БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Каменка Атанасова Стайкова

ЛИНГВИСТИЧНИ И СЕМАНТИЧНИ РЕСУРСИ ПРИ КОМПЮТЪРНО ГЕНЕРИРАНЕ И АНОТИРАНЕ НА БЪЛГАРСКИ ТЕКСТОВЕ

АВТОРЕФЕРАТ

за придобиване на образователната и научна степен „доктор“

по специалност 01.01.12 Информатика

професионално направление 4.6 Информатика и компютърни науки

Научен ръководител:
доцент д-р Данаил Дочев

София, 2015 г.

Дисертацията е обсъдена и допусната до защита на разширено заседание на секция „Лингвистично моделиране и обработка на знания” на ИИКТ-БАН, състояло се на 28.04.2015 г.

Дисертацията съдържа 180 стр., от които 51 стр. в приложения, 27 фигури, 12 таблици и 5 стр. литература, включваща 83 заглавия.

Защитата на дисертацията ще се състои на _____20___ г. от _____ ч. в зала _____ на блок _____ на ИИКТ – БАН на открито заседание на научно жури в състав:

1. проф. Радослав Павлов, ИМИ - БАН
2. проф. Мария Нишева, СУ „Кл. Охридски”, ФМИ
3. проф. Иван Койчев, СУ „Кл. Охридски”, ФМИ
4. проф. Галя Ангелова, ИИКТ - БАН
5. доц. Данаил Дочев, ИИКТ - БАН, научен ръководител

Материалите за защитата са на разположение на интересуващите се в стая 215 на ИИКТ – БАН, ул. „Акад. Г. Бончев”, бл. 25А.

Автор: Каменка Атанасова Стайкова

Заглавие: Лингвистични и семантични ресурси при компютърно генериране и анотиране на български текстове

Съдържание

Обща характеристика на дисертационния труд

Компютърното генериране на текст (Natural Language Generation) е научна дисциплина, чиято цел е продуцирането на разбираеми текстове на естествен език. Генерирането и анотирането на естествено-езикови текстове е особено актуално в съвременния свят, когато комуникацията е изключително бърза благодарение на Глобалната мрежа. От една страна съществува необходимост информация във вътрешно компютърно представяне да бъде поднесена в разбираем за хората вид, от друга страна, огромният обем достъпна текстова информация би била по-разбираема и полезна ако е систематизирана и подредена по някакви (семантични) критерии. В първия случай са приложими стратегиите на компютърното генериране на текст, във втория случай са полезни техниките за обработване на естествен език, като една от тях е семантичното анотиране.

В Глобалната мрежа са достъпни текстови представяния на всички съвременни езици и това парави обработването на естествен език научна област от изключителна важност и с необозрим потенциал. Предмет на настоящата дисертация са ресурсите за генериране и анотиране на български текстове. Научната общност занимаваща се с компютърни обработки на български език все още се нуждае от стабилни и пре-използваеми ресурси, отворени за допълване и развитие. В този смисъл лексико-граматическият ресурс за генериране на български текстове разглеждан в настоящата дисертация дълго ще бъде ценен и актуален с предлаганите възможности за по-нататъшно разширяване и развитие. Семантичните ресурси използвани за анотиране на български текстове също представят една от най-актуалните теми в обработването на естествено-езикови текстове, а именно- семантичните технологии. Освен, че са свързани с едни от най-интересните съвременни научни изследвания такива ресурси биха могли да имат и други приложения, тъй като по природа са пре-използваеми.

Обзор на основните резултати в областта

Компютърното генериране на текст се определя като „задача интензивно използваща знания” (a knowledge-intensive problem) тъй като изисква много ресурси и различни видове знания. За компютърно генериране на текст са необходими знания за предметния свят описван в текстовете, знания за естествения език, на който се генерира (лексика, граматика, семантика), стратегически реторически знания (как се постигат определени комуникационни цели, как се построяват различни видове текст, стил на текста) и т.н. От една страна за компютърното генериране на текст (КГТ) са важни солидните теоретични изследвания свързани с лингвистичната същност на изходния продукт- текстът на естествен език. От друга страна, своите аргументи има и едно по-технократско отношение към компютърното генериране на текст, което въвлича по специфичен начин някои от съвременните семантични технологии. Най-многобройни са създадените системи и приложения за компютърно генериране на текст на английски език. Все още продукции на друг естествен език са сериозно изследователско предизвикателство.

Абстрактно описани, задачите на компютърно генериране на текст имат две измерения на лингвистична спецификация. От една страна, информацията в бъдещия текст може да бъде по-конкретна или по-абстрактна (стратификация с нива: регистър/ стил на текста, семантика, лексико-граматика, графология/фонология). От друга страна, информацията може да кореспонсира с различни смислови стойности или мета-функции (пропозиционална, интер-персонална или текстова).

Така се очертава представата за същността на генерирането на текст и се определят присъщите задачи при компютърно генериране:

1) селектиране и интерпретиране на съдържанието,

2) планиране на текста,

3) лексико-граматическа реализация на текста.

Техниките за селектиране на съдържанието се свързват с вземането на решение коя част от представената на входа информация да се включи в подготвяния текст и коя- да се пропусне. Класическият начин за определяне на съдържанието е въведен със схемите на текста [McKeown, 1985] и играе важна роля при структурирането на текстове в много от системите за КГТ. Някои от съвременните приложения за КГТ ползват съвсем различни техники за селектиране на съдържание, когато са свързани с т.н. «отворено планиране на текста». Тези техники зависят пряко от специфичното представяне на знанията на входа, а именно представяне в RDF-графи.

Техниките за интерпретиране на съдържанието варират от създаването на таблици съпоставящи понятията от приложната област и лингвистичните ресурси на генератора, до различни по сложност онтологични модели. Ролята на онтологиите тук е да свързват понятията от предметната област на текста с лингвистичната природа на езиковите единици, чрез които се изразяват тези понятия в текста. В [Bouayad-Agha et.al., 2013] се твърди, че една от най-обещаващите такива онтологии е Обобщеният модел (Generalized Upper Model), чиято еволюция продължава повече от 20 години [Bateman et.al, 1990], [Bateman et.al., 2010].

Класически техники за планиране на текста са шаблонните структури, които могат да бъдат използвани за текстове със стереотипни конструкции. Сравнително гъвкав тип текстови шаблони са предложени от [McKeown, 1985] и представени в термините на мрежи на преходите (transition networks). Този подход е една от най-разпространените техники за организиране на текст въпреки явните си ограничения. Създаване на текстови конструкции с по-голяма гъвкавост е възможно чрез прилагане на Теорията за реторичната структура (TPC) [Mann and Thompson, 1988], която предлага общо описание на релациите съществуващи между текстови сегменти като показва дали тези релации са граматически или лексикално сигнализирани.

По отношение на лексико-граматическата реализация на текста съществуващите техники прилагани в КГТ могат да бъдат класифицирани в скала, като в единия й край са множества от шаблони, а в другия- реализациите с граматики. Структурните шаблони могат да бъдат фиксирани или параметризирани, и се задействат от специфични комуникационни цели или семантични спецификации. Граматиките имат за цел да съпоставят на всеки добре форматиран семантичен вход кореспондиращ му низ или последователност от низове представящи генерирания изходен текст. Подходите за тази задача съдържат огромна част от работата по КГТ. Най- известните формализми и съответстващи им граматически ресурси за КГТ са: Опорна фразова граматика [Pollard and Sag, 1994] и семантичното генериране управлявано от опорния елемент (semantic head-driven generation) [Shieber et.al, 1990], Модел на смисъла на текста разработен от Мелчук и повърхностния генератор RealPro [Lavoie and Rambow, 1997], формализмът за функционално унифициране (Functional Unification Formalism, FUF) [Elhadad, 1990] и граматиката за английски език SURGE (Systemic Unification Realization Grammar for English), генериране направлявано от съобщението (message-directed processing) [McDonald, 1983] и лексико-граматическия ресурс MUMBLE, генериране с контрол направляван от граматиката, предложен за системата Penman [Mann, 1983] и граматиките в Средата за многоезиково генериране KPML [Bateman, 1997].

Парадигмата на Системично-функционалната лингвистика е една от основните най-успешни постановки за компютърно генериране на текст. Системично-функционалната лингвистика предлага изключителни възможности за паралели при системично-функционалните описания на различни естествени езици. Такива паралели позволяват по-бърза и по-ефективно формализиране на лингвистичните знания при създаване на система за компютърно генериране на текст за новоразглеждан език.

Специфичното при системично- функционалния възглед за КГТ е организирането на целия процес на генериране около комуникационните цели, а не около граматическите структури на конкретния естествен език. На това се базира идеята за изграждане пространство на търсене от системични мрежи за генериране, което представлява приложен ресурс на системично-

функционалната граматика за даден език. За реализиране на такава граматиката е необходимо моделиране на (някои) лингвистични феномени на естествения език по отношение на Системично-функционалната лингвистика.

Интересна характеристика на компютърното генериране е фактът, че от изходните текстове не може да се оцени какви технологии са използвани в процеса на генериране - елементарни текстови обработки или дълбочинни методи с богати ресурси. Затова е важен обхватът езикови феномени, с които работи генераторът на естествено-езиков текст. Този обхват дава представа за богатството на възможните вариации на изходните текстове.

Новите тенденции в компютърно генериране на текст са свързани с нарастващото значение на семантичните технологии в компютърните обработки. Естествено за компютърното генериране на текст е използването на онтологии за предметната област на генерирането, а също и обобщаващи онтологични модели от по-висок ред за лингвистичните знания. Следователно, онтологичното инженерство е до голяма степен свързано с компютърното генериране на текст. Семантичното анотиране на текстове на естествен език има приложение в областта на генерирането като практически значима подпомагаща дейност. Семантичното анотиране представлява интерес особено за естествени езици различни от английски поради по-малкия брой изследвания и работещи реализации.

Цели и задачи на дисертацията

Основна цел на настоящата дисертация са изследвания и експерименти за реализация на компютърно генериране на текстове на български език чрез създаване на лингвистични ресурси за основни обекти на българския език и на семантични ресурси за конкретни предметни области.

Изследователските задачи на дисертационния труд са конкретизирани на базата на анализа на съвременните подходи, модели и техники за формиране на лингвистични и семантични ресурси, отчитащи проблемите на съвременното генериране на текстове, отразен в обзорната глава. Те са съобразени и с нуждите на изследователската среда, в която е работил авторът.

Задачите на дисертацията са следните:

1. Създаване на формални описания на основни обекти на българския език в рамките на Системично- функционалната теория на Халидей с цел компютърно генериране на технически текстове.
2. Реализация на създадените формални описания във вид на компютърен ресурс в Средата за многоезиково генериране KPML.
3. Апробация на разработения ресурс за компютърно генериране на кохерентни технически текстове на български език в избрана предметна област.
4. Разработване и реализация на схема за анотиране на специализирани текстове на български език на базата на съвременни семантични технологии.

В дисертацията са използвани съществено научните изследвания извършени с активното участие на автора по проектите AGILE и СИНУС.

AGILE: „Automatic Generation of Instructions in Languages of Eastern Europe" (Автоматично генериране на инструкции на три източно- европейски езика) е изследователски проект финансиран от Европейската комисия и реализиран по програмата *INCO-Copernicus* през 1998-2001г. от партниращи си организации от пет държави: Великобритания, Германия, Чехия, България и Русия.

СИНУС: „Семантични технологии за Интернет-услуги и технологично поддържано обучение" е изследователски проект № Д-002-189 финансиран от Националния фонд „Научни изследвания" през 2009-2012г. Резултатите от проекта са достъпни на адрес: sinus.iinf.bas.bg.

Методология на изследването

Компютърното генериране на текст (Natural Language Generation) е под-област на Обработването на естествен език (Natural Language Processing). Научните изследвания и разработки в компютърното генериране на текст са съсредоточени върху създаването на компютърни системи продуциращи разбираем текст на естествен език. Започвайки обикновено от някакво семантично представяне на информацията на входа, системите генериращи естествен език използват знания за езика и знания за приложната област, за да създадат и оформят документи, отчети, рапорти, обяснения, помощни съобщения или други видове естествено- езикови текстове.

В настоящата дисертация са представени научно-приложни изследвания свързани с процеса на компютърно генериране на български текстове. От гледна точка на информатиката основни за компютърното генериране на текст са формализацията и обработката на лингвистични знания (лексико-граматика). В настоящата дисертация някои лингвистични феномени на българския език са обект на анализ и моделиране в парадигмата на Системично – функционалната граматика на [Halliday, 1994]. Показано е изграждането на приложен лексико-граматически ресурс за генериране на български език. Достъпът до Средата за многоезиково генериране KPML подпомага съществено работата по създаване, тестване и настройка на приложна системично- функционална граматика за генериране на български език.

Системично- функционалните граматики създадени на базата на теорията на [Halliday, 1994] кодират семантичните връзки в текста във функционална форма и са насочени към директни съответствия между по-високите нива на организация на текста и граматическия компонент. Това е предпоставка за научно-приложна изследователска работа по генериране на различни типове/ стилове на изходни текстове от едно и също входно представяне.

С развитието на семантичните технологии в последите години се наблюдават интензивни научно-приложни изследвания свързани с подпомагане компютърното генериране на текст, например за по-удобно представяне на знания в онтологични конструкции или по-ефективни семантични обработки на естествено-езикови текстове. Изследователско направление в настоящата дисертация е семантичното анотиране на специализирани текстове на български език чрез прилагане на подхода „Релация: От онтология към текст“. Основна идея на метода е специално разработени ресурси (терминологичен лексикон, анотационна граматика) да се използват в комбинация при разпознаването на срещания на онтологичните понятия от дадена онтология в текст. Изследването е реализирано с частични граматики на базата на регулярни изрази разработени в системата CLaRK.

Структура на съдържанието

Първата глава на дисертацията представлява обзор, който представя по същество компютърното генериране на текст като научно-приложна област и описва лингвистичните и семантичните ресурси, които се използват в процеса на генериране. Първите два под-раздела 1.1 и 1.2 се занимават с постановката на задачата за КГТ и със съвременните техники и методологии използвани в процеса на КГТ. В раздел 1.3 са разгледани накратко семантичните технологии характерни за областта Обработване на естествен език, тъй като КГТ е нейна под-област. Коментирани са също съвременните идеи и перспективи на компютърното генериране на текст от данни на Семантичната мрежа.

Втора глава представя работата по създаването на ресурс за генериране на български изречения. В раздел 2.1 са показани теоретични модели за някои граматически явления в българския език на базата на Системично-функционалната граматика на Халидей. В раздел 2.2 са описани принципите за създаване на приложна българска системично-функционална граматика. Показани са резултати от автоматичното генериране на български изречения със създадената приложна българска системично-функционална граматика.

Трета глава демонстрира използването на приложната граматика като лексико-граматически ресурс за генериране на български език в конкретна приложна област. Описан е процесът на планиране на текст. Показани са резултати от автоматично генериране на кохерентни текстове на български език в различни стилове.

В четвърта глава се обръща внимание на семантичните технологии като представяне на знания и семантично анотиране, които биха подпомогнали подготвителните работи при компютърно генериране на текст. Използвана е конкретна приложна област за изследване на семантично анотиране с частични граматики на базата на регулярни изрази.

1. Компютърно генериране на текст, лингвистични и семантични ресурси

Компютърното генериране на текст (Natural Language Generation) е под-област на компютърната обработка на естествен език (Natural Language Processing). Това е дисциплина, в която научните изследвания са съсредоточени върху създаването на компютърни системи продуциращи разбираем текст на естествен език. За да разграничим Генериране на естествен език от научната област Синтезиране на говор, в тази дисертация ще използваме превода Компютърно генериране на текст.

Дефиниране на задачата

Най-общо основната задача на компютърното генериране на текст може да се формулира като превръщане на някакъв вид не-лингвистична информация чрез работата на компютърна система в писмен текст на естествен език, подходящ за възприемане от човека. Робърт Дейл[1] дава следната дефиниция[2]:

"Компютърното генериране на текст е процес на целенасочено построяване на текст на естествен език, за да бъдат постигнати определени комуникационни цели."

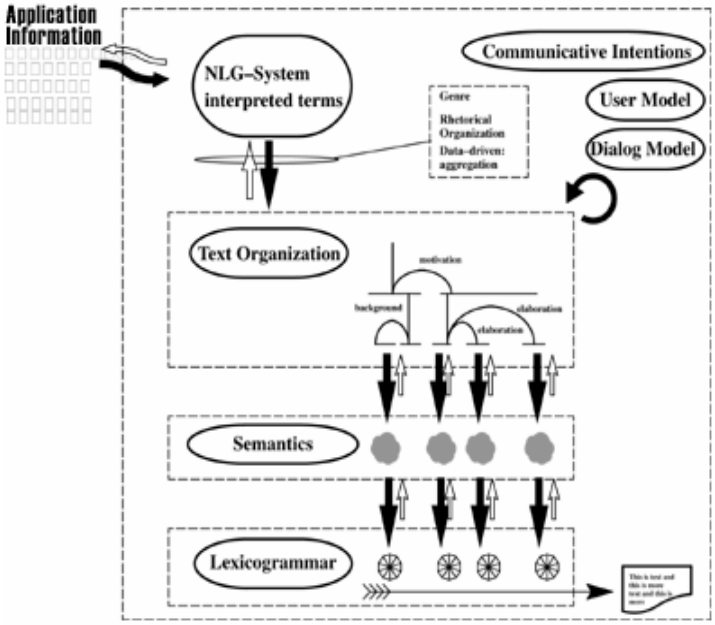
КГТ има приложен аспект с определена практическа стойност. Тъй като компютърните системи използват вътрешни представяния на информацията (бази-данни, счетоводна информация, бази-знания, данни от тестове за работата на машини и съоръжения и т.н.), то съществува необходимост от компютърни програми, които да представят такава информация или обобщения на най-важните й характеристики в разбираем вид за не- специалисти в конкретната област. Технологията за КГТ се използва за представяне на данните в текстов вид в подходящ и удобен за човека формат. Дефинирането на конкретната задача насочва прилагането на технологията за КГТ или към автоматично генериране без участието на човека или към полуавтоматично генериране с участие на потребителя.

От научно-изследователска гледна точка КГТ се определя като „задача интензивно използваща знания" (a knowledge-intensive problem) тъй като изисква много ресурси и различни видове знания За КГТ са необходими знания за предметния свят описван в текстовете, знания за естествения език (лексика, граматика, семантика на съответния естествен език), стратегически реторически знания (как се постигат определени комуникационни цели, как се построяват различни видове текст, стил на текста) и т.н.

Модел на процеса за КГТ

Предложената от Джон Бейтман Абстрактна система за КГТ[3] представена на Фигура 1, е резултат от анализ на голям брой приложни системи за КГТ от перспективата на обхвата езикови феномени, с които системите се занимават.

[1] Робърт Дейл е професор в Macquarie University, Австралия: http://web.science.mq.edu.au/~rdale/
[2] http://web.science.mq.edu.au/~rdale/teaching/esslli/part01.pdf
[3] http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/nlg/ATG01/ATG01.html



Фигура 1: Абстрактна система за КГТ

При КГТ не е установен общ, външно мотивиран източник на генерирането. Съществува консенсус, че „входът" трябва да бъде някакво семантично представяне. За различните системи решенията за „вход" са съвсем различни. Входните данни могат да бъдат, например, минимално структурирани числови данни, статистически рапорти, съдържанието на бази-данни. Разнообразни семантични представяния са били изпробвани през годините – логики от първи ред и свързаните с тях формализми [Basile and Bos, 2011], сценарии на Шанк (Schank's scripts) [Hovy, 1988], концептуални графи на Сова [Nogier and Zock, 1992], различни фреймови представяния, а напоследък и RDF-представянията типични за Семантичната мрежа.

С изхода на процеса на генериране ситуацията е сходна. Генерираните текстове са разнообразни: с голяма дължина или кратки, самостоятелни параграфи или единични изречения (напр. отговорите на запитвания към бази-данни), могат да са низове или да е приложено по-сложно текст-форматиране (пунктуация, оформяне на страници), могат да имат линейна структура, да бъдат част от диалог или да бъдат организирани като хипертекст. Текстовете могат да бъдат насочени към аудитории различаващи се по опит, знания, интереси или по когнитивно натоварване; може да се изисква изход на различни естествени езици и т.н. Систематично представяне на цялостната картина за входа и изхода на системите за генериране на текст все още не е предложено.

Техниките, които се използват за генериране на текст могат да бъдат съпоставени и оценени спрямо изискваното езиково многообразие в изходните текстове. Колкото повече гъвкавост на текстовете е търсена, толкова по-общи лингвистични знания са необходими и е по-голяма мотивацията за обръщане към характерните техники на КГТ. Разглеждането на приложните системи от перспективата на обхвата обработвани езикови феномени позволява вариациите на изходните текстове да бъдат поставени в общ план, като се абстрахираме от реализационното ниво на системите.

Абстрактно описани задачите на КГТ имат две измерения на лингвистична спецификация: разслояване/ стратификация (stratification) и мета-функция (metafunction). Тези две измерения са извлечени от функционалния семиотичен подход на Системично- функционалната лингвистика: информацията може да бъде по-малко абстрактна или много абстрактна (стратификация с нива: регистър, семантика, лексико-граматика, графология/фонология), а също така информацията може да се отнася за различни смислови видове (мета-функции: пропозиционална, интер-персонална или текстова). Така се очертава представата за същността на генерирането на текст и се определят присъщите му задачи:

1) селектиране и интерпретиране на съдържанието,

2) планиране на текста,

3) лексико-граматическа реализация на текста.

За реализиране на КГТ трябва да бъдат направени съвместими избори пропозиционално, интерперсонално и текстово, за да се конструират последователности от семантични спецификации, които поддържат текстовата семантика и я изразяват в граматически конструкти. Без значение как е реализирана една система за КГТ, тези абстрактни задачи са същностни за генерирането. Ако някаква част от поставените от задачите проблеми не са решени, то гъвкавостта на системата за КГТ е ограничена по отношение на текстовото разнообразие.

На Абстрактната система за КГТ (Фигура 1) можем да погледнем и като на приложен модел на процеса на КГТ. Входът за процеса на КГТ е наречен *информация за приложението* и е отбелязан в горния ляв ъгъл на изображението на Абстрактната генерираща система. Можем да приемем текст-планировчика и лексико-граматиката като отделни модули. След селектиране на съдържанието текст-планировчикът разполага със целево ориентирани техники конструиращи последователности от не-строги семантични спецификации с размерите на изречение. Лексико-граматиката преработва такива сравнително абстрактни спецификации в не толкова абстрактни граматически конфигурации и свързан с тях лексически материал. На изхода имаме последователност от граматически спецификации превърната в текст съставен от символите на дадения естествен език.

Взаимодействието между компонентите е показано със стрелки в двете посоки, въпреки че в много приложни архитектури не е поддържана такава двупосочност. Много разлики между подходите за КГТ се дължат на гъвкавостта, която те придават на съпоставянето на представянията във вътрешните нива. Отделни листа на текстовата структура в някои системи се съпоставят на отделна семантична спецификация, а другаде- на последователност от семантични спецификации. По същия начин, отделни семантични спецификации могат да съответстват на последователности от лексико-граматически спецификации, но на практика най-често са реализирани прости връзки и съответствия.

За целите на настоящата дисертация моделът на Абстрактната генерираща система показан на Фигура 1 е използван като основа за изследване на различните нива на лингвистична абстракция на текстове- инструкции на български език. Основната мотивация на изследването е на практика да бъде реализирано генериране на текстове на български език, като се създадат нужните за това ресурси. Настоящият обзор е насочен към преглед на прилаганите техники за решаване на задачите на КГТ поставени от лингвистичната природа на изходния продукт – текстът като такъв.

Техники за селектиране и интерпретиране на съдържание

Селектиране на съдържанието е вземането на решение коя част от информацията да се включи в подготвяния текст и коя- да се пропусне. Това е сложна задача зависеща от много фактори като представата какво желае да знае целевата аудитория или степента на детайлизиране на текста изисквано от аудиторията. Тези фактори могат да бъдат групирани в три области:

1. Общи свойства на предвижданото взаимодействие потребител - система за КГТ: какви потребители се имат предвид, каква е темата на интеракцията по същество и т.н.
2. Специфични за интеракцията с потребителя ситуации, които възникват по време на генерирането, например, някоя конкретна част от текста трябва да бъде обяснена или разширен. Ако потребителят е поставен в конкретна експертна аудиторна група, той ще има нужда от по-малко обяснения за термините в текста.
3. Конкретните текстове се приспособяват към естествените очаквания на потребителя, за да бъдат разпознати. В лингвистиката това се нарича жанр или тип на текста.

Например, текст- прогноза за времето има разпознаваеми лингвистични характеристики, които го отличават от текст- техническо ръководство или текст от речник.

Представянето и използването на знания за типа на текста обикновено се прави чрез явно задаване на макроструктура на текста. Това се дължи на факта, че конкретни типове текст имат често срещани конкретни структури, към които трябва да се придържа и системата за генериране. Това създава полезна организационна схема за процеса на КГТ.

Изключително важно за селектиране на съдържанието е това, че конкретни части от макро-структурата на текста изискват изразяването на определен тип информация. Характерни структури от този вид са въведени в КГТ като схеми на текста [McKeown, 1985] и играят важна роля при структурирането на текстове в много от системите за КГТ. Този традиционен начин за определяне на съдържанието със шаблони на базата на правила прилага парадигмата на т.н. „затворено планиране", както например в [Duboue and McKeown, 2003]. Парадигмата на „отворено планиране" е приложена, както например в [Dai et.al., 2010], когато входните данни имат мрежово представяне в граф и съществено се използва топологията на графа за "информирано търсене" на релевантни възли.

Интерпретиране на съдържанието. Най-простият подход към интерпретиране на съдържанието е да се създадат таблици съпоставящи понятията от приложната област и лингвистичните ресурси на генератора

Ако адаптираното лингвистично представяне е достатъчно абстрактно, то е осигурена и възможност за гъвкава реализация. Проблематично е специфичното за приложната област кодиране на информацията. Една техника намерила широко приложение е ограничаването на възможните съпоставяния между термините от приложната област и тези от КГТ само до релацията наречена логическо включване (logical subsumption). Това се осигурява, когато всеки конкретен екземпляр на факт, състояние, ситуация и т.н., които се срещат в приложението могат да бъдат класифицирани в термините на йерархия от общи понятия и релации, които имат систематично поведение по отношение на възможната за тях лингвистична реализация.

Пример за цялостна висша йерархия с търсените качества представлява онтологията описана в [Bateman et.al, 1990] и наречена Висш модел (The Upper Model), която е извлечена от системично-функционалната парадигма за естествения език [Halliday, 1994]. Висшият модел е всъщност изчислителен ресурс за представяне на знания специално разработен за компютърно генериране на текст. Абстрактната организация на знанията е лингвистично мотивирана, за да бъде ограничена лингвистичната реализация при генерирането [Bateman, 1990]. Висшият модел е проектиран така, че да бъде преносим, пре-използваем и външен за лексико-граматиката. Той може да бъде считан за вътрешна връзка между специфична за приложната област информация и лингвистичното граматическо ядро на една система за КГТ. Еволюцията на Обобщения модел (Generalized Upper Model) продължава в посока детайлизиране на някои от под-йерархиите [Bateman et.al., 2010] или превеждане на онтологията на онтологичния език на Семантичната мрежа OWL. Обобщеният модел тепърва привлича интереса на изследователите занимаващи се с генериране от данни на Семантичната мрежа, като в [Bouayad-Agha et.al., 2013] се твърди, че той е най-обещаващата многоезикова лингвистично ориентирана онтология поддържаща процеса на текст-генериране.

Обобщеният модел е основната онтология от високо ниво на средата за многоезиково генериране KPML [Bateman, 1997]. Направени са анализи за адекватността на онтологията и към други езици като напр. арабски [Al-Muhtaseb and Mellish, 1997], както и частични анализи от прагматични съображения при генериране на текст в средата KPML. В [Kruijff et.al, 2000] се докладва за реализираното по проекта AGILE многоезиково генериране на текстове- инструкции на руски, български и чешки в средата KPML извършено със

съществено прилагане на Обобщения модел за тези езици.

Техники за планиране на текста

Задачата за планиране на текста представлява организиране на избраното съдържание в подходящи текстови структури, които да бъдат трансформирани в кохерентен текст. Прилаганите подходи към тази задача се различават много по гъвкавост и изчислителна сложност.

Най-простият подход към организацията на текста като цяло са фиксираните структури или шаблоните. Почти всяка текстова структура може да бъде замразена. При генериране с шаблони се жертва огромна част от гъвкавостта на текста, но не всички приложения за КГТ имат нужда от голяма гъвкавост. Замразяването на различни аспекти от цялостния процес на КГТ съответства точно на основните свойства на самия език, тъй като при съставянето на текстове хората понякога използват наготово определени набори от направени лингвистични избори, вместо винаги да правят тези избори „на живо". Това може да варира от изречения-формули и цели текстове, през различни степени на идиоматичност (на синтаксиса, на начина за структуриране на аргументи, на използваните семантични конфигурации и т.н.). Затова структуриращите шаблони могат да бъдат разглеждани като частично замръзнали резултати от текст-планирането.

Разработки, които могат да ограничат своите основни описания и да компилират полезни шаблони са обещаващи за спестяване на време и труд. Шаблонната структура може да бъде използвана за текстове, които демонстрират стереотипни конструкции. Сравнително гъвкав тип текстови шаблони са предложени от [McKeown, 1985] и представени в термините на мрежи на преходите (transition networks). Този подход се е превърнал в една от най-разпространените техники за организиране на текст въпреки явните си ограничения.

Създаване на текстови конструкции с по-голяма гъвкавост е възможно чрез прилагане на Теорията за реторичната структура (TPC) [Mann and Thompson, 1988], която предлага общо описание на релациите съществуващи между текстови сегменти като показва дали тези релации са граматически или лексикално сигнализирани. Текстовете анализирани чрез TPC са йерархично декомпозирани на вложени множества от свързани текстови елементи (spans). Теорията за реторичната структура дефинира около 25 релации, които важат между текстовите елементи. Дефинициите на Теорията на реторичната структура поставят изискването да се поддържа връзка със смисъла, който трябва да носи даден елемент на кохерентния текст. Има също ограничения върху комуникационния ефект постигнат чрез комбинираното множество от текстови елементи. Конструирането на дискурсна структура на базата на TPC има доказан ефект при поддържане на селекции от свързващи форми и текстови съюзи.

TPC има своето компютърно представяне [Moore and Paris, 1988] и е използвана в много системи за КГТ. Обикновено TPC се прилага за генерирането на текст чрез представяне на реторичните релации като тип комуникационни цели и прилагане на стандартни за изкуствения интелект планиращи стратегии, за да се продуцират текстови структури.

Техники за лексико-граматически реализации

Шаблони

Най-простият метод за конструиране на изречения представляват параметризираните шаблони. Заради простотата си шаблонното генериране е прилагано в някои практически ориентирани системи за КГТ. Вече усложнена тази техника смесва текстови фрагменти – резултат от пълно генериране с предварително заложени шаблони, като по този начин и двата метода имат място в една и съща система за КГТ.

Реализации с граматики

Гръбнака на граматиките са приложните граматически конструкции. С увеличаването им в пространството на търсене се увеличават и възможните граматически реализации. Решаваща роля имат решения, които предлагат подходяща навигация за това пространство на търсене. Различните видове граматически описания могат да доведат до различни възможности за претърсване на пространството.

Структурната граматика е по същество организирана около описания на фразовата структура. Тя обикновено е претърсвана за приложими правила ограничени от семантиката, която трябва да бъде реализира. Най-солидната стратегия тук е алгоритъмът за семантично генериране управлявано от опорния елемент (semantic head-driven generation) [Shieber et.al, 1990]. Този алгоритъм генерира стрингове от логически форми за сравнително широк клас граматически формализми.

Техниката работи по същество следвайки последователности от граматически правила свързани чрез техните синтактични главни или опорни елементи (heads), които споделят обща семантика, за да достигнат до приложими лексикални елементи. Ако не са намерени такива правила или последователността стигне до края, се избира някое правило, което декомпозира семантиката недетерминистично. В случай, че се стигне до лексикални елементи, алгоритъмът работи обратно „нагоре" по дървовидната структура като налага ограниченията намерени в лексикона.

Приема се, че елементите на лексиката предлагат най-богатия източник на ограничения за синтактичната структура и затова те са търсени първи, за да се избегне построяването на неприложими структури. Въпреки елегантността и формалната спецификация на алгоритъма той не е използван извън формалното теоретично генериране на изречения. Остават много отворени въпроси по отношение на работата му с големи лексико-граматики със съществени не-пропозиционални семантични изисквания. Недетерминизмът на алгоритъма също е критикуван в средите на КГТ като неподходящо свойство при реално генериране.

Алтернативата наречена *обработка направлявана от съобщението* (message-directed processing) [McDonald, 1983] е предпочетена за построяване на лексико-граматическия ресурс MUMBLE за английски език. Тук детерминистичната и инкрементална конструкция на фразата се контролира директно от входните спецификации. Тези входни спецификации изискват конкретни синтактични фрагменти на дървовидната структура изразени чрез Tree Adjoining Grammars: [Joshi, 1987]. Такъв вход експлицитно идентифицира конкретните граматически конструкции, които трябва да бъдат селектирани за резултатното изречение.

В подобен стил са входните спецификации за повърхностния генератор RealPro [Lavoie and Rambow, 1997]. Входът му представлява представяне, което е синтактична структура на зависимостите (syntactic dependency structure), генераторът я попълва, за да я превърне в напълно определено изречение. Тази техника е повлияна от многослойния Модел на смисъла на текста разработен от Мелчук и колеги. Генераторът RealPro не наследява по-дълбоките и по-абстрактни лингвистични нива предложени от Мелчук, но като следствие е много бърз.

Контрастираща алтернатива, наречена контрол направляван от граматиката, се предлага от Penman [Mann, 1983] и неговия наследник KPML [Bateman, 1997], които са генератори за системично-функционални граматики. Системичните граматики организират своето пространство на търсене около възможни комуникационни цели, а не около граматически структури. Фрагменти на структурите са локализирани в това пространство на характеристики и сами по себе си имат много ограничен статут. Това е изключително ефективно за нуждите на КГТ.

КГТ изисква описания на причините защо да се използват дадени структури (синтактични, текстови и т.н.), а не само формални описания на използваните структури. Това е естествена територия за функционалната лингвистика, която оказва далеч по-голямо влияние върху широкоспектърните системи за КГТ, отколкото върху анализа на естествен език, където са норма структурните подходи към синтаксиса. Системично- функционалните граматики, с

теоретична основа изложена в [Halliday, 1994], се фокусират точно върху прекодиране на връзките във функционална форма и, следователно, са насочени към директен интерфейс между по-високите нива на организация на текста (планиращите процеси) и граматическия компонент. Традиционно в тези граматики се обръща повече внимание на не-пропозиционалните (текстовите и интерперсонални) аспекти на смисъла [Matthiessen and Bateman, 1991].

Практическа възможност за разработване на приложни системично-функционални граматики дава Средата за многоезиково генериране KPML [Bateman, 1997]. Тя предлага стабилна платформа за работа с широкообхватни граматики и е специално ориентирана към многоезиковото генериране на текст. Основна идея за създаването на Средата KPML е да се предложат ресурси за реалистично, но същевременно широкообхватно генериране, при което се търси както гъвкавост на изходните текстове, така и бързина при генерирането.

Алгоритъмът за генериране в Средата KPML се състои от преходи през пространство на характеристики на генерирания елемент, като преходите са последователни и с нарастваща специфичност. Всеки такъв преход създава множество ограничения определящи един структурен фрагмент. Този фрагмент може да включва граматически конституенти, които да изискват по-нататъшни преходи, за да се специфицират. Макар много прост, алгоритъмът има предимството, че е доста бърз дори за големи граматики като NIGEL за английски език, граматиката KOMET за немски език [Teich, 1999] или AGILE-граматиките за български, руски и чешки [Bateman et.al., 2000]. В алгоритъма няма връщане назад (backtracking).

Формализмът за функционално унифициране (Functional Unification Formalism, FUF) [Elhadad, 1990] предлага по-мощно трасиране на пространството от системични характеристики чрез използване на не-детерминистична експанзия с унификация. Не-детерминизмът е направен по-ефективен чрез няколко допълнителни механизма за направляване процеса на унификация [Elhadad and Robin, 1992]. Граматиката за английски език SURGE (Systemic Unification Realization Grammar for English) има много голямо покритие и е създадена за прилагане на FUF.

Процесът на генериране с FUF се състои от унифициране на такъв вход с подобни на него дефиниции от граматиката. Входните данни направляват процеса на унификация, за да се намерят тези части от граматиката, с които те са съпоставими и да се специфицира нататък структурата в зависимост от ограниченията поставени от граматиката. Подходът с унификация неутрализира до известна степен разделението между контрол направляван от граматиката и контрол направляван от съобщението, доколкото избираният път при унификацията е чувствителен и към двата източника на контрол.

В заключение можем да отбележим, че експерименти за КГТ на най- много различни естествени езици са правени в Средата за многоезиково генериране KPML. Това се дължи на характерния системично- функционален стил при анализа на естествения език улесняващ и насърчаващ аналогиите и паралелните изводи като се спазва посоката от най-общите семантични дефиниции към спецификацията на по-фини нюанси изразявани в конкретния естествен език. Този стил на теоретично и особено на реализационно ниво подпомага пре-използването на базови семантични конструкции и улеснява построяването на лексико-граматики за различните естествени езици. В средата за многоезиково генериране KPML са построени лексико-граматически ресурси за английски, немски, холандски, китайски, испански, руски и други от естествените езици. На базата на този изследователски опит е изграден и ресурсът за генериране на български език Приложна българска системично-функционална граматика.

Компютърно генериране на текст, семантични технологии и ресурси

Семантичните технологии атакуват проблема формулиран като “липса на семантика при изпълняване на компютърните програми от машините”. Компютрите „не разбират”, „нямат понятие” за смисъла на елементите, с които оперират. Хипотезата, върху която се градят

семантичните технологии в компютърната наука е предположението, че компютрите ще демонстрират „по-интелигентно поведение” ако са снабдени с явни, формални описания, въвеждащи семантично ниво на оперативните единици.

Според популярната дефиниция в [Polikoff and Allemang, 2003] семантичните технологии са такива софтуерни технологии, които позволяват смисълът на информацията и асоциациите между информационни единици да бъдат достъпни за обработка по време на изпълнение на програмата. Като под-област на Обработването на естествен език, компютърното генериране на текст е повлияно от семантичните технологии в областта, а те са: извличане на информация от текст, представяне на знания и онтологични конструкции, извличане на знания.

Важна за настоящата дисертация е технологията на семантично анотиране. Според дефиницията в [Erdmann et.al, 2000], при семантичното анотиране на текстове на естествен език към текстовете се прикачат метаданни, които трябва да направят семантиката на термините в текста „разбираема” за машините. При този процес, който е по принцип полу-автоматичен, се извличат знания, в смисъл, че между лексически термини от текста и, например, онтологични понятия се установява връзка. Така се придобиват знания, чиито смисъл е бил търсен в обработвания контекст.

При подхода за семантично анотиране на текст предложен в [Simov and Osenova, 2007] и доразработен в [Simov and Osenova, 2008] към семантичното анотиране се подхожда в перспектива от онтологията към текста, оттам и наименованието му „Релация: От онтология към текст”. При него се използват лексикони, основани на онтология (Ontology-Based Lexicons).

Основната идея на метода е специално разработени ресурси, терминологичен лексикон и анотационна граматика, да се използват в съчетание при разпознаването на срещания на онтологичните понятия в текст. Дефинираната задача никак не е тривиална, тъй като (1) не всички онтологични понятия задължително имат лексикализация, (2) онтологичните понятия не винаги се срещат в текстовете във вида, в който са лексикализирани в онтологията от експерти или специалисти в предметната област и (3) едни и същи онтологични понятия могат да бъдат изразени в естествено-езиков текст по множество различни начини и представени по смисъл със свободни фрази.

Терминологичният лексикон е множество от лексикални еквиваленти на понятията от дадена онтология. Той изпълнява двустранна роля. Първо, лексиконът свързва понятията на онтологията с лексическото знание, използвано от граматиката за разпознаване на ролята на понятието като езиков елемент на текста. Второ, лексиконът представлява основа за създаване на удобен интерфейс между потребителя и онтологията, който позволява онтологията да бъде представена по естествен за потребителя начин.

Анотационната граматика е средство за разпознаване на онтологични термини в целевите текстове. В идеалния случай тя представлява разширение на една обща дълбочинна граматика за даден език, адаптирана към конкретната анотационна задача. Като минимум анотационната граматика представлява частична граматика за анотиране на понятията с добавени правила за разрешаване на многозначност. Частичната граматика съдържа за всеки термин от лексикона най-малко едно граматическо правило за разпознаване на понятието. За да работи анотационната граматика е необходима предварителна обработка на текста, т.е. анотирането му с граматически характеристики и лематизация. Настройването на граматиката може да се счита за втори етап на обучение при прилагане на подхода „Релация: От онтология към текст”.

В сравнение с класическия метод за семантично анотиране при следване на релацията „От онтология към текст“ се използват съществено онтологии като ресурс за съставяне и обогатяване на терминологичния лексикон, което е много удачно при работа с текстове класифицирани тематично в определена област на знанието. КГТ изисква представяне на знанията в тематичната област на бъдещите текстове, затова семантично анотиране с метода „От онтология към текст“ е подходяща подготвителна дейност за генерирането. Ползата от

пре-използване на семантично анотирани текстове би била по-голяма, както за настройване на процеса на КГТ, така и за оценяването му.

Можем да направим извода, че семантичните технологии като цяло реализират алгоритми и решения, които дават семантична структура на информацията, за да се ползва тя и от хората и от компютрите. Това е интересна перспектива по отношение на Компютърното генериране на текст, като се има предвид важността на семантичните представяния за процеса на КГТ. Семантичното анотиране на текст е полезно при предварителните обработки на големи обеми от текстова информация и може да послужат за подготовка на ресурси за КГТ.

Изводи

От направения обзор е видно, че компютърното генериране на текст е област със специфична теоретична платформа и полезни прагматични реализации, област, която тепърва предстои да разгърне потенциала си. От една страна за КГТ са важни солидните теоретични изследвания свързани с лингвистичната същност на изходния продукт- текстът на естествен език, от друга страна своите аргументи има и едно по-технократско отношение към КГТ, което въвлича по специфичен начин някои от съвременните семантични технологии. Безспорно е, че и двата подхода изискват солидни лингвистични ресурси представени в една или друга форма според избрания метод за генериране. Без такова представяне на лингвистични знания, което по същество са лингвистичните ресурси, не може да съществува компютърно генериране на текст.

Най-богати и детайлизирани са наличните лингвистичните ресурси, чрез които се продуцират текстове на английски език и все още КГТ на друг естествен език е сериозно изследователско предизвикателство. Поради това основна идея на настоящата дисертация е изследването как българският език може да бъде обект на анализ и моделиране, за да се създаде лингвистичен ресурс за КГТ на български език.

Паралелното изследване на няколко естествени езика през призмата на дадена теоретична парадигма, както и подпомагането при паралелно създаване на приложни лингвистични ресурси е изключително полезно. На базата на един съществуващ солиден линвистичен ресурс повторението на процедурата по анализ и формализиране на езиковите феномени е значително улеснена. Това дава възможност за обръщане на специално внимание на разликите в естествените езикови системи и фина настройка на специфичните явления за всеки добавен език.

За настоящата дисертация като теоретична основа на КГТ на български език е избрана парадигмата на Системично-функционалната лингвистика, тъй като тя е една от основните най-успешни постановки за КГТ. Системично-функционалната лингвистика предлага изключителни възможности за паралели при системично-функционалните описания на различни естествени езици, по-бърза и по-ефективна формализация и реализация на лингвистични знания при създаване на система за КГТ за новоразглеждан език. Специфичното при системично- функционалния възглед за КГТ е организирането на целия процес на генериране около комуникационните цели, а не около граматическите структури на конкретния естествен език. На това се базира идеята за изграждане пространство на търсене от системични мрежи за генериране на български език, което да представлява ресурс наречен Приложна системично-функционалната граматика за български език. За реализиране на граматиката е необходимо моделиране на някои лингвистични феномени на българския език по отношение на Системично-функционалната лингвистика. На базата на теорията на [Halliday, 1994] системично- функционалните граматики са фокусирани върху кодиране на семантичните връзки в текста във функционална форма и са насочени към директни съответствия между по-високите нива на организация на текста, т.е. планиращите процеси, и граматическия компонент. Това е предпоставка за научноприложно изследване за генериране на няколко типа текстове- технически инструкции на български език. Достъпът до Средата за многоезиково генериране KPML подпомага работата по създаване, тестване и настройки на приложни системично- функционални граматики на различни естествени езици.

Както става ясно от направения обзор на областта, компютърното генериране на текст не може да бъде реализирано без семантични ресурси. Естествено за КГТ е използването на онтологии за предметната област на генерирането, а също и висши онтологични модели за лингвистичните знания, така че онтологичното инженерство е до голяма степен свързано с КГТ. Семантичното анотиране на текстове на естествен език има приложение в КГТ като практически значима подпомагаща дейност. Семантичното анотиране представлява интерес особено за естествени езици различни от английски поради по-малкия брой изследвания и работещи реализации. Поради това изследователско направление в настоящата дисертация е представяне на знания в онтологични модели и семантично анотиране на специализирани текстове на български език.

2. Подготовка на лексико-граматически ресурс за компютърно генериране на български текстове

Интерес към компютърното генериране на български език са проявявали изследователи от областта на компютърната лингвистика още от времето на първите по-сериозни приложни опити в обработването на естествен език. Идеята на Руслан Митков за генерираща система [Mitkov, 1990] е свързана с генериране на описания на основните геометрични термини на български език. По отношение на моделирането на българския език, особено през последните години има интересни разработки свързани с машинния превод, например [Ranta, Angelov, Hallgren, 2010]. Съществува опит в генерирането на изречения на български език от концептуални графи [Bontcheva and Angelova, 1996]. Генериране на текстове на български език, при това като един от естествените езици в паралелно многоезиково генериране, е реализирано по проекта AGILE (1998-2001), което е обсъдено подробно в дисертационния труд.

2.1. Моделиране на някои езикови явления в термините на Системично-функционалната граматика

За да се генерира автоматично отделно изречение на даден език е достатъчно генераторът да разполага само с нужната лексическа и граматическа информация за реализираните езиковите явления в това изречение. По тази логика е определен неголям корпус от целеви текстове по отношение на генерирането, за да се изследват специфичните езикови явления проявени конкретно в избрания тип текстове, а също да се фиксира нужният лексически материал за генериране в избраната област- инструкции за CAD/CAM софтуер.

Текстовият корпус съдържа за българската си част 9 процедури, 194 изречения (клаузи), 1219 думи. Извършен е детайлен анализ на текстовия корпус в три стъпки:

Първо, определяне на граматическите конструкции използвани в процедурите.

Второ, проучване на начините за описване на езиковите явления в термините на Системично-функционалната лингвистика.

Трето, разработване на формални спецификации за основните граматически функции.
Разполагайки с дадения обем моделирани езикови явления работата нататък е организирана в посока създаване на принципно построено множество от ресурси за описание на една цялостна граматика, като това множество представлява детайлизация, работеща в избраната предметна област.

Системично-функционалната граматика като системична мрежа дава възможност да се определят регионите, в който се сглобяват типичните конструкции на предметната област и в същото време да се моделират така, че да бъдат пре-използваеми за всеки друг процес на генериране, тъй като отразяват принципен граматически модел. Да се развиват, обаче, изчерпателно всички клонове на достигнатите граматически системи е подход, който отнема

много време и затова в рамките на проекта AGILE не се цели построяване на цялостни приложни граматики. Това е причината предложеното тук моделиране на езиковите явления да е частично, но обхващащо всички регистрирани в текстовия корпус явления, за да бъде възможно автоматичното генериране на текст на български език.

В [Стайкова и Пенчев, 2000] е направена първата малка стъпка за въвеждане на терминологията за системично-функционален анализ на български език, но тук ще се придържаме към утвърдената английска терминология за описване на езиковите явления като системични мрежи. Теорията на системичните мрежи представя естествения език като ресурс за създаване на смисъл. Всяка система в системичната мрежа дава избор да се изрази един или друг смисъл на разглеждания в системата аспект. Системата се състои от (1) входно условие, което показва къде се прави избора, (2) множество от възможни изходи, и (3) реализации, които показват какви са структурните следствия в езика за всеки от възможните изходи на системата. Като пример за записване на система по-долу имаме системата-вход към дадена граматиката, която определя какъв по ранг е описваният езиков елемент.

RANK:
(start) →
[clauses],
[groups-phrases],
[words] (+Stem) (Stem:Morphems),
[morphems] (+Head).

Ниво изречение

Съставени са системични мрежи след анализ на следните явления в българския език и в частност в текстовия корпус::

Определени са типовете процеси в текстовия корпус

Ключови системи:

PROCESS-TYPE:
(transitivity-unit) →
[material] (Process::do-verb),
[mental] (Process::experience-verb),
[verbal] (Process::symbolic-verb),
[relational] (Process::relational-verb).

AGENCY:
(transitivity-unit) →
[middle] (Process::middle-verb),
[effective] (Process::effective-verb).

Обстоятелства на процесите

Ключова система е TYPE-OF-CIRCUMSTANCE. Примерите за обстоятелства от текстовия корпус са най-често в материални процеси, към които е асоциирано обстоятелствено пояснение за място. Отбелязани са също срещания на изрази за релацията част-цяло (*крайна точка на дъгата*) и на изрази за релацията за съотнасяне (*Въведете ъгъл спрямо допирателната.*)

Диатезност

Според системично-функционалния анализ ядро на функционално-семантичното поле Диатезност (Diathesis) е категорията залог. Залогът може да бъде описан като връзка между транзитивните функции изразени чрез ролите на участниците в процеса и функциите за

деятелността в процеса: Агент/Agent и Медиум/Medium [Halliday, 1994, стр.161].

Освен чрез морфологичната категория за залог пасивност може да бъде изразена и чрез "средна конструкция":

Точката	*се задава*	*от потребителя.*
Цел/Goal	Процес:материален	Агент/Agent
Медиум/Medium	Финитив	Актор/Actor
Подлог/Subject	Залог: страдателен	

Модалност

Модалността има най-характерно проявление на нивото на изречението в императивното или индикативно наклонение на глагола. В текстовия корпус броят на императивните и на индикативните изречения е почти изравнен заради стила на текстовете-инструкции, които го съставляват. В други текстове императивното наклонение на глаголите не би било толкова застъпено. Използвана е само учтивата заповедна форма на глагола, 2-ро лице, мн. число. В корпуса няма въпросителни изречения, нито реализации на условно наклонение.

Ключовата система е наречена MOOD-TYPE. Чрез нея изреченията се разделят на индикативни/ indicative и императивни/ imperative.

MOOD-TYPE:
(independent-clause-simplex) →
    [indicative],
    [imperative] (+Finite).

Анализирани са характериатики на индикативните и на императивните изречения.

Темпоралност

В съвременното езикознание се приема, че българският език притежава девет глаголни времена: сегашно, минало свършено, минало несвършено, минало неопределено, минало предварително, бъдеще, бъдеще в миналото, бъдеще предварително и бъдеще предварително в миналото [Бояджиев, Куцаров, Пенчев, 1999г.]. От тях само две се срещат в изследвания текстов корпус- сегашно време и минало неопределено. Срещанията на сегашно време са най-многобройни. Ключова система:

TENSE-SYSTEM:
(clause-simplex) →
    [past],
    [future],
    [present] (Finite:::present-form).

Завършеност на процеса

Видът на глагола е лексикално-граматична категория характерна за славянските езици, в това число и за българския език. Приема се, че видът на глагола е лексикално-граматична характеристика [Бояджиев, Куцаров, Пенчев, 1999]. В Системично-функционалната теория видът на глагола е функция на Предикатора.

По отношение на използването на различни по вид глаголи в изследвания корпус може да се каже, че както несвършения вид, така и свършения вид се наблюдават в разглежданите текстове. Вида на глагола не се влияе от преобладаващото използване на повелително наклонение.

ASPECT е наименование на новата система за моделиране на тази характеристика и се отнася към Предикатора, следователно, трябва да бъде добавена на нивото на изречението:

ASPECT:
(clause-simplex) →
[perfective] (Process::perfective-verb)
[imperfective] (Process::imperfective-verb)

Ниво сложно изречение

В системично-функционалната граматика на Халидей се поддържа представата, че на сложното изречение/ clause complex може да се гледа като на комплекс от изречения, точно както групата може да бъде възприемана като „комплекс от думи“. Сложността в организацията на сложните изречения произтича от различните начини, по които простите изречения могат да бъдат свързвани.

Различават се две измерения, за да се предложат по-детайлни описания за това как точно простите изречения модифицират главното изречение. Първото измерение се занимава със системата на взаимозависимост, дали е паратактична (съчинение) или хипотактична (подчинение). Другото измерение е логико-семантично и се занимава с експанзията и проекцията. Те са разгледани подробно в дисертацията като се подчертава идеята, че сложните изречения възникват като резултат от взаимодействието на двете измерения.

Сложните изречения от корпуса с текстове-инструкции са анализирани и са обсъдени системите изграждащи системична мрежа за моделиране на сложни изречения според особеностите на българския език. Заради семантичната абстракция при свързване на прости изречения в паратактични или хипотактични комплекси, както и боравейки с логико-семантичното измерение, се твърди, че при съставянето на целевите сложни изречения в избрания контекст на текстове-инструкции няма специални случаи за езиковата система на българския език. Направен е изводът, че бихме могли до голяма степен да се възползваме и за българския език от разработената в СФГ теоретична системично-функционална мрежа.

2.2. Лексико-граматически ресурс за генериране на български текстове: Приложна българска системично-функционална граматика

Приложната българска системично-функционална граматика е ресурс за генериране на български текстове създаден в средата за многоезиково генериране KPML като резултат от работата по проекта AGILE. В средата KPML при създаване на нови граматики се насърчава подход на сравнителен анализ с приложната английска граматика NIGEL или с подходящи приложни ресурси за други езици, например, немски или френски.

Изключително полезно се оказа сътрудничеството при сравнителния анализ на трите славянски езика- руски, български и чешки. Цел на изследователската работа е създаването на ресурси за трите славянски езика, които да бъдат пре-използваеми за генериране в различни предметни области. Изчерпателният анализ за всички възможни изходи на системичната мрежа отнема много време и труд, така че в рамките на проекта приложният резултат засяга основно идентифицираните функционални полета при анализа на конкретния текстов корпус. Този подход предлага възможност във всеки момент приложният ресурс да бъде детайлизиран и обогатен с нови възможности за генериране.

Преди проекта AGILE не са известни генериращи системи, които да предлагат пре-използване на ресурсите за генериране в нови приложни области, така че приложеният подход към създаване на граматиките е от полза за цялата общност. Приложната българска системично-функционална граматика е достъпна в Банката по генериране[4] поддържана за средата за многоезиково генериране KPML. Ресурсът притежава представително множество от генерирани примери, които дават представа за функционално-системичните полета, които са разработени.

Като изчислителен ресурс една приложна граматика в средата KPML представлява

[4] http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/genbank/generation-bank.html

системична мрежа организирана в полета (regions) от свързани системи, като полетата са условно разделени и системите в едно поле имат отношение към едно и също граматическо явление. Начало на системичната мрежа е системата Ранг, в която се правят първите избори за подадения като вход семантичен израз.

Изборите се направляват от входните данни и от логиката за правене на автоматични избори, която се реализира чрез избирателите /choosers. Всяка система има свой избирател който насочва процеса на генериране през някой от изходите на системата. Понякога избирателите се нуждаят от специфична или динамична информация, за да направят избор и ползват свои проучватели/ inquiry, за доставяне на такава информация за текущия процес.
Генерирането се инициира и направлява от системичните характеристики на очаквания генериран текст, зададени във входно описание –израз на езика за планиране на изречения Sentence Planning Language. Всички термини, които присъстват във входните изрази са дефинирани или в онтологията от високо ниво Обобщен модел (GUM) или в предметната онтология за конкретната приложна област.

В настоящия раздел от дисертацията са показани системи и избиратели от Приложната българска системично-функционална граматика за някои от специфичните езикови явления разгледани в раздел 2.1. Чрез различни примери е демонстрирано генерирането на изречения съдържащи разнообразни езикови явления, чието моделиране е показано в предходния раздел. Описаните модели са приложени в граматиката и тя ги обработва адекватно. Това включва генериране с деятелен и страдателен залог, със свършен и несвършен вид на глагола, с императивно и индикативно наклонение, номинализация на глаголната група, членуване, съгласуване на елементите в групата на съществителното по род и число. Накрая са показани синтактичните конструкции на генерирани сложни изречения.

Фигура 2 Генерирана структура за „Начертайте линия.“

Типове процеси. Транзитивност и диатезност (залог)

В дисертацията е показано формалното програмно представяне на системата и избирателя за тип на процес, от които започва спецификацията в граматическото поле транзитивност, и по-точно транзитивност на не-релационни процеси (nonrelationaltranzitivity).

Примерите за генериране на изречение с ефективен материален процес са най-често срещани в целевия корпус. С дадения по-долу SPL за изречението „Начертайте линия.“ се генерира синтактичната структура показана на Фигура 2, където като характеристики на процеса присъстват do-verb и effective-verb. Първата характеристика е резултат от преминаване на процеса на генериране през изхода MATERIAL на системата PROCESS-TYPE (Тип на процес), чието аналитично представяне е показано в предходния раздел.

```
(S/ DM::DRAW
:SPEECHACT          IMPERATIVE
:ACTOR (HEARER / DM::USER     IDENTIFIABILITY-Q        IDENTIFIABLE)
```

За изречението „Начертайте линията.“ Избирателят е насочил процеса на генериране към изхода EFFECTIVE на системата AGENCY. Със следващия SPL се демонстрира преминаване на процеса на генериране през алтернативния изход: MIDDLE. Това е нужно, когато се реализират средни процеси, какъвто е случая с процеса появявам-се /appear, например в изречението „Линията се появява.“

```
(S / DM::APPEAR
:ACTOR  (L / DM::LINE
        :IDENTIFIABILITY-Q  IDENTIFIABLE))
```

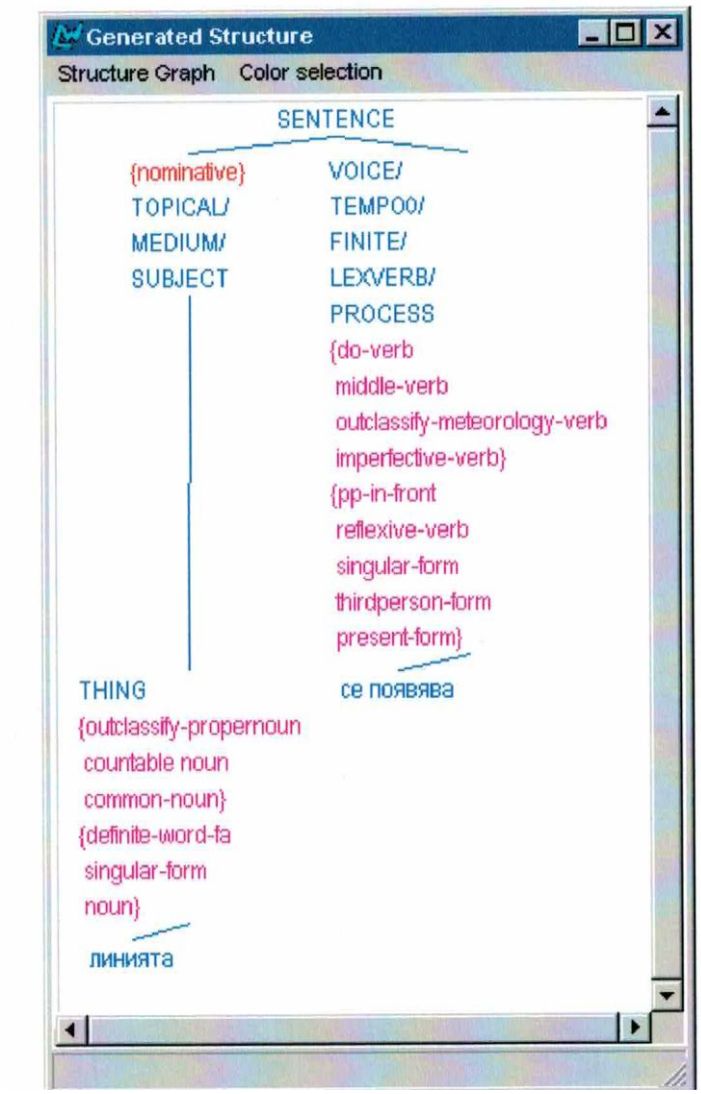
По този начин е проследен пътят на процеса на генериране в системичната мрежа на ресурса с такива входни SPL-и, които отразяват анализираните езикови явления за българския език. Алтернативните избори на пътища през системичната мрежа дават разнообразието на възможни продукции при генериране. В дисертацията са показани съответните изходни продукции на зададените SPL-изрази.

Входното SPL- задание за генериране изречението „Чертае се линия.“ :

```
(S / DM:: DRAW
:PREFER-MENTION-AGENT-Q    WITHHOLD
:ACTEE   (AE / DM::LINE))
```

SPL-израз задаващ реализация на деятелен залог в изречението “Потребителят чертае линия.“:

```
(S / DM::DRAW
:ACTOR        (AR / DM::USER
              :IDENTIFIABILITY-Q  IDENTIFIABLE)
:ACTEE        (AE / DM::LINE))
```



Фигура 3 Генерирана структура за „Линията се появява.“

Модалност. Наклонение

Реализацията на речевия акт се отразява от елемента на Модуса: двойката Subject-Finite. Когато Актьорът е и слушател при речевия акт в семантичния вход за генерирането се задава роля на слушател. SPL за генериране на изречението „Вие чертаете линия.“:

```
(S / DM::DRAW
:ACTOR      (HEARER / DM::USER
            :IDENTIFIABILITY-Q  IDENTIFIABLE)
:ACTEE      (AE / DM::LINE))
```

Българската приложна граматика има възможност за генериране на функцията модалност, т.е. изразяване на възможност, необходимост и т.н. Въведена е типичната „да-конструкция“ за реализиране на модални глаголни групи като „мога да...“, „трябва да...“ и т.н. Това се демонстрира от следващият семантичен вход, чрез които се реализира изречението „Вие можете да чертаете линия.“

(S / DM::DRAW
:MODAL-PROPERTY-ASCRIPTION GENERAL-POSSIBILITY
:ACTOR (HEARER / DM::USER
:IDENTIFIABILITY-Q IDENTIFIABLE)
:ACTEE (AE / DM::LINE))

Темпоралност

Моделираното граматическо време в рамките на поставените цели е сегашно време. Характеристиката present-form за глаголите, реализиращи процеси в дадените примери присъства в синтактичната структура на всяко от изреченията (виж Фигура 2 и Фигура 3 по-горе).

Завършеност на процеса. Вид на глагола

Разнообразието: свършен и несвършен вид за един и същ процес водят към различни реализации. Например, за процеса DRAW са достъпни две реализации свързани с алтернативите за вид: за несвършен вид процесът се реализира с глагола „чертая“, за свършен вид се използва реализация с глагола „начертая“.

В SPL-а това се сигнализира с израза

:ASPECT-Q PERFECTIVE-ASPECT

Съответните характеристики уточнени по време на генерирането (perfective-verb и imperfective-verb) могат да бъдат видени при реализациите на изреченията „Начертайте линия.“ и „Чертае се линия.“

Текстова кохерентност/ свързаност

Текстова кохерентност/ свързаност, реализирана с текстови съюзи / Textual Conjunctive, е демонстрирана чрез вариации на изречения като: „Първо изберете точка.“ , „След това въведете ъгъл“, „Накрая начертайте линия.“ SPL-представянето за последното изречение е следното:

(S/ DM::DRAW
:SPEECHACT IMPERATIVE
:ACTOR (HEARER / DM::USER IDENTIFIABILITY-Q IDENTIFIABLE)
:ACTEE (AE / DM::LINE)
:CONJUNCTIVE SEQUENCE-LAST)

Номинализация

Номинализирана глаголна група „чертая линия“ се реализира със сления SPL-вход:

(S / DM::DRAW
: EXIST-SPEECH-ACT-Q NOSPEECHACT
:ACTEE (AE / DM::LINE))

Членуване и съгласуване по род и число в групата на съществителното

SPL-изразът идентифициращ членуването в дадена група на съществителното е
IDENTIFIABILITY-Q IDENTIFIABLE
Генерираните структури на изречения показват с разнообразни случаи по отношение на членуването (пълен и непълен член за съществително, пълен и непълен член за прилагателно от мъжки род в групата на съществителното, членуване при женски род).

Сложни изречения

Полето от приложната граматика, което съдържа системите, избирателите и проучвателите за генериране на сложни изречения се нарича **CLAUSECOMPLEX.** Преминаването на процеса на генериране през това поле е онагледено в дисертацията с генерираните синтактични конструкции от SPL-ите за следните изречения:

За изречението „Натиска се RETURN, за да се затвори полилинията.“:

```
(S / RST-PURPOSE
:PREFER-MENTION-AGENT-Q  WITHHOLD
:DOMAIN (D / DM::PRESS
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::RETURN))
RANGE (R / DM::CLOSE-SCREEN-OBJECT
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::POLYLINE   IDENTIFIABILITY-Q       IDENTIFIABLE)))
```

За изречението „Натиснете RETURN, за да затворите полилинията.“:

```
(S / RST-PURPOSE
:SPEECHACT  IMPERATIVE
:DOMAIN (D / DM::PRESS
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::RETURN))
RANGE (R / DM::CLOSE-SCREEN-OBJECT
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::POLYLINE   IDENTIFIABILITY-Q       IDENTIFIABLE)))
```

За изречението „Натиснете RETURN и затворете полилинията.“:

```
(S / CONJUNCTION
:SPEECHACT  IMPERATIVE
:DOMAIN (D / DM::PRESS
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::RETURN))
RANGE (R / DM::CLOSE-SCREEN-OBJECT
        :ACTOR (HEARER / DM::USER    IDENTIFIABILITY-Q       IDENTIFIABLE)
        :ACTEE (AE1 / DM::POLYLINE   IDENTIFIABILITY-Q       IDENTIFIABLE)))
```

Приложната българска системично-функционална граматика е създадена на базата на обширния ресурс за генериране на английски изречения NIGEL. Използвани са съществено изложените в предходния раздел 2.1 системично-функционални модели за някои граматически явления в българския език.

Демонстрирани са възможностите на създадения ресурс да генерира сравнително широк обхват от алтернативни граматически характеристики в изреченията (напр. деятелен и страдателен залог, императивно и декларативно изречение, просто и сложно изречение, няколко вариации на хипотактични и паратактични връзки в сложното изречение, номинализация на глаголна група и др.)

Постигнатият обхват на генериране е прагматично ограничен от конкретния стил на текстовете от корпуса: процедурни текстовете-инструкции. От друга страна, ресурсът е достъпен за пре-използване и разширяване на генерираните вариации, тъй като се базира стриктно на системично-функционалното моделиране на българския език, което отчита лингвистичната природа на изходния продукт от генерирането- добре формулираното

българско изречение.

3. Реализация на процеса на генериране на български текстове

Тази глава от дисертацията описва самия процес на автоматично генериране на текстове-инструкции на български език, както той е конструиран и изпълнен от системата AGILE.

Процесът на задаване на параметри за автоматичното генериране започва с въвеждане на понятията от предметната област на бъдещите текстовете, това представлява дефинирането на модела на предметната област на CAD/CAM-приложенията. Осъществява се свързване на тези понятия с Обобщения модел, онтологията от високо ниво, която осигурява съответствието между понятията от предметна област и тяхната системично-функционална проекция по отношение на българския език (раздел 3.1).

Определянето на съдържанието на бъдещия текст се извършва от потребителя на системата с помощта на текст-структуриращи елементи, които са описани в раздел 3.2.

След получаване на входните данни за процеса на автоматично генериране работи модулът за планиране на текст, а след него и планировчикът на изречение, които създават плановете на всяко отделно изречение като част от цялостен кохерентен текст. Механизмът е разяснен в раздел 3.3 заедно с идеята за създаване на няколко текстови стилови разновидности на изходния текст.

Лексико-граматическата реализация на автоматичното генериране на текст се извършва в средата за многоезиково генериране KPML от приложната българска системично-функционална граматика. Описанието на реализирания процес на автоматично генериране на текстове- инструкции на български език е завършено с коментар на резултата от автоматичното генериране на 5 различни по стил текстове на български език от конкретен зададен вход за системата (раздел 3.4)

3.1 Представяне на знания за генерирането на български текстове

Избраната предметна област е свят на CAD/CAM термини и инструкции. Източник на текстове-инструкции за текстовия корпус са ръководствата за ползване на CAD/CAM софтуер, които се състоят главно от указания за създаване на различни видове проекти и чертежи, за съхраняване на файловете на даден проект, за повторното зареждане на тези файлове и т.н.

Базови елементи на тази предметна област са понятия като ФАЙЛ, КЛАВИШ, ЕКРАН, различни видове графични обекти: ЛИНИЯ, ДЪГА и т.н. основните действия в областта се извършват от потребителя: ЧЕРТАЯ, ЗАПАЗВАМ (файл), НАТИСКАМ (клавиш) и т.н. Моделът на предметната област съдържа също текст-структуриращи елементи, които са градивни блокове на бъдещите процедурни текстове.

Елементите от предметната област са разделени в две групи:
ПРОЦЕСИ (ДЕЙСТВИЯ_на_ПОТРЕБИТЕЛЯ, СЪБИТИЯ и т.н.) и
НЕЩА (КЛАВИШ, ЕКРАН, ЛИНИЯ, МУЛТИЛИНИЯ и т.н.)
Текст-структуриращите елементи са четири на брой и предоставят контекст, в който действията и събитията могат да се случат. Текст-структуриращите елементи са: ПРОЦЕДУРА, МЕТОД, СПИСЪК_от_ПРОЦЕДУРИ и СПИСЪК_от_МЕТОДИ и са формално дефинирани, така че да моделират структурите на текстовете от целевия корпус.
В текста на дисертацията и в Приложение 1 са дадени тези формални дефиниции и конструкцията на модела на предметната област.

Използване на Обобщения модел

В обзорната част на настоящата дисертация (раздел 1.2.1) е посочена ролята и важността на онтологията от високо ниво Обобщен модел в процеса на интерпретиране на съдържанието

при автоматично генериране на естествен език. Обобщеният модел е практически използван като посредник между нивата с по-висока абстракция в процеса на текст-генериране (виж Фигура 1: Абстрактна система за КГТ) и "по-ниските" нива на лексико-граматическата реализация.

В работата по компютърното генериране на български език е направен системично-функционален анализ на понятията от предметната област на CAD/CAM - приложения и беше установено, че понятията могат адекватно да бъдат класифицирани в онтологията по отношение на реализацията си с езикови конструкции на български език. Това означава, че не е необходима никаква корекция в Обобщения модел, за да бъде той използван в процеса на автоматично генериране на български текстове.
По отношение на един цялостен анализ за съответствие на конструкциите на Обобщения модел и езиковите конструкции на българския език е необходима още сериозна изследователска работа. Можем да твърдим единствено, че е направена първата малка стъпка с въвеждане на терминологията на Системично-функционалния анализ на български език в [Стайкова и Пенчев, 2000]. Това е съществено, тъй като конструкциите на Обобщения модел кореспондират пряко с постановките на Системично-функционалната граматика на [Halliday, 1994].

3.2. Определяне на съдържанието

Едно от предизвикателствата пред създателите на системата AGILE е автоматичното генериране на кохерентни текстове от съдържание зададено от потребителя на системата. На евентуалния потребител е предоставен интерфейс, чрез който да бъдат съчетани и подредени термини от модела на предметната област на CAD/CAM приложения. Това са понятия за елементи описани в раздел 3.1, например: ЛИНИЯ, ДЪГА, КЛАВИШ, ПОКАЗВАМ, ЧЕРТАЯ, НАТИСКАМ, ПОТРЕБИТЕЛ и т.н.

Потребителят на системата може да съставя текстове- инструкции. За да се групират и подредят някакъв набор от термини на предметната онтология, така че да изразяват търсено съдържание в модела на областта са предоставени и *текст-струкуриращи елементи*. От анализа на текстовия корпус изложен в раздел 2.1 става ясно, че целевите текстове-инструкции най-често се състоят от определени изграждащи ги елементи, например цел, списък от стъпки за постигане на целта, странични ефекти и т.н. Чрез входния интерфейс потребителят на системата задава комбинация от определените текст-структуриращи блокчета като ги изпълва с определено съдържание - конкретни понятия от модела на предметната област на CAD/CAM приложенията. Така чрез конкретен набор от текст-структуриращите елементи се описва *какво* да бъде реализирано в изходния текст.

Определянето на съдържанието на бъдещите текстове в системата се извършва чрез затворено планиране, което означава, че то изцяло зависи от решенията и намесата на потребителя на генериращата система. Потребителят задава комбинация от текст-структуриращи елементи и ги запълва със желаното съдържание изразено чрез понятия дефинирани в онтологията на предметната област.

Следващият пример представя структурата от дефинираните текст-структуриращи елементи, с която се определя връзката цел-подцел в изречението „Изберете Save, за да запазите чертежа.":

```
процедура
 ЦЕЛ 'запазвам чертеж'
 МЕТОДИ списък_от_методи
         ПЪРВИ метод
                ПОД-СТЪПКИ списък_от_процедури
                          ПЪРВИ процедура
                                 ЦЕЛ 'избирам Save'
```

Формално представяне на съдържанието на примерната процедура наречена „Чертане на дъга“ е предоставено в приложение към дисертацията.

3.3. Планиране на текста

От всеки зададен вход за процеса на генериране се очаква системата AGILE да предостави на изход автоматично генериран кохерентен текст, а ако е възможно и няколко типа автоматично генерирани текстове. Обработването на входното съдържание, т.е. на запълнените текст-структуриращи елементи зададени от потребителя, се извършва от модул за текст-структуриране, чиято работа има два етапа. Първо от зададеното съдържание се създава план на бъдещия изходен текст като цяло, а след това планировчикът на изречения създава и плановете на отделните изречения (т.н. SPL-и) на базата на готовия текстов план.

Текст-структуриращите елементи описват само съдържанието на бъдещия текст, т.е. *какво* да бъде реализирано. За да се контролира *как* да бъде реализирано съдържанието са въведени текстови шаблони, които дефинират *стил на текста* като задават параметрите на реализацията в лексико-граматиката за определени елементи на текстовата структура. Характерен пример е заглавието на текста-инструкция съдържащо се в слота ЦЕЛ на текст-структуриращия елемент ПРОЦЕДУРА, което се реализира чрез номинализация на зададения глагол. В края на този раздел са изложени възможностите предлагани от системата AGILE за отразяване на определени стилови предпочитания за изходния текст.

Модулът за текст-структуриране работи с две групи понятия: текст-структуриращите елементи описани в предходния раздел и текстови шаблони, които представляват градивните единици на текстовия план. Текстовите шаблони дават разпределението на съдържанието и дефинициите им са извлечени от терминологията ориентирана към потребителя, към която обикновено се придържат авторите на технически ръководства.

Елементи на текстовите шаблони:

ДОКУМЕНТ-ЗАДАНИЕ
Елементът е дефиниран чрез два слота:
ЗАГЛАВИЕ_на_ЗАДАНИЕТО (задължително)
ИНСТРУКЦИИ_за_ ЗАДАНИЕТО (задължително)
ИНСТРУКЦИИ_за_ ЗАДАНИЕТО означава поне една инструкция.

ИНСТРУКЦИЯ
Елементът е дефиниран чрез три слота:
ЗАДАЧИ (задължително)
ОГРАНИЧЕНИЕ (незадължително)
ПРЕДУСЛОВИЕ (незадължително)

ЗАДАНИЕ
Елементът е дефиниран чрез два слота:
ИНСТРУКЦИИ (задължително)
СТАРНИЧЕН_ЕФЕКТ (незадължително)

Съпоставянето на текстовите шаблони и текст-структуриращите елементи на входния интерфейс за съдържанието дава възможност за създаване на правила за текст-структуриране.

Изречението „Изберете Save, за да запазите чертежа.“ може да бъде представено чрез елементите на текстовите шаблони по следния начин:

ЗАГЛАВИЕ_на_ЗАДАНИЕТО “запазване на чертеж”
ИНСТРУКЦИИ_за_ЗАДАНИЕТО
ИНСТРУКЦИЯ
ЗАДАЧИ
ЗАДАНИЕ “избирам Save”

Текстовите шаблони се използват за предварително определяне на избора на граматически средства за реализиране на съдържанието и по този начин може да бъде определян стила на текста. С такава цел в шаблоните могат да се въведат и допълнителни маркери на дискурса, за да се направи експлицитна текстовата структура. Независимо от конкретния текстов шаблон могат да съществуват избори, които да влияят на нивото на експлицитност, например, за информиране на потребителя на генерирания текст за наблюдаваните странични ефекти.

При анализа на текстовия корпус за системата AGILE са забелязани следните характеристики и варианти:

-Явно изразяване или скриване на агента извършващ действието;

-Различни начини да се изрази връзката с читателя – дали читателят да е специално адресиран или не;

-Режим на реализиране на инструкции;

-Сложност на лингвистичните изрази.

Прецизната работа с тези аспекти дава възможност за определяне на вариациите в текстовия стил. Въведени са два различни стила, в които могат да бъдат реализирани инструкциите: *персонален императивен* и *безличен изявителен*.

Определени са изразите, които се включват в SPL-заданията, за да настроят различните аспекти на определените стилове. Получаването на различни SPL-конструкции от зададения семантичен вход за различните стилове се контролира чрез средствата за текст-структуриране на средата KPML, а именно израза REALIZE-WITH. Например, ако искаме да видим съдържанието на елемента TASK-TITLE реализирано с номинална група, то можем да зададем следния израз SPL-конструкцията:

(REALIZE-WITH TASK-TITLE
(:EXIST-SPEECH-ACT-Q NOSPEECHACT))

В допълнение към ограниченията за лексико-граматическите реализации средата KPML дава възможност да се специфицира *layout* за определени елементи на текстовата структура. Модулът за текст-планиране е реализиран в стила на приложните системично-функционални граматики на средата за многоезиково генериране KPML.

След като разполагаме с план на текста, който включва ограниченията за начина на реализация на неговите съставни части вече могат да бъдат създадени SPL-и за отделните изречения като се използват тези ограничения. Това се извършва от Планировчика на изречения.

Една интересна задача на Планировчика на изречения е агрегацията на изреченията. Тук агрегацията представлява всъщност комбиниране на два или повече SPL-фрагмента в по-голям фрагмент от SPL-код. Всяка отделна част определя плана за реализиране на елемент от потребителския вход, който може да има и зададен стил. На практика всеки SPL-фрагмент определя изречение. Когато става въпрос за комбиниране на SPL-и, то трябва да се вземе решение за сложността на изречението, което да изрази зададеното съдържание.

Планировчикът на изречения преминава през листата на създадената дървовидна структура на текста последователно като създава SPL-код за семантиката идентифицирана в тези листа.

Първо зададеното съдържание се превежда в чисто същностно съответстващ му SPL-код, след това се определят границите на отделните изречения и се формира SPL-код за всяко изречение.

След това се добавят ограниченията в реализацията изисквани от стила на текста чрез описаните по-горе специфични SPL-фрагменти. Така плановете на изреченията носят информация не само какво да бъде реализирано, но и как да стане това. По този начин подадените за генериране от тактическия генератор SPL-и на изречения представляват всъщност код на кохерентен текст.

3.4. Лексико-граматическа реализация

Стиловото разнообразие на текстовете генерирани на български език е представено в следващата Таблица 1.

Стил на текста	Кратки команди номериран списък	По-дълги изречения	По-дълги изречения с обяснения
Персонален императивен	Вариант 1	Вариант 2	Вариант 4
Безличен изявителен		Вариант 3	+ допълнителна агрегация Вариант 5

Таблица 1 Стил на генерирания текст

Двете основни вариации представени от редовете на таблицата, персонален императивен стил и безличен изявителен стил, се отнасят за начина на реализиране на инструкциите в текстовете, например „Въведете ОК.“ или „Въвежда се ОК“.

Колоните отразяват различни варианти на организацията на текста - представянето на списък от команди като номериран списък с елементи записани всеки на отделен ред, или вариации на агрегацията. Вариациите на агрегацията са всъщност различно разпределение на съдържанието в изречения. Това може да включва генериране на „изречения с обяснения“, или, с други думи, генериране на информацията за страничните ефекти.

Следват генерираните изходни текстове в петте вариации за зададено едно и също примерно съдържание на желания текст.

Вариант 1: Персонален-императивен текст с кратки команди

Чертане на полилиния, съставена от отсечки и дъги

1. Стартирайте командата PLINE.
2. Задайте началната точка на отсечка.
3. Задайте крайната точка на отсечката.
4. Въведете **a**. Изберете OK.
5. Задайте крайната точка на дъгата.
6. Задайте трета точка на дъгата.
7. Въведете **l**. Изберете OK.
8. Въведете дължина на отсечка**.**
9. Въведете ъгъл на отсечката спрямо допирателната в крайната точка на дъгата.
10. Натиснете Return.

Вариант 2: Персонален-императивен текст

Чертане на полилиния, съставена от отсечки и дъги

1. Стартирайте командата PLINE, като използвате един от следните методи:

Windows: От плаващото меню Polyline на функционалния ред Draw изберете Polyline.

DOS и UNIX: От менюто Draw изберете Polyline.

2. Задайте началната точка на отсечката.
3. Задайте крайната точка на отсечката.
4. Въведете **a**, за да превключите на режим Arc**.** След това изберете OK в диалоговия прозорец на режима Arc.
5. Задайте крайната точка на дъгата.
6. Задайте трета точка на дъгата.
7. Въведете **l**, за да се върнете в режим Line. След това изберете OK в диалоговия прозорец на режима Line.
8. Въведете дължина на отсечка от крайната точка на дъгата.
9. Въведете ъгъл на отсечката спрямо допирателната в крайната точка на дъгата.
10. Натиснете Return, за да завършите полилинията.

Вариант 3: Безличен-изявителен текст

Чертане на полилиния, съставена от отсечки и дъги

1. Стартира се командата PLINE, като се използва един от следните методи:

Windows: От плаващото меню Polyline на функционалния ред Draw се избира Polyline.

DOS и UNIX: От менюто Draw се избира Polyline.

2. Задава се началната точка на отсечката.

3. Задава се крайната точка на отсечката.

4. Въвежда се **a** за превключване на режим Arc. След това, в диалоговия прозорец на режима Arc се избира OK.

5. Задава се крайната точка на дъгата.

6. Задава се трета точка на дъгата.

7. Въвежда се **l** за връщане в режим Line. След това, в диалоговия прозорец на режима Line се избира OK.

8. Въвежда се дължината на отсечката от крайната точка на дъгата.

9. Въвежда се ъгълът на отсечката спрямо допирателната в крайната точка на дъгата.
10. Натиска се Return за завършване на полилинията.

Вариант 4: Персонален-императивен текст с обяснения

Чертане на полилиния, съставена от отсечки и дъги

Първо начертайте отсечката.

1. Стартирайте командата PLINE, като използвате един от следните методи:

Windows: От плаващото меню Polyline на функционалния ред Draw изберете Polyline.

DOS и UNIX: От менюто Draw изберете Polyline.

2. Задайте началната точка на отсечката.

3. Задайте крайната точка на отсечката.

4. Въведете **a**, за да превключите на режим Arc**.** Появява се диалоговият прозорец на режима Arc. Изберете OK.

5. Задайте крайната точка на дъгата.

6. Задайте трета точка на дъгата.

7. Въведете **l**, за да се върнете в режим Line. Появява се диалоговият прозорец на режима Line. Изберете OK.

8. Въведете дължината на отсечка от крайната точка на дъгата.

9. Въведете ъгъла на отсечката спрямо допирателната в крайната точка на дъгата.

10. Натиснете Return, за да завършите полилинията.

Вариант 5: Безличен-изявителен текст с обяснения и агрегация

Чертане на полилиния, съставена от отсечки и дъги

Първо се чертае отсечката.

1. Стартира се командата PLINE, като се използва един от следните методи:

Windows: От плаващото меню Polyline на функционалния ред Draw се избира Polyline.

DOS и UNIX: От менюто Draw се избира Polyline.

2. Задава се началната точка на отсечката и се задава крайна точка на отсечката.

3. Въвежда се **a** за превключване на режим Arc. Появява се диалоговият прозорец на режима Arc. Избира се OK.

4. Задава се крайната точка на дъгата и се задава трета точка на дъгата.

5. Въвежда се **l** за връщане в режим Line. Появява се диалоговият прозорец на режима Line. Избира се OK.

6. Въвежда се дължината на отсечка от крайната точка на дъгата и се въвежда ъгъла на отсечката спрямо допирателната в крайната точка на дъгата.

7. Натиска се Return за завършване на полилинията.

С изложеното съдържание в Глава 3 от дисертацията се проследява логиката на реално организиран процес за генериране на текст на български език. Създаденият ресурс за генериране на български изречения, Приложна българска системично-функционална граматика, е използван при тактическата генерация. Обърнато е внимание на семантичните технологии, които имат отношение към реалното конструиране и изпълнение на такъв процес за текст-генериране. Показан е нужния семантичен модел на предметната област на генерирането, показано е как се осигурява съответствие между понятията от този модел и висшата онтология кореспондираща с лингвистичните конструкции за лингвистична реализация на зададените понятия. Споделен е изследователският опит при текст-планирането извършено в рамките на проекта AGILE, за да се генерират 5 различни типа текстове от едно и също входно семантично съдържание. Демонстрирани са постигнатите резултати при генерирането на кохерентен текст, които са уникални като опит за автоматично генериране на български език.

4. Работа с български текстове, семантични технологии и ресурси

В тази глава се разглеждат приложения на семантичните технологии за семантично анотиране на специализирани текстове на български език, които биха подпомогнали процес на компютърно генериране на български текстове. Използвани са разработки по проекта СИНУС, който има за цел създаването на семантична платформа за технологично-поддържано обучение чрез динамично композиране на учебните материали. Това изисква да са на разположение поддържащи информационни модели за динамично създаване и адаптиране на учебни обекти, за да се осигури многократното им използване в процеса на обучение.

Семантичната платформа на проекта СИНУС е тествана с демонстрационни примери за използване на предлаганите от проекта технологи, които включват работа с мултимедийни учебни обекти анотирани с мета-данни за съдържанието им. Разглежданата тук научно-приложна задача се състои в това да се намерят методи за преодоляване на ограниченията на

стандартните мета-данни за учебни обекти, а именно, да се използват онтологии, които позволяват пряка компютърна обработка на знанията кодирани в мета-данните.

Чрез платформата на проекта СИНУС се достъпват и използват различни съществуващи мултимедийни библиотеки като в разглежданите тук приложения мултимедийните обекти са от мултимедийната цифрова библиотека "Виртуална енциклопедия на Източно-Християнското изкуство" [Pavlova-Draganova et. al., 2007]. Това е мултимедийната библиотека, която съдържа мултимедийна информация за иконографски обекти (икони, миниатюри, стенописи и т.н.) създадени на територията на България от VII до XIX век. Един мултимедиен обект е съвкупност от дигитални изображения и различни описателни текстове.

Семантичното търсене от платформата на проекта СИНУС в самата мултимедийна библиотека е осигурено от онтология, която за целите на проекта е наречена *базова онтология*. Възприетият подход осигурява намиране, визуализация и използване на мултимедийно съдържание чрез прилагане на различни схеми на мета-данните описващи мултимедийните обекти от различни перспективи според различните интереси на потребителите. Това се постига чрез формализирани експертни знания структурирани в онтологии, които са условно наречени *специализирани онтологии*.

4.1 Формализиране на знания от областта на иконографията

Задачата свързана с формализирането на знания от областта на иконографията е породена от задачата за създаване на семантични модели на основните мултимедийни обекти нужни за обучителната среда на проекта СИНУС. Основните мултимедийни обекти използвани в демонстрационните примери на проекта са ИКОНА, СТЕНОПИС, МИНИАТЮРА.

Достъпването на такива обекти от платформата на проекта СИНУС се осигурява от създадените в средата семантични модели на мултимедийните обекти. Семантичните модели от своя страна се изграждат върху онтологични конструкции. Спецификата на проекта, както и предварителният анализ на данните доведоха до решение онтологичните концептуални знания за мултимедийните обекти да се групират в една *базова онтология* и три *специализирани онтологии*.

През последните години в областта на онтологичното инжинерство е извършена значителна по обем работа, в резултат на която се увеличават областите на информационното пространство снабдени с онтологични стандарти при Интернет- обработки на информация. Такива области са медицината, културното наследство на човечеството, управлението на проекти и т.н.

За приложните изкуства и по-общо в областта на културното наследство на човечеството такъв фундаментален онтологичен модел е CIDOC – CRM, разработен от Групата по стандартизиране на документацията към Международния съвет на музеите (International Council of Museums - ICOM). От септември 2006-та година онтологията CIDOC CRM е приета за стандарт ISO 21127 на Международната организация по стандартизация (ISO).

Основната роля на CIDOC CRM е да служи като база за свързване на информацията за културното наследство и да представлява семантично „лепило" необходимо при трансформирането на съвременните разпръснати локални информационни източници в кохерентен и стойностен глобален ресурс."

Затова понятията от базовата онтология на проекта СИНУС- ОБИО са създадени да бъдат съвместими с понятията дефинирани в CIDOC CRM. Това е нужно при достъпване на хетерогенни данни от различни източници в Интернет-пространството. Така използваният в средата СИНУС семантичен модел е пре-използваем и достъпен за всеки проект базиран на CIDOC CRM, като публикуваните знания могат да се обработват от субекти в широкото Интернет-пространство.

След анализ на понятията от онтологията ОБИО-СИНУС, с добавените специализации на класове за ОБИО-СИНУС, йерархията на класовете от CIDOC CRM изглежда така:

E1 CRM Entity
E2 - Temporal Entity
E4 - - Period
E5 - - - Event
OBIO - - Important Event for ECR
E7 - - - - Activity
E11 - - - - - Modification
E12 - - - - - - Production
OBIO - - - - - - Iconographical Object Production
E13 - - - - - Attribute Assignment
E65 - - - - - Creation
E63 - - - - Beginning of Existence
E12 - - - - - Production
OBIO - - - - - Iconographical Object Production
E65 - - - - - Creation
E64 - - - - End of Existence
E77 - Persistent Item
E70 - - Thing
E72 - - - Legal Object
E18 - - - - Physical Thing
E24 - - - - - Physical Man-Made Thing
E90 - - - - Symbolic Object
E71 - - - Man-Made Thing
E24 - - - - Physical Man-Made Thing
OBIO - - - - Base of Iconographical Object
E22 - - - - - Man-Made Object
OBIO - - - - - Iconographical Object
OBIO - - - - - - Icon
OBIO - - - - - - Wall-Painting
OBIO - - - - - - Miniature
OBIO - - - - - - Mosaic
OBIO - - - - - - Витраж
OBIO - - - - - - Plastic Iconographical Object
OBIO - - - - - - - Иконостас
OBIO - - - - - - - Престол
E84 - - - - - Information Carrier
E25 - - - - - Man-Made Feature
E78 - - - - - Collection
E28 - - - - Conceptual Object
E90 - - - - - Symbolic Object
E73 - - - - - - Information Object
E33 - - - - - - - Linguistic Object
E35 - - - - - - - - Title
E36 - - - - - - - Visual Item
E38 - - - - - - - - Image
OBIO - - - - - - - - Iconographical Image
OBIO - - - - - - - - - One Figure Composition
OBIO - - - - - - - - - Many-figures Composition
OBIO - - - - - - - - - Composition of Complete Compositions
E41 - - - - - - Appellation
E42 - - - - - - - Identifier
E35 - - - - - - - Title
E89 - - - - - Propositional Object
E73 - - - - - - Information Object
E33 - - - - - - - Linguistic Object
E35 - - - - - - - - Title
OBIO - - - - - - - Biography
OBIO - - - - - - - IO Identification Note
OBIO - - - - - - - IO Description
OBIO - - - - - - - Iconographical Technique Description
OBIO - - - - - - - Base Description
OBIO - - - - - - - IO Condition State

E36 - - - - - - - Visual Item
E38 - - - - - - - - Image
OBIO - - - - - - - - Iconographical Image
OBIO - - - - - - - - - One Figure Composition
OBIO - - - - - - - - - Many-figures Composition
OBIO - - - - - - - - - Composition of Complete Compositions
E55 - - - - - Type
E56 - - - - - - Language
E57 - - - - - - Material
E58 - - - - - - Measurement Unit
OBIO - - - - - Iconographic School
OBIO - - - - - Iconographical Technique
OBIO - - - - - Canonic Type of Iconographical Image
E39 - - Actor
E74 - - - Group
OBIO - - - Iconographic Clan
E21 - - - Person
OBIO - - - Iconographer
OBIO - - - Important Person for ECR
E52 - Time-Span
OBIO - Year
OBIO - Month
OBIO - Day
OBIO - Century
OBIO - Part of Century
E53 - Place
OBIO - State
OBIO - Region
OBIO - Town
OBIO - Village
OBIO - Monastery
OBIO - Church
OBIO - Chapel
OBIO - Museum
OBIO - Gallery
E54 - Dimension
OBIO - Height
OBIO - Width
OBIO - Thickness
E59 Primitive Value

Приложение 3 към дисертацията съдържа представянето на онтологията OBIO на онтологичния език OWL, както и конструкциите на три специализирани онтологии формирани с участието на експерти в областта на иконографията [Paneva-Marinova et.al., 2010]. Онтологиите са във формата използван на практика в сценария за експлоатация на средата Синус.

4.2 Семантично анотиране на специализирани български текстове

В този раздел на дисертацията е описано решението на задача за семантично анотиране на специализирани текстове на български език. В приложния контекст на проекта Синус задачата се дефинира така:

Разполагаме с мултимедийни обекти, които са анотирани в семантичната платформа Синус и представляват екземпляри на класа *Иконографски обект* от базовата онтология ОБИО. Разполагаме с базов семантичен модел и разширен семантичен модел на *Иконографски обект*, поддържани от описаните в предходния раздел онтологии.
В Базовия семантичен модел връзките между понятията представляват или обектна релация или релация за данни. Обектната релация свързва две онтологични понятия. Релациите за данни предоставят достъп до текстови данни, които представляват кратки описателни

текстове на български език за конкретния иконографски обект – икона, стенопис, миниатюра. Текстовете са част от описанията на мултимедийни обекти от цифровата библиотека „Виртуална енциклопедия на българската иконография [Pavlova-Draganova, et.al., 2007].

Задачата за семантично анотиране е насочена именно към тези текстове. Тя се състои в това в описателните текстове на български език да бъдат добавени анотации, които фиксират всички споменавания на онтологични понятия от специализираната онтология „Технология на иконографски обект“.

Тези семантични анотации след това позволяват полуавтоматично разширяване на семантичния модел на мултимедийния обект в семантичното пространство на средата Синус, където семантичното анотиране е част от по-голяма прагматична задача. Чрез разширения семантичен модел става възможно семантично търсене на анотирани мултимедийни обекти в платформа за технологично поддържано обучение [Agre, 2012], [Dochev and Agre, 2012].

Семантично анотиране на български текстове чрез следване на релацията “От онтология към текст”

Подходът за семантично анотиране чрез следване на релацията “От онтология към текст” е представен в обзорната част на дисертацията. Използването му в конкретната задача за анотиране дефинирана в проекта Синус е изключително подходящо, защото са изпълнени всички изисквания за прилагане на този подход:

1) в семантичното пространство на Синус-платформата се използва ***онтологията*** „Технология на иконографски обект”;

2) онтологичните й понятия са лексикаризирани на български език и на тази основа е възможно създаването на ***терминологичен лексикон*** на български език**;**

3) могат да бъдат построени ***частични граматики*** на български език, които да разпознават споменавания на онтологичните понятия в българските текстове и да ги анотират като онтологични термини.

Настройване на средствата за семантично анотиране

Прилагането на метода “От онтология към текст” започва със съставяне на онтологичен лексикон. В ***терминологичния лексикон*** се записват лексикализациите на наименованията на онтологичните понятия в дефинициите на онтологията. Например, в онтологията е дефиниран клас с уникален идентификатор *#OWLClass_Lacquering* и лексикализации за наименованието на класа на български език: *лак* и *лаково покритие*.

```
<owl:Class rdf:about="#OWLClass_Lacquering">
    <rdfs:label xml:lang="bg">лак</rdfs:label>
    <rdfs:label xml:lang="bg">лаково покритие</rdfs:label>
    ...
</owl:Class>
```

Лексиконът има два дяла, като единият съдържа наименования на онтологичните индивиди, другият - наименованията на онтологичните класове. Към тези думи и словосъчетания са добавени някои вариации.

Анотационни граматики

За реализиране на анотационните граматики е използвана системата CLaRK [Simov et.al., 2001] предназначена за създаване и работа с корпуси от XML-документи. Името на системата идва от съкращение на Computational Linguistics and Represented Knowledge. Системата CLaRK разполага с инструмент за работа с крайни автомати, който се използва за проверка на валидността на XML-документите, в токанизаторите и в каскадните регулярни граматики.

Граматиката работи детерминистично над входната дума. Резултатът от прилагането й е копие на входната дума, в което разпознатите под-думи са заменени с категории на

граматиката. Резултатът е наречен изходна дума за граматиката. В този смисъл този вид регулярни граматики могат да бъдат наречени крайни трансдюсери [Simov et al., 2002].

За настройване на системата ClaRK е необходим корпус от текстове, за да бъде разширен лексикона и да бъде „обучена" анотационната граматика в разпознаване на срещанията на лексическите елементи в целевите текстове. В общия случай е важно текстовете от корпуса действително да съдържат лексикализации на онтологичните понятия, които ще бъдат разпознавани.

В работата по настройване на анотационните граматики на CLaRK са използвани част от наличните описателни текстове на български език достъпни чрез свойствата за данни на базовия семантичен модел:

1)#OWLDataProperty_base_has_Description с описания на основата на иконографски обект,

2)#OWLDataProperty_iconographicalTechnique_has_Description с описания на иконографската техника характерна за иконографския обект,

3)#OWLDataProperty_conditionState_has_Name с описание на текущото състояние на иконографски обект.

Корпусът се състои от 434 описателни текста, разпределени според съдържанието си както следва:

Описания на основата на иконографски обект	189 текста
Описания на иконографска техника	124 текста
Описания на състоянието на иконографски обект	121 текста

За съставянето на златен стандарт е извършено ръчно маркиране на срещанията на онтологичните термини в текстовете. Златният стандарт задава целта, към която се стремим с настройването на програмата: маркираните срещания на онтологични елементи в текстовете от корпуса да се разпознаят автоматично от анотационната граматика.

В корпуса са отбелязани 899 срещания на търсените термини, т.е. общ брой лексикализации за индивиди и класове от онтологията „Технология на иконографски обект".

Създаване на анотационни граматики

Както е споменато по-горе, граматиките в системата CLaRK се състоят от правила представени чрез регулярен израз и маркер. За решаване на конкретната задача са създадени две граматики - едната разпознава индивидите от онтологията, а другата – онтологичните класове. Изграждането на всяка от граматиките започва с елементите на съответния лексикон. Всеки елемент на лексикона се лематизира с „Българския морфологичен лексикон" [Попов, Симов, Видинска, 1998] и лематизираната форма се превръща в регулярен израз на съответното граматическо правило.

Съществената част от настройването или „обучението" на граматиката се състои в лематизирането на регулярните изрази, които трябва да разпознаят търсените фрази в текстовете. Най-простият случай е обобщаване или заместване на част от фразата с подходящ системен символ.

Създаване на каскадно задание за претърсване на текст

Върху входния XML-документ съдържащ целеви текстове се извършват следните обработки формирани на етапа на обучението на системата:

- токанизация на текстовете
- вмъкване на XML-елемент за изходния списък с маркери
- парсиране с граматиката SINUS TSO basic
- парсиране с граматиката SINUS TSO I-2
- парсиране с граматиката SINUS TSO Cl

Описаните обработки образуват едно каскадно задание в системата CLaRK, чиито изход е XML-документ, с добавени към него списъци с маркери, получени чрез разпознаването на заложените в граматиките фрази (регулярни изрази). Обученото състояние на системата

CLaRK се запомня, за да бъде извиквано от процесите на средата СИНУС.

Резултати от работата върху текстовия корпус

Програмата за разпознаване на термините с така настроените граматики е приложена върху всеки текст от корпуса. Резултатите, изложени в приложение към дисертацията, показват за всеки онтологичен термин до каква степен е било ефективно "обучението" на граматиките. От обобщението е видно, че програмата разпознава правилно общо 757 срещания на термини и 12 пъти погрешно предлага разпознаване на термин. Неразпознати са 142 срещания на термини в текстовете от корпуса. С тези данни са пресметнати стойностите на Точност (Precision) и Покритие (Recall):

Точност = 757 / 769 = 0.984

Покритие = 757 / 899 = 0.842

За поставените прагматични цели определени в проекта Синус тези резултати са много добри. Приложение 4 съдържа таблица с информация за неразпознатите срещания на онтологичните термини в корпуса.

Проведеното научноприложно изследване показва, че макар да са възможни по-нататъшни подобрения, следването на релацията „От онтология към текст" води до ефективно решаване на поставената задача за семантично анотиране на български текстове.

Реализираната процедура разпознава най-често срещаните в корпуса онтологични термини и създава адекватни семантични анотации. Като основна насока за бъдеща работа по развитие на избрания метод можем да отбележим, че съществува широко поле за експерименти с различни стратегии за лематизиране на регулярните изрази на частичните анотационни граматики. Съществен принос би дало пре-използването на създадените с метода ресурси във вид на терминологични лексикони, особено ако могат да са полезни и в други приложни области.

5. Заключение

Настоящият дисертационен труд съдържа описание на научно-приложната и изследователска работа по подготовката и използването на ресурси за компютърно генериране на български текстове. Теоретичната парадигма за компютърното генериране е Системично-функционалната лингвистика на Халидей. Избраният приложен подход към генерирането е така наречения контрол направляван от граматиката. Както е показано в обзорната част, необходимите ресурси в конкретния формализъм за компютърно генериране на даден естествен език са лексико-граматически и семантични.

Показано е изграждането на приложна лексико-граматика за компютърно генериране на български език в последователните стъпки на процеса:

1) Частичен анализ и моделиране на някои езикови явления в българския език. Създадени са модели на езиковите явления, които са идентифицирани в корпус от целеви текстове за генерирането. Текстовете са процедури-инструкции от ръководство за ползване на CAD/CAM системи. Системично-функционалната граматика на Халидей е парадигмата на лингвистичното моделиране. В дисертацията са показани съставените модели за езиковите явления диатезност, модалност, темпоралност, завършеност на процеса, номинализация на глаголната група, членуване. Моделирани са и някои типични за корпуса връзки в сложните изречения.
2) На основата на лингвистичните модели и в паралел със съществуващ приложен ресурс за английски език е съставена приложна българска системично-функционална граматика. Приложният ресурс е изграден със средствата на Средата за многоезиково генериране KPML. Съществени характеристики на приложния ресурс са достъпност, отворената възможност за разширяване и надграждане, както и за пре-използване в други предметни области.

В дисертацията е описан процесът на генериране на български текстове със създадения лексико-граматически ресурс. Докладвана е работата с нужните семантични ресурси за генерирането: проверки за настройване на лингвистичната онтология от високо ниво Обобщен модел (Generalized Upper Model) и на модела на предметната област за работа с български език.

Показан е резултат от компютърно генериране на български език: 5 варианта на текстове-инструкции генерирани от едно и също представяне на информацията на входа.

Последната част на дисертацията показва работата по формализиране на знания в семантични модели свързани със семантично анотиране на български текстове. С идея, че семантичното анотиране би подпомогнало значително подготовката за компютърно генериране в нови предметни области е описан процес на семантично анотиране на описателни текстове на български език от областта на иконографията. Приложен е подходът „Релация: От онтология към текст“. Използвани са частични граматики на базата на регулярни изрази разработени в системата CLaRK.

Дисертацията инспирира идеи за бъдеща работа. Възможностите за експерименти и изследвания на компютърно генериране на български текстове с приложната системично-функционална граматика са огромни. Интересни са и перспективи като многоезиковото генериране или компютърно генериране от различни входни представяния на информацията.

Основни научни и научно-приложни приноси

В дисертацията са докладвани резултати от осъществяване на процес на компютърно генериране на български текстове. Реализирани са следните научни и научно-приложни приноси:

1. Направено е формално описание на базови обекти от българския език в съответствие с Системично-функционалната теория за естествения език. Разработеното описание позволява компютърно генериране на процедурни текстовете (технически инструкции) за създаване на техническа документация на български език.

2. Разработеното формално описание е реализирано във вид на приложен компютърен ресурс (Приложна системично-функционална граматика на български език). Реаризацията е извършена в Средата за многоезиково генериране KPML и позволява пре-използване, разширяване и по-нататъшно развитие.

3. Разработената Приложна системично-функционална граматика на български език е реализирана като модул в многоезиковата система AGILE позволяваща компютърно генериране на технически текстове в различни стилове (персонален- императивен, безличен-изявителен). Системата е приложена за генериране на свързани текстове на български език при създаване на технически ръководства за CAD/CAM системи.

4. Въз основа на разработено онтологично описание (специализация на CIDOC- CRM) в областта на Източно-християнско иконографско изкуство е предложена и реализирана схема за семантично анотиране на специализирани текстове на български език. Реализацията използва частични граматики на базата на регулярни изрази.

Списък на публикациите по дисертацията

1. Kamenka Staykova: *Natural Language Generation and Semantic Technologies*, Cybernetics and Information Technologies, Volume 14, No2, 2014, pp. 3-24.

2. Kamenka Staykova, Gennady Agre: *Use of Ontology-to-Text Relation for Creating Semantic Annotation*, In Proceedings of 13th International Conference on Computer Systems and Technologies - CompSysTech 2012, Ruse, Bulgaria, June 22 - 23, 2012, pp. 64-71.

3. Kamenka Staykova, Petya Osenova, Kiril Simov: *New Applications of "Ontology-to-Text Relation" Strategy for Bulgarian Language*, Cybernetics and Information Technologies, Bulgarian Academy of Sciences, Sofia, Vol.12, No 4, 2012, pp. 43-52.

4. Kamenka Staykova, Gennady Agre, Kiril Simov, Petya Osenova: *Language Technology Support for Semantic Annotation of Iconographic Descriptions*, In Proceedings of the International Workshop "Language Technologies for Digital Humanities and Cultural Heritage", Sept. 2011, Hisar, Bulgaria, 16 Sept. 2011, pp. 51-57.

5. Kamenka Staykova: *Exercise in Conceptualization*, Cybernetics and Information Technologies, Bulgarian Academy of Sciences, Sofia, Vol. 5, No 2, 2005, pp. 69-83.

6. Kamenka Staykova and Sergey Varbanov: *The Globe: Representation of Linguistic Knowledge and Knowledge about the World Together*, In Proceedings of the Workshop "Language and Speech Infrastructure for Information Access in the Balkan Countries", part of Fifth International Conference on Recent Advances in Natural Language Processing, RANLP-2005, Borovec, Bulgaria, 25 September 2005, pp. 68-74.

7. Danail Dochev, Kamenka Staykova: *A Multilingual System for Automatic Generation of Technical Manual Texts*, In Proceedings of the International Conference on Computer Systems and Technologies CompSysTech'2001, Sofia, 21-22 June 2001, pp. II.14.1-5.

8. Kamenka Staykova, Danail Dochev: *Development of Lexico-Grammar Resources for Natural Language Generation (Experience from AGILE Project)*, In: Cerry S. and D. Dochev (Eds.), Proceedings of the International Conference "Artificial Intelligence: Methodology, Systems, Applications 2000", Varna, September 2000, Lecturer Notes in Artificial Intelligence 1904, Springer-Verlag, 2000, pp. 242-251.

9. Kruijff GJ, E. Teich, J. Bateman, I. Kruijff-Korbayová, H. Skoumalová, S. Sharoff, L. Sokolova, T. Hartley, K. Staykova, J. Hana: *Multilinguality in a Text Generation System for Three Slavic Languages*, In Proceedings of the 18th Conference on Computational Linguistics - Volume 1, Saarbrücken, Germany, 2000, pp. 474 – 480..

10. Staykova K.: *Bulgarian Resource for Generation of Instructional Texts: Result of AGILE Project*, IIT Working Papers, IIT/WP-109, 2000.

11. Стайкова К., Й. Пенчев: *Системично- функционалната лингвистика и българският език*, "Български език", София, XLVIII, том 4-5, 1999/2000, стр. 5-24.

12. Dochev D., N. Gromova, K. Staykova - Lexico-Grammatical Characteristics of Bulgarian Software Instructional Texts. Problems of Engineering Cybernetics and Robotics, No 49, pp. 11-19, 1999.

Публикация 1. кореспондира с обзорната част на дисертацията, Глава 1, където се разглежда отношението на компютърното генериране на текст и съвременните семантични технологии. Публикации 10. и 11. са свързани с моделирането на някои езикови явления в българския език (раздел 2.1), докато публикации 7. и 9. отразяват работата по приложната граматика (раздел 2.2). Към изследванията описани в Глава 3 имат отношение публикации 5., 6. и 8. Това са доклади за реализирания процес на компютърно генериране на български текстове. Публикации 2., 3. и 4. се отнасят към Глава 4 и отразяват работата по семантично анотиране на специализирани български текстове.

Апробация на резултатите

Доклади по темата на дисертацията са изнесени на следните научни форуми и семинари:

- Международна конференция по компютърни системи и технологии CompSysTech'2012, Русе юни, 2012.

- Международен семинар “Language Technologies for Digital Humanities and Cultural Heritage”/ „Езикови технологии за електронната хуманитаристика и културното наследство“ Хисаря, септември, 2011.

- Международен семинар "Language and Speech Infrastructure for Information Access in the Balkan Countries", Езикова инфраструктура за достъп до информацията в Балканските страни, Боровец, септември 2005г.

- постер-сесия на международната конференция RANLP 2001, Цигов чарк, 2001.

- Международна конференция по компютърни системи и технологии CompSysTech'2001, София, юни 2001г.

- 18та международна конференция по компютърна лингвистика, Саарбрюкен, Германия, 2000.

- Девета международна конференция AIMSA: Изкуствен интелект- методология, системи, приложения, Варна, септември 2000г.

- Текущи семинари на секция “Изкуствен интелект” в Института по информационни технологии на БАН, 1999-2004.

Литература

[Бояджиев, Куцаров, Пенчев, 1999] Бояджиев Т, И. Куцаров, Й. Пенчев: *Съвременен български език*, ИК Петър Берон, София, 1999.

[Попов, Симов, Видинска, 1998] Попов Д., К. Симов и Св. Видинска: *Речник за правоговор, правопис и пунктуация*, Атлантис, 1998, София.

[Стайкова и Пенчев, 2000] Стайкова К., Й. Пенчев: *Системично- функционалната лингвистика и българският език*, "Български език", София, XLVIII, том 4-5, 1999/2000, стр. 5-24.

[Agre, 2012] Agre, G.: *SINUS – A Semantic Technology Enhanced Environment for Learning in Humanities*, Cybernetics and Information Technologies, Vol. 12, 2012, No 4, 5-24.

[Al-Muhtaseb and Mellish, 1997] Al-Muhtaseb, Husni and Chris Mellish: From the Generalized Upper Model Towards an Arabic Upper Model, In Proceedings of The 4th IEEE International Conference on Electronics, Circuits and Systems ICECS'97, Cairo, Egypt, 1997.

[Basile and Bos, 2011] Basile V. and J. Bos: Towards *Generating Text from Discourse Representation Structures*, In Proceedings of the 13th European Workshop on Natural Language Generation (Nancy, France, 2011), pp. 145–150.

[Bateman, 1990] Bateman J.: *Upper Modelling: Organizing Knowledge for Natural Language Processing*, In Proceedings of the 5th International Workshop on Natural Language Generation, 3-6 June 1990, pp. 54–60.

[Bateman, 1997] Bateman J. A.: *Enabling Technology for Multilingual Natural Language Generation: The KPML Development Environment*,. Natural Language Engineering 3, 1 (1997), 15–55.

[Bateman, 2002] Bateman J. A.: *Natural Language Generation: an Introduction and Open-Ended Review of the State of the Art, 2002*, http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/nlg/ATG01/ATG01.html .

[Bateman and Teich, 1995] Bateman, J. A. and Teich, E.: *Selective Information Presentation in an Integrated Publication System: an Application of Genre-Driven Text Generation*, Information Processing and Management, 1995, 31(5), 753-767.

[Bateman and Paris, 1989] Bateman, J.A. and C.L. Paris: *Phrasing a Text in Terms the User Can Understand*, In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, IJCAI'89, Detroit, Michigan, 1989, pp. 1511-1517.

[Bateman et.al, 1990] Bateman J., R. Kasper, J. Moore, R. Whitney: *A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model*, Technical report, USC/Information Sciences Institute, Marina del Rey, California, 1990.

[Bateman et.al., 1994] Bateman, J., B. Magnini and F. Rinaldi: *The Generalized {Italian, German, English} Upper Model*, The ECAI94 Workshop: Comparision of Implemented Ontologies, Amsterdam, 1994.

[Bateman et.al., 1995] Bateman J. A., B. Magnini and G. Fabris: *The Generalized Upper Model Knowledge Base: Organization and Use*, In Towards Very Large Knowledge Bases, N. Mars, Ed. IOS Press, Amsterdam, 1995, pp. 60–72.

[Bateman et.al., 2000] Bateman, J., Teich, E., Kruijff-Korbayová, I., Kruijff, G.-J., Sharoff, S. and Skoumalová, H.: *Resources for multilingual text generation in three Slavic languages*, in Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000), European Language Resources Association (ELRA), 2000, Athens, Greece, pp.1763-1768.

[Bateman et.al., 2010] Bateman, J. A., J. Hois, R. Ross, and T. Tenbrink: *A Linguistic Ontology of Space for Natural Language Processing*, Artificial Intelligence 174, 14 (2010), 1027 – 1071.

[Berners-Lee et.al., 2001] Berners-Lee T., J. Hendler and O. Lassila: *The Semantic Web*, Scientific American, 2001, pp. 29–37.

[Bontcheva and Angelova, 1996] Bontcheva, K. and Angelova, G.: *Planning and Generating Hypertext Documentation.* In: Proceedings. of the Workshop "Gaps and Bridges in Natural Language Genaration", European Conference on Artificial Intelligence ECAI-96, Budapest, Hungary, August 1996, pp. 25-28.

[Bontcheva and Cunningham, 2003] Bontcheva, D. K. and H. Cunningham: *The Semantic Web: A New Opportunity and Challenge for Human Language Technology,* In Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services, ISWC'03 ,Florida, USA, October 2003, pp. 89–96.

[Bontcheva, 2005] Bontcheva K.: *Generating Tailored Textual Summaries from Ontologies*. In The Semantic Web: Research and Applications, Proceedings of the Second European Semantic Web Conference, ESWC, Heraklion, Crete, Greece, 2005, pp. 531–545.

[Bontcheva and Wilks, 2004] Bontcheva K. and Wilks Y.: *Automatic Report Generation from Ontologies: the MIAKT Approach*. In Natural Language Processing and Information Systems, Proceedings of the Nineth International Conference on Applications of Natural Language to Information Systems (NLDB), 2004, vol. 3136 of Lecture Notes in Computer Science, Springer, pp. 324–335.

[Bouayad-Agha et.al., 2012] Bouayad-Agha N., Casamayor G., Mellish C., and Wanner L.: Content Selection from Semantic Web Data. In Proceedings of the Seventh International Natural Language Generation Conference (INLG), Special Track on Future Generation Challenges Proposals (2012), pp. 146–149.

[Cimiano and Kopp, 2010] Cimiano P.and S. Kopp: *Accessing the Web of Data through Embodied Virtual Characters*, Semantic Web 1, 1-2, 2010, 83–88.

[Crofts et.al., 2011] Crofts, N., Doerr M., Gill T., Stead S., Stiff M.: *Definition of the CIDOC Conceptual Reference Model*, 2011, available: http://www.cidoccrm.org/docs/cidoc_crm_version_5.0.4.pdf

[Dai et.al., 2010] Dai Y., S. Zhang, J. Chen, T. Chen, W. Zhang: *Semantic Network Language Generation Based on a Semantic Networks Serialization Grammar*, World Wide Web 13, 3, 307-341.

[Dannells et.al., 2012] Dannells D., M. Damova, R. Enache, and M. Chechev: *Multilingual Online Generation from Semantic Web Ontologies*. In Proceedings of the 21st International Conference Companion on World Wide Web, New York, USA, 2012, WWW '12 Companion, ACM, pp. 239–242.

[Davis et.al., 2008] Davis B., A. Iqbal, A. Funk, V. Tablan, K. Bontcheva, H. Cunningham, S. Handschuh: *RoundTrip Ontology Authoring*. In Proceedings of the International Semantic Web Conference (ISWC), vol. 5318 of Lecture Notes in Computer Science. Springer Berlin /Heidelberg, 2008, pp. 50–65.

[Davis et.al., 1993] Davis R., Shrobe H., and Szolovits P.: *What Is a Knowledge Representation?*, AI Magazine Volume 14 Number 1, 1993, http://www.aaai.org/ojs/index.php/aimagazine/article/view/1029/947

[Demir et.al, 2010] Demir S., S. Carberry, and K. McCoy: *A Discourseaware Graph-based Content Selection Framework,* In Proceedings of the 6th International Natural Language Generation Conference (INLG) (2010), pp. 17–27.

[Dochev and Agre, 2009] Dochev D., Agre G.: *Towards Semantic Web Enhanced Learning*, In: Proceedings of the International Conference on Knowledge Management and Information Sharing, Madeira, 2009, pp. 212-217.

[Dochev and Agre, 2012] Dochev D. and G. Agre: *Supporting Learning-by-Doing Situations by Semantic Technologies***,** In: Stefanoui D., J Culita (Eds.). Proceedings of 17th Annual Conference on Media and Web Technology EUROMEDIA'2012, Bucharest, April, 2012, ISBN 978-90-77381-69-4, pp. 49-53.

[Duboue and McKeown, 2003] P. A. Duboue and K. R. McKeown: *Statistical acquisition of content selection rules for natural language generation*. In Proceedings of 2003 Conference on Empirical Methods for Natural Language Processing, (EMNLP 2003), Sapporo, Japan, July.

[Elhadad, 1990] Elhadad, M.: *Types in Functional Unification Grammars*, in Proceedings of the 28th. Annual Meeting of the Association for Computational Linguistics, ACL, 1990, pp. 157-164.

[Elhadad and Robin, 1992] Elhadad, M. and Robin, J.: *Controlling Content Realization with Functional Unification Grammars*, In R. Dale, E. H. Hovy, D. Rösner and O. Stock, editors, "Aspects of automated natural language generation", 6th. international workshop on natural language generation, Springer, Berlin/Heidelberg, 1992, pp. 89-104.

[Erdmann et.al, 2000] Erdman M., A. Maedche, H.-P. Schnurr, S. Staab: *From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools*, In P. Buitelaar and K. Hasida (eds) Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, 2000.

[Galanis and Androutsopolos, 2007] Galanis, D. and I. Androutsopoulos: *Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: the NaturalOWL System*, in Proceedings of the 11th European Workshop on Natural Language Generation, ENLG '07, pp. 143–146, 2007.

[Gardent et.al., 2011] Gardent C., E. Banik, and L. Perez-Beltrachini: *Natural Language Generation and Natural Language Interfaces to Knowledge Bases*, Tutorial at the Sixth International Conference on Knowledge Capture (K-CAP), 2011.

[Gruber, 1993] Gruber, T.*: A translation approach to portable ontology specifications,* Knowledge Acquisition 5(2), pp. 199–220, 1993.

[Halliday, 1994] M A K Halliday: *Introduction to Functional Grammar*, Edward Arnold, London, Second Edition, 1994.

[Henschel, 1993] Renata Henschel: *Merging the English and the German Upper Model*, Darmstadt, Germany, GMD/ Institute fur Integriente Publikation-and Informationssysteme, 1993.

[Hewlett et.al., 2005] Hewlett D., A. Kalyanpur, V. Kolovski, C. Halaschek-Wiener: *Effective NL Paraphrasing of Ontologies on the Semantic Web*. In Workshop on End-User Semantic Web Interaction, 4th Internatioanl Semantic Web Conference (ESWC), Galway, Ireland, 2005.

[Hielkema et.al., 2007] Hielkema F., C. Mellish, P. Edwards: *Using WYSIWYM to Create an Open-ended Interface for the Semantic Grid*, In Proceedings of the Eleventh European Workshop on Natural Language Generation, ENLG '07, Stroudsburg, PA, USA, Association for Computational Linguistics, 2007, pp. 69–72.

[Hovy, 1988] Hovy, E.: *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

[Hovy and Nirenburg, 1992] Eduard Hovy and Sergei Nirenburg: *Approximatingan Interlingua in a Principled Way*, the DARPA Speech and Natural Language Workshop, Arden House, New York, 1992.

[Joshi, 1987] Joshi, A. K.: *The Relevance of Tree Adjoining Grammar to Generation*, In G. Kempen, ed., "Natural Language Generation: Recent Advances in Artificial Intelligence, Psychology, and Linguistics", Kluwer Academic Publishers, Boston/Dordrecht, 1987.

[Kruijff et.al., 2000] Kruijff, G.-J., Teich, E., Bateman, J., Kruijff-Korbayová, I., Skoumalová, H., Sharoff, S., Sokolova, L., Hartley, T., Staykova, K. and Hana, J., *A multilingual system for text generation in three Slavic languages*, in Proceedings of the 18th. International Conference on Computational Linguistics (COLING'2000)', Saarbrücken, Germany, 2000, pp. 474-480.

[Lavoie and Rambow, 1997] Lavoie, B. and Rambow, O.: *A fast and portable realizer for text generation systems*, in Proceedings of the 5th. Conference on Applied Natural Language Processing, ACL, Washington, 1997, pp. 265-268.

[Mann, 1983] Mann, W.C., *An overview of the PENMAN text generation system*, in Proceedings of the National Conference on Artificial Intelligence, AAAI, 1983, pp.261-265.

[Mann and Matthiessen, 1985] Mann, W.C. and Matthiessen, C.M.I.M.: *Demonstration of the Nigel text generation computer program*, in J.D.Benson and W.S.Greaves, eds, "Systemic Perspectives on Discourse", Volume 1, Ablex, Norwood, New Jersey, 1985, pp.50-83.

[Mann and Thompson, 1988] Mann, W.C. and S.A. Thompson: *Rhetorical structure theory: Toward a functional theory of text organization*, Text 8(3), 243-281.

[Matthiessen and Bateman, 1991] Mattiessen C. and J. Bateman: *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*, Frances Pinter Publishers and St. Martin's Press, London and New York, 1991.

[McDonald, 1983] McDonald, D.D. *Description directed control: its implications for natural language generation*, Computers and Mathematics, 9(1), 1983, 111-129.

[McKeown, 1985] McKeown, K: *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*, Cambridge University Press, Cambridge, England, 1985.

[Mellish and Pan, 2006] Mellish C., and J. Pan: *Finding Subsumers for Natural Language Presentation.* In Proceedings of the International Workshop on Description Logics, The Lake District, England, 2006, pp. 127–134.

[Mellish and Pan, 2008] Mellish C., and J. Pan: *Natural Language Directed Inference from Ontologies.* Artificial Intelligence 172, 10, June 2008, 1285–1315.

[Mellish and Sun, 2006] Mellish C., and X. Sun: *The Semantic Web as a Linguistic Resource: Opportunities for Natural Language Generation.* Knowledge-Based Systems 19, 5, Sept. 2006, 298–303.

[Meteer, 1992] Meteer, M. W.: *Expressibility and the Problem of Efficient Text Planning*, Pinter Publishers, London, 1992.

[Mitkov, 1990] Mitkov, R.: *Generating Explanations of Geometrical Concepts,* Computers and Artificial Intelligence, 1990, 9(6), 579-589.

[Moore and Paris, 1988] Moore, J. D. and Paris, C. L.: *Constructing Coherent Texts Using Rhetorical Relations*, In Proceedings of the Tenth Annual Conference of the Cognitive Science Society', Cognitive Science Society, 1988.

[Nogier and Zock, 1992] Nogier J.-F. and M. Zock, *Lexical Choice as Pattern Matching*, Knowledge Based Systems 5, 3 (1992), 200–212.

[Paneva-Marinova et.al., 2010] Paneva-Marinova D., R. Pavlov, M. Goynov, L. Pavlova-Draganova, L. Draganov: *Search and Administrative Services in Iconographical Digital Library*, In Proceedigs of the International Conference "Information Research and Applications", July 2010, Varna, Bulgaria, 177-187.

[Pavlova-Draganova, et.al., 2007] Pavlova-Draganova L., V. Georgiev, L. Draganov: *Virtual Encyclopaedia of Bulgarian Iconography,* Information Technologies and Knowledge, Vol. 1, 2007, No 3, 267-271.

[Pollard and Sag, 1994] Carl J. Pollard and Ivan A. Sag: Head-Driven Phrase Structure Grammar, University of Chicago Press, Chicago, Illinois, USA, 1994.

[Polikoff and Allemang, 2003] I. Polikoff and D. Allemang: *Semantic Technology*, TopQuadrant Technology Briefing, v1.1, September 2003, https://lists.oasis-open.org/archives/regrep-semantic/200402/pdf00000.pdf .

[Power, 2009] Power, R.: *Towards a Generation-based Semantic Web Authoring Tool*, in Proceedings of the 12th European Workshop on Natural Language Generation ENLG '09, Stroudsburg, USA, ACL, 2009, pp. 9–15.

[Power, 2010] Power R.: *Complexity Assumptions in Ontology Verbalisation*, In Proceedings of 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Stroudsburg, PA, USA, 2010, ACLShort '10, pp. 132–136.

[Power and Third, 2010] Power R. and A. Third: *Expressing OWL Axioms by English Sentences: Dubious in Theory, Feasible in Practice*, In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Stroudsburg, PA, USA, 2010, COLING'10, Association for Computational Linguistics, pp. 1006–1013.

[Ranta, Angelov, Hallgren, 2010] Ranta, A., K. Angelov, T. Hallgren: *Tools for Multilingual Grammar-Based Translation on the Web,* In Proceedings of the ACL 2010 System Demonstrations, pp. 66-71, 2010.

[Rao et.al., 2011] Rao D., P. McNamee, M. Dredze: *Entity Linking: Finding Extracted Entities in a Knowledge Base*, Multi-source, Multi-lingual Information Extraction and Summarization, 2011.

[Reiter, 1995] Reiter, E., *NLG vs. Templates*, in Proceedings of the Fifth European Workshop on Natural Language Generation, Faculty of Social and Behavioural Sciences, University of Leiden, Leiden, The Netherlands, 1995, pp. 95-105.

[Reiter and Dale, 2000] Reiter E. and R. Dale: *Building Natural Language Generation Systems (Studies in Natural Language Processing)*, Cambridge University Press, 2000.

[Shieber et.al., 1990] Shieber, S.M., van Noord, G., Pereira, F. C. N. and Moore, R.C., *Semantic head-driven generation*, Computational Linguistics 16(1), 1990, 30-42.

[Simov and Osenova, 2007] Simov K., P. Osenova: *Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects*. In: Proceedings of the Workshop on NLP and Knowledge Represenattion for eLearning Environments, RANLP-2007, 49-55.

[Simov and Osenova, 2008] Simov K., P. Osenova: *Language Resources and Tools for Ontology-Based Semantic Annotation*, In: Al. Oltramari, L. Prévot, Chu-Ren Huang, P. Buitelaar, P. Vossen, Eds. Proc. of the OntoLex Workshop at LREC'2008, 2008, 9-13.

[Simov et.al., 2001] Simov K., Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, A. Kiryakov: *CLaRK - an XML-based System for Corpora Development*, In: Proceedings of the Corpus Linguistics Conference, 2001, pp. 558-560.

[Simov et.al., 2002] Kiril Simov, Milen Kouylekov, Alexander Simov, *Cascaded Regular Grammars over XML Documents*, In: Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002), Taipei, Taiwan. September, 2002. pages 51-58.

[Staykova, 2014] K. Staykova: *Natural Language Generation and Semantic Technologies*, Cybernetics and Information Technologies, Volume 14, No2, 2014, pp. 3-24.

[Staykova and Agre, 2012] K. Staykova, G. Agre: *Use of Ontology-to-Text Relation for Creating Semantic Annotation*, In Proceedings of 13th International Conference on Computer Systems and Technologies - CompSysTech 2012, Ruse, Bulgaria, June 22 - 23, 2012, pp. 64-71.

[Staykova et.al., 2012] K. Staykova, P. Osenova, K. Simov: *New Applications of "Ontology-to-Text Relation" Strategy for Bulgarian Language*, Cybernetics and Information Technologies, ISSN 1311-9702, Bulgarian Academy of Sciences, Sofia, Vol.4, 2012, pp. 43-52.

[Stevens et.al., 2011] Stevens R., J. Malone, S. Williams, R. Power, A. Third: *Automating Generation of Textual Class Definitions from OWL to English*. Journal of Biomedical Semantics 2, 2:S5, 2011.

[Teich, 1999] Elke Teich: *Systemic functional grammar in Natural Language Generation: linguistic description and computational representation*, Cassell, London, 1999.

[Turner et.al., 2006] Turner R., S. Sripada, E. Reiter, I. Davy: *Generating Spatio-Temporal Descriptions in Pollen Forecasts.* In Proceedings of the 11th Conference of European Chapter of the Association for Computational Linguistics, Trento, Italy, EACL06, 2006, pp. 163-166.

[Wimalasuriya and Dou, 2010] Wimalasuriya D. C. and D. Dou: *Ontology- based Information Extraction: An Introduction and a Survey of Current Approaches*, Journal of Information Science, 36 (3), 2010, pp. 306-323.