ОТ СИНТАКСИСА К СЕМАНТИКЕ – О ВЫБОРЕ ФОРМАЛИЗМОВ И ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

FROM SYNTAX TO SEMANTICS – CHOOSING FORMALISMS AND LANGUAGE RESOURCES

Койт М.Э*. (mare.koit@ut.ee),* ***Роосмаа Т.А****. (tiit.roosmaa@ut.ee),* ***Ыйм Х.Я.*** *(haldur.oim@ut.ee)*
Тартуский университет

Рассматриваются формализмы, методы и лингвистические ресурсы, применяемые в вычислительной лингвистике для моделирования синтаксиса и семантики. Дается обзор работ над автоматическим анализом синтаксиса и семантики эстонского языка, которые ведутся в Тартуском университете.

1. Введение

Автоматический анализ естественного языка (ЕЯ) важен для многих практических приложений: проверки правописания и грамматики текста, составления аннотации документа, машинного перевода (МП), поиска информации в Интернете, общения с базами данных на ЕЯ и др. Известны два подхода к обработке ЕЯ – эмпиризм и рационализм. Первый делает упор на эмпирическом материале (корпуса текстов и устной речи, лексические базы данных, базы звуковых сигналов и др.) и применяет различные статистические методы; второй берет за основу правила, описывающие язык, и занимается разработкой формальных грамматик [5].

В Ассоциации Вычислительной Лингвистики (ACL)[1] действуют специальные группы, объединяющие людей, которые интересуются тем или иным аспектом обработки языка, в т.ч. SIGPARSE[2], занимающаяся проблемами разработки синтаксических анализаторов (парсеров), и SIGSEM[3], занимающаяся вычислительной семантикой. Созданы международные агентуры, поддерживающие сбор, хранение и распределение ресурсов для различных языков, напр., Консорциум лингвистических данных (LDC)[4] и Европейская ассоциация лингвистических ресурсов (ELRA)[5]. Многие фирмы занимаются исследованием и созданием програмного обеспечения для обработки ЕЯ, в т.ч. Google, IBM, Microsoft.

В данной статье мы сосредотачиваемся на синтаксисе и семантике – даем короткий обзор о формализмах, методах и ресурсах, применяемых для синтаксического и семантического анализа, среди которых нам пришлось выбирать, когда мы приступили к моделированию эстонского языка. Во второй части статьи мы рассматриваем имеющиеся или разрабатываемые средства анализа эстонского языка.

2. Автоматический синтаксический анализ

При выборе грамматики для синтаксического анализа следует оценить как ее лингвистические, так и вычислительные аспекты. Некоторые практические системы применяют лексико-функциональную грамматику (LFG), однако анализ с помощью такой грамматики является NP-полным [1]. Другой популярный формализм – вершинная грамматика составляющих (HPSG). Еще одна известная стратегия – анализ синтаксических зависимостей. Многие современные парсеры применяют статистические методы к данным из размеченных корпусов, т.е. эмпирический подход к анализу (напр., вероятностные контекстно-свободные грамматики, метод максимальной энтропии и нейронные сети). Парсеры могут выполнять анализ предложения либо сверху вниз, либо снизу вверх.

[1] http://www.aclweb.org/
[2] http://hmi.ewi.utwente.nl/sigparse/
[3] http://mcs.open.ac.uk/pp2464/sigsem/
[4] http://www.ldc.upenn.edu/
[5] http://www.elra.info/

2.1. Формализмы

2.1.1. Грамматики составляющих

Самая важная математическая система для моделирования структуры составляющих ЕЯ – контекстно-свободная грамматика Хомского (КСГ). Такие грамматики являются ядром многих формальных моделей синтаксиса естественных (а также формальных) языков и могут быть включены в разные приложения [2]. Они имеют достаточную выразительную силу, чтобы представить сложные отношения между словами в предложении, а также чтобы реализовать эффективные алгоритмы анализа. Введение вероятностных расширений в КСГ позволяет добиться результатов анализа, сопоставимых с человеческими.

Правила КС могут применяться, чтобы снабдить любое предложение древовидной синтаксической структурой и, тем самым, образовать корпус, где каждое предложение размечено его деревом анализа. Такой синтаксически аннотированный корпус называется банком синтаксических деревьев (treebank).

Существует большое количество расширений КСГ, которые позволяют учитывать далекие отношения зависимости между составляющими – в том числе, грамматика объединения деревьев (Tree Adjoining Grammar) и различные унификационные грамматики, напр., вышеупомянутая ВГС/HPSG.

2.1.2. Грамматики зависимостей

Наряду с КСГ, применяются и грамматики синтаксических зависимостей (ГЗ). Здесь структура предложения описывается в терминах слов и бинарных синтаксических (или семантических) отношений между ними.

Преимущество формализма зависимостей заключается в его строго предсказуемой силе. Так, зная глагол, мы можем определить, является ли данное существительное его субъектом или объектом. ГЗ позволяют обрабатывать языки со свободным порядком слов, в том числе и эстонский язык (в котором порядок слов относительно свободен, напр., объект может находиться перед глаголом или после него). Грамматика составляющих требовала бы отдельных правил для обоих положений, а в ГЗ достаточно одной ссылки. Таким образом, ГЗ абстрагируется от вариаций порядка слов и представляет только ту информацию, которая нужна для анализа.
Существует много реализаций ГЗ, в том числе грамматика Мельчука (1979), Link Grammar[6] (1993), Constraint Grammar [7] (1995) и др. ГЗ часто применяются для языков, отличных от английского, хотя и для английского языка создано несколько анализаторов.

2.2. Ресурсы

Существует несколько синтаксически размеченных корпусов – банков синтаксических деревьев. Среди них различаются банки составляющих и банки зависимостей. Самый известный банк составляющих – Penn Treebank[7] в Пеннсильванском университете; построены синтаксические деревья для известных корпусов английского языка (Brown, Switchboard, ATIS, Wall Street Journal), а также для арабского и китайского языков. Известный банк зависимостей – Пражский банк чешского языка (Prague Dependency Bank)[8]. Он использовался для разработки вероятностного анализатора чешского языка.

Кроме того, имеются банки деревьев, в которых принято гибридное представление синтаксической структуры, напр., TIGER Treebank для немецкого языка.

Созданы специальные средства визуализации и редактирования деревьев, напр., Annotate, WordFreak.[9]

Для осуществления поиска в банках деревьев создано несколько инструментальных средств, напр., tgrep, Tgrep2.[10]

3. Автоматический семантический анализ

Представление значения впервые было использовано в вопросно-ответных системах 1960-ых годов.

3.1. Формализмы

Выбор формализма для представления значения, естественно, зависит от постановки задачи. Часто применяется предикатная логика первого порядка (ЛП1). После того, как формализм выбран, нужно решить, как

[6] http://www.link.cs.cmu.edu/link/
[7] http://www.cis.upenn.edu/~treebank/
[8] http://ufal.mff.cuni.cz/pdt/
[9] http://www.annotate.org/, http://wordfreak.sourceforge.net/
[10] http://tedlab.mit.edu/~dr/Tgrep2/

привести формулы в соотвествие выражениям ЕЯ. Существуют два основных метода: унификация и лямбда-исчисление. Для выводов применяются метод автоматического доказательства теорем или метод порождения модели [1].

3.1.1. Представление значения

Хорошо известными формализмами для представления значения являются, наряду с ЛП1, семантические сети, концептуальные зависимости и фреймы. Хотя между этими формализмами имеются большие различия, они совпадают на глубинном уровне представления, имея общую предпосылку: значение состоит из структур, образованных из символов некоторого множества.

Семантическая структура ЕЯ в целом имеет предикатно-аргументное строение. Это означает, что между различными понятиями, лежащими в основе составляющих предложение слов и фраз, имеют место особые отношения зависимости.

3.1.2. Семантические роли

Так как нас интересуют взаимосвязи между синтаксисом и семантикой, мы не будем останавливаться на проблемах лексической вычислительной семантики, а переходим к семантическому анализу предложения. Первая задача здесь заключается в определении семантических ролей синтаксических составляющих. Исходя из того, что семантическая структура ЕЯ имеет предикатно-аргументное строение, в предложении необходимо 1) выявить единицы – предикаты, 2) для каждого предиката определить составляющие, функционирующие как его аргументы, и 3) определить, какую семантическую роль выполняет каждая составляющая.

Задача заключается в автоматическом определении того, какие составляющие предложения являются аргументами данного предиката, а затем – в определении роли каждого аргумента. Разметка семантических ролей может улучшить результаты любой отрасли автоматической обработки ЕЯ, актуальными задачами являются поиск информации и ответы на вопросы. Большинство применяемых методов связано с корректируемым машинным обучением. Следовательно, нужны большие размеченные корпуса. В этой роли могут выступать, например, FrameNet[11] или PropBank[12]. В FrameNet для выражения семантических ролей применяется большое количество фреймо-специфических фреймо-элементов (рис. 1), а в PropBank – меньшее количество неспецифических меток аргументов.

```
Frame: STATEMENT
SPEAKER Evelyn said she wanted to leave.
MESSAGE Evelyn announced that she wanted to leave.
ADDRESSEE Evelyn spoke to me about her past.
TOPIC Evelyn's statement about her past.
MEDIUM Evelyn preached to me over the phone.
————————————————————————————————
acknowledge.v, acknowledge.n, add.v, address.v, admission.n, admit.v, affirm.v,
affirmation.n, allegation.n, allege.v, announce.v, announcement.n, assert.v,…
```

Рис. 1. Фрейм STATEMENT в базе фреймов FrameNet. В конце описания фрейма приводится перечень т.н. фреймо-запускающих элементов – слов, наличие которых в тексте запускает процесс заполнения ролей этого фрейма при семантическом анализе

В ходе семантического анализа в предложении находятся предикаты, и для каждого предиката определяется, какую тематическую роль по отношению к нему выполняет та или иная составляющая предложения (рис. 2).

Системы разметки семантических ролей встречаются с многими трудностями. Как FrameNet, так и PropBank предполагают, что связи между конституентами не должны пересекаться, и тематические роли составляющих являются независимыми. На практике это не всегда возможно соблюсти.

Иногда семантические роли размечают в тексте, в котором определены лишь части речи, а синтаксические деревья предложений не построены. Такой подход полезен в случае, если существующий парсер, обычно тренированный на газетных текстах, не может выполнить качественный синтаксический анализ.

[11] http://www.icsi.berkeley.edu/~framenet/

[12] http://verbs.colorado.edu/~mpalmer/projects/ace.html

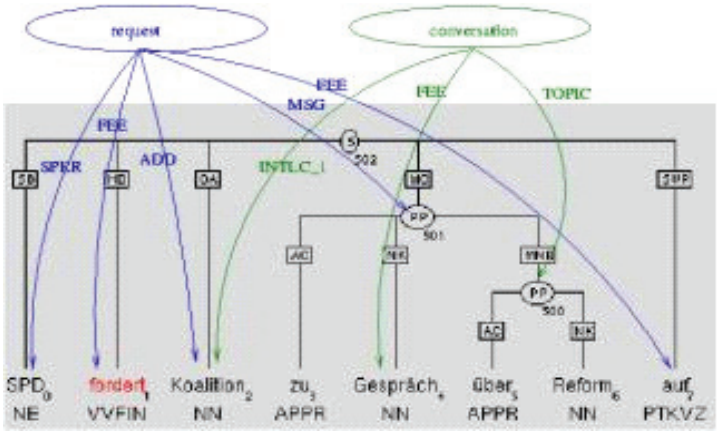

Рис. 2. Анализ предложения "SPD fordert Koalition zu Gespräch" (СПГ призывает коалицию к переговорам). Применяются два фрейма из FrameNet – REQUEST и CONVERSATION[13]

3.2. Ресурсы

FrameNet и PropBank – два грандиозных проекта (предпринятых, соответственно, в Международном институте информатики Великобритании и в Пеннсильванском университете США), общей целью которых является исследование синтаксической реализации семантических ролей в английском языке и разметка семантических ролей в синтаксически размеченном корпусе (соответственно, в Британском национальном корпусе BNC и Penn Treebank). Размеченные корпуса затем могут быть использованы для тренировки семантических анализаторов.

При создании такого корпуса для конкретного языка возникает вопрос: можно ли применять базу фреймов, созданную на основе одного (к примеру, английского) языка, для других языков? Проблема аналогична переводу лексикальной базы данных WordNet на другие языки: оказывается, что языки, хотя и пересекаются, классифицируют мир по-разному. В Саарлендском университете Германии предпринят проект SALSA (SAarbrucken Lexical Semantics Annotation and analysis)[15] с целью создания семантически размеченного корпуса немецкого языка. За основу взят синтаксически размеченный корпус TIGER в объеме 1,5 млн. словоформ газетных текстов. Для разметки семантических ролей применяются фреймы FrameNet (см. рис. 2). В корпусе дополнительно размечены значения слов и анафорические ссылки. Для облегчения разметки, разработана программа SALTO, дающая возможность визуализировать результаты анализа и редактировать схемы деревьев. Результаты представлены в формате XML.

4. Автоматический анализ эстонского языка

4.1. Синтаксический анализ

4.1.1. Грамматика ограничений: поверхностный анализатор

Входом для синтаксического анализатора является выход морфологического анализатора, в нашем случае EstMorf [3]. Основой синтаксического анализа была взята грамматика ограничений (ГО) [4], с учетом того, что этот формализм подходит для языков со свободным порядком слов (EstCG, Estonian Constraint Grammar [6]).
Первый этап анализа – снятие морфологической омонимии (disambiguation), в котором словоформы с неоднозначным морфологическим анализом (таких слов в эстонском языке выше 45% [7]), получают однозначный анализ. Второй этап – определение синтаксических функций слов (предикат, субъект, объект и др.). В результате получается поверхностный синтаксический анализ предложения: определены члены предложения, но не связи между ними (рис. 3).

```
Aknas
    aken+s //_S_ com sg in **CLB // @ADVL
kustus
    kustu+s //_V_ main indic impf ps3 sg ps af  #Intr // @+FMV
tuli
    tuli+0 //_S_ com sg nom // @SUBJ
$.
    . //_Z_ Fst //
```

Рис. 3. Синтаксический анализ предложения „Aknas kustus tuli" (В окне погас свет): определены члены предложения

13 http://www.coli.uni-saarland.de/projects/salsa/

Оценочные показатели парсера неплохие: точность (precision) 78,09–87,57%, полнота (recall) 96,41–98,53%, в зависимости от того, выполняется ли снятие морфологической омонимии автоматически или вручную. Тем не менее, работа над анализатором продолжается.

4.1.2. Корпуса

Для развития синтаксического анализатора создано несколько корпусов. Корпус EstCG, размеченный по принципам ГО, содержит 350 000 словоформ (25 000 предложений).[14]

Кроме того, начаты работы над двумя банками синтаксических деревьев. Первый из них, Sofie Parallel Treebank, является совместной работой исследователей стран Северной Европы.[15] Банк содержит древовидные схемы предложений первой главы романа Гаардена "Мир Софии" на семи языках. Текущий объем эстонской части банка – 50 предложений (рис. 4).[16]

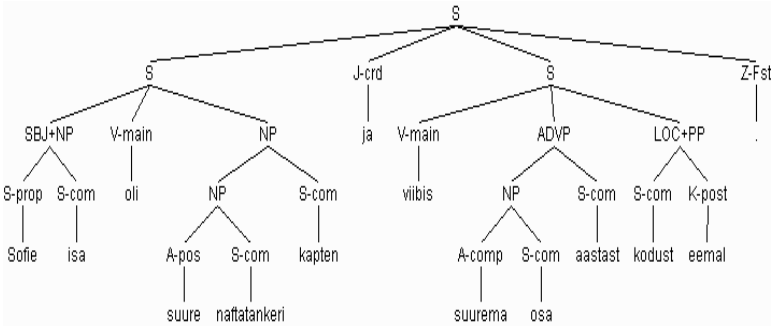

Рис.4. Sofie Parallel Treebank: анализ предложения „Sofie isa oli suure naftatankeri kapten ja viibis suurema osa aastast kodust eemal" (Отец Софии был капитаном большого нефтяного танкера и проводил большую часть года вдали от дома)

Второй банк синтаксических деревьев, Arborest[17], содержит 150 предложений, для которых построены деревья в стиле VISL.[18] Размечены как функции (S=subject, P=predicate и т.д.), так и форма (np, vp и т.д., см. рис. 5).

Деревья построены полуавтоматически: сначала к предложениям, размеченным по EstCG, были применены правила перехода от ГО к фразовой структуре (созданные Экхардом Биком из Южно-Датского университета), затем полученные схемы деревьев были проверены и исправлены вручную. Примерно одна треть структур была построена безошибочно и не требовала исправления.

Начато создание банка деревьев Arborest-2, в который (с учетом того, что современные семантические исследования сосредоточены на предложениях, выражающих движение), будут входить 1000 простых предложений, включающих глаголы движения. Предложения взяты из [8].

Вместе с тем, происходит переход к дополненной версии ГО (EstCG-2), которая, в отличие от EstCG, позволяет выражать и отношения зависимости между членами предложения. Дальнейшая цель – получение фразовой структуры предложения в стиле VISL с применением программы Э. Бика. Деревья VISL, в свою очередь, можно автоматически перевести в формат XML.

4.2. Семантический анализ простого предложения

Работа в области вычислительной семантики до последнего времени концентрировалась на лексической семантике: создании тезауруса типа WordNet и разрешении мнозначности слов. Лексическая база данных для эстонского языка (EstWordNet) содержит 11 000 записей (синонимических множеств, т.н. синсетов), включающих в себя глаголы, существительные и прилагательные.[19]

[14] http://math.ut.ee/~heli_u/syntkorpus.html
[15] http://w3.msi.vxu.se/~nivre/research/nt.html
[16] http://omilia.uio.no/sofie
[17] http://corp.hum.sdu.dk/arborest.html
[18] VISL =visual interactive language learning http://beta.visl.sdu.dk
[19] http://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=en

Составлена программа разрешения многозначности существительных, которая преобразует синтаксические категории EstCG в отношения зависимостей и затем применяет отношения гипоним/гипероним из EstWordNet для нахождения самого “близкого“ значения.[20]

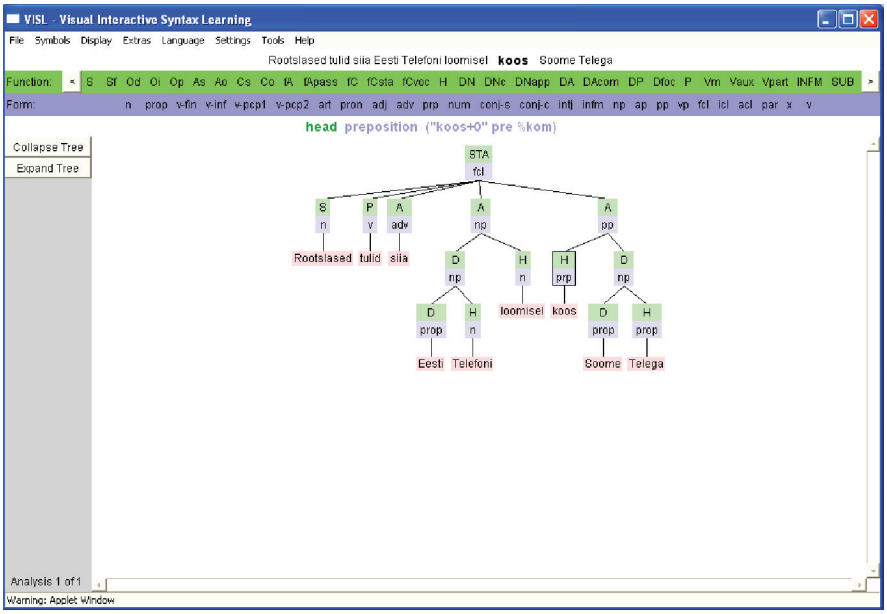

Рис. 5. Arborest: анализ предложения “Rootslased tulid siia Eesti Telefoni loomisel koos Soome Telega” (Шведы пришли сюда при создании «Эстонского телефона» вместе с финским «Теле»)

Изучив имеющиеся лексические ресурсы, мы решили не идти по пути немцев и не перенимать описания понятий из FrameNet или др. баз данных, но, по их опыту, составлять такие описания самостоятельно. В качестве формализма мы применяем фреймы. Тем не менее, где это возможно, мы используем лингвистическое обеспечение, разработанное другими. На рис. 6 приведен анализ простого предложения, в состав которого входит глагол движения (*сесть*). Синтаксический анализ выполнен с помощью EstCG-2, переход к фразовой структуре в стиле VISL произведен с помощью программы Бика, заполнение фреймов и соотнесение их с синтаксическими составляющими предложения выполнено вручную при помощи программы SALTO.

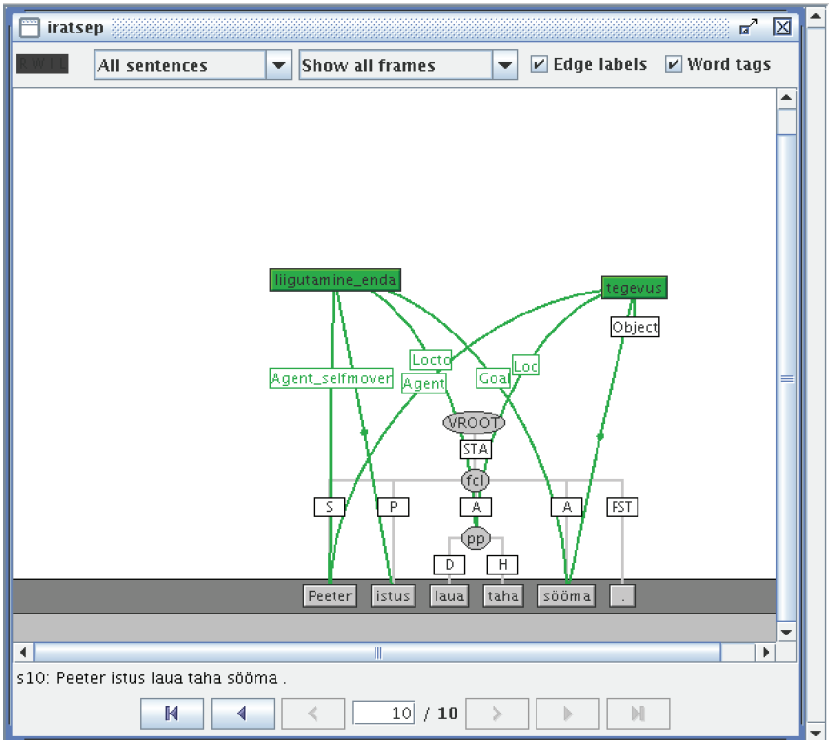

Рис. 6. Анализ предложения “Peeter istus laua taha sööma” (Пеэтер сел за стол есть). В семантическом представлении участвуют два фрейма: liigutamine_enda (двигаться) и tegevus (действие). Исполнителями семантических ролей являются синтаксические составляющие предложения.

Имеющийся семантически размеченный корпус еще мал (несколько десяток предложений), и наша ближайшая работа – дополнение корпуса. Пока это происходит вручную, но в перспективе будут применяться методы машинного обучения.

[20] http://www.cs.ut.ee/~kaarel/NLP/Programs/Semyhe/

5. Заключение

Автоматический анализ текста имеет большое количество важных практических приложений – компьютерные системы, общающиеся с человеком на ЕЯ, МП, поиск документов, автоматическое составление резюме, и многое другое. В данной статье мы рассматривали автоматический анализ синтаксиса и семантики – формализмы, которые применяются для представления структуры предложения или текста, основные методы анализа, применяемые языковые ресурсы. Наш опыт автоматической обработки эстонского языка показывает, что, находясь на перепутье и делая выбор формализма, метода или ресурса, всегда нужно смотреть как назад – на предыдущий опыт (и собственный, и чужой), так и вперед – какие последствия влечет за собой то или иное решение. Переход от синтаксиса к семантике означал для нас построение моста между двумя исследовательскими направлениями, которые до этого развивались сравнительно самостоятельно, но после того, как пробел между ними был заполнен, получили дополнительную мотивацию продолжать свою работу.

Список литературы

1. Blackburn P., Bos J. Computational Semantics // Theoria 2003. 18(1), 27-45. http://homepages.inf.ed.ac.uk/jbos/pubs/theoria.pdf
2. Jurafsky D., Martin J.H. An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall: 2000.
3. Kaalep H., Vaino T. Complete Morphological Analysis in the Linguist's Toolbox // Congressus Nonus Internationalis Fenno-Ugristarum. Tartu: 2001. Pars V, 9-16.
4. Karlsson F., Voutilainen A., Heikkilä J., Anttila A. (Eds.) Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text // Natural Language Processing, No 4. Berlin and New York, Mouton de Gruyter: 1995.
5. Manaris B. Natural Language Processing: An Human-Computer Interaction Perspective // *Advances in Computers* (Marvin V. Zelkowitz, ed.). New York, Academic Press: 1998. Vol. 47, pp. 1-66. http://www.cs.cofc.edu/~manaris/publications/advances-in-computers-vol-47.pdf
6. Müürisep K. Syntactic analysis of Estonian using Constraint Grammar // Proc. DIALOG'98. Kazan: 1998. Vol. 2, 619-625.
7. Puolakainen T. Developing Constraint Grammar for Morphological Disambiguation of Estonian // Proc. DIALOG'98. Kazan: 1998. Vol. 2, 626-630.
9. Rätsep H. Eesti keele lihtlausete tüübid (Типы простых предложений эстонского языка). Tallinn, Valgus: 1978.