *В една от кабините видя, че управлението отказва.* думите *в* и *от* са разпознати като предлог (P---), думата *една* – като числително бройно в единствено число, женски род, нечленувана форма CQ--:sf0, думата кабините – като съществително нарицателно от женски род в множествено число, нечленувана форма (NCF-:pd), думата *видя* – като личен преходен глагол от свършен вид в трето лице единствено число минало свършено време (VPPT:3sa), препинателните знаци , и . – като пунктуация (PU), думата *че* – като подчинителен съюз (CS--), думата *управлението* – като съществително нарицателно в среден род единствено число, членувана форма (NCN-:sd), и думата *отказва* – като личен преходен глагол от несвършен вид в трето лице единствено число (VPIT:3sp). Извлечените триграми от части на речта

и прилежащите им граматични характеристики са илюстрирани в примера по-долу.

P---	CQ--:sf0	P---
в	една	от
CQ--:sf0	P---	NCF-:3pd
една	от	кабините
P---	NCF-:3pd	VPPT:3sa
от	кабините	видя
NCF-:3pd	VPPT:3sa	PU
кабините	видя	,
VPPT:3sa	PU CS--	
видя	,	че
PU	CS--	NCN-:3sd
,	че	управлението
CS--	NCN-:3sd	VPIT:3sa
че	управлението	отказва
NCN-:3sd	VPIT:3sp	PU
управлението	отказва	.

Могат да бъдат формулирани следните правила:

CS--, CLAUSE, PU --> CS--, NCN-:sd VPIT:3sp, PU (съществително и глагол образуват подчинено изречение, ако са между подчинителен съюз и точка)

PP(NCF-:3pd), V--- --> P--- NCF-:3pd, V--- (предлог и съществително пред глагол образуват предложна фраза)

Правилата могат да се формулират, така че да свързват не само лексикални, но и синтактични категории, получени като резултат от действието на други правила. Например:

NP(NCF-:3s0), V--- --> CQ--:sf0 PP(NCF-:3pd), V--- (числително бройно в женски род единствено число и предложна група със съществително в женски род пред глагол образуват словосъчетание с главна част съществително в женски род единствено число – *една кабина от кабините*)

PP, V--- --> P--- NP(NCF-:3s0), V--- (предлог и именна група пред глагол образуват предложна група)

Правилата, формулирани на базата на статистически данни, извлечени от големи по обем корпуси от текстове, и съдържащи информация за неednозначен маркер за десен или десен и ляв край на фраза, както и достатъчно детайлна информация за част на речта и

прилежащите ѝ граматични характеристики, от една страна – предоставят достоверно описание на синтактичната структура на българския език, от друга страна – предлагат детайлна информация за възможните комбинации и наличните ограничения при съчетаемостта.

5. Заключение

Граматика от контекстно-зависими бинарни правила като описаните по-горе не може да опише еднозначно структурата на всички български изречения, защото ще останат случаи на неоднозначно интерпретиране, в които само човек може да отстрани многозначността. Независимо от това правилата са подходящи за автоматичен анализ, при което ще осигурят голяма точност и относително добро покритие.

Литература

Бъркалова 1997: Бъркалова, П. Българският синтаксис – познат и непознат. Пловдив.

Илиев 1888: Илиев, Ат. Синтаксис на българския език. Пловдив.

Калканджиев 1936: Калканджиев, П. Кратка българска граматика. Пловдив.

Коева 1999: Коева, Св. Синтаксис и пунктуация. Пловдив.

Koeva 2010: Koeva, Sv. Syntactic Annotation in Bulgarian National Corpus. In: Proceedings from the seventh international conference Formal Approaches to South Slavic and Balcan Languages, Dubtovnik, 35–41.

Коева и Стоянова, 2011: Коева, Св. и Ив. Стоянова. Български национален корпус, в: Български език, кн. 3, 137–146.

Кръпова 2000: Кръпова, Ил. Лекции по езикознание. Пловдив.

Пашов 1994: Пашов, П. Практическа българска граматика. София.

Пенчев 1984: Пенчев, Й. Строеж на българското изречение. София.

Пенчев 1993: Пенчев, Й. Български синтаксис. Управление и свързване. Пловдив.

Петрова 2009: Петрова, Ив. Синтактичен анализ на простото съобщително изречения. Дисертация за присъждане на образователната и научна степен „доктор“, София.

Чомски 1957: Chomsky, Noam. Syntactic Structures. The Hague/Paris: Mouton.

Чомски 1965: Chomsky, Noam. Aspects of the theory of syntax. Cambridge, MA: MIT Press.