

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ФІЛОЛОГІЇ**

**Наталія ДАРЧУК
Ярина ХОДАКІВСЬКА**

**АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ
методичні рекомендації до курсу**

КИЇВ – 2023

Рецензенти:

Зубань О.М. – кандидат філологічних наук, доцент кафедри української мови та прикладної лінгвістики ННІ філології КНУ імені Тараса Шевченка;

Цигвінцева Ю.Л. – доктор філософії, молодший науковий співробітник відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України.

Рекомендовано до друку та оприлюднення на сайті Навчально-наукового інституту філології Вченою радою Навчально-наукового інституту філології, протокол № 4 від 28 листопада 2023 р.

Дарчук Н.П., Ходаківська Я.В.

Автоматичний синтаксичний аналіз: методичні рекомендації до курсу (електронне видання). – К., 2023. – 26 с.

Методичні рекомендації систематизують організацію семінарських та лабораторних занять до курсу «Автоматичний синтаксичний аналіз» комплексної багатосеместрової дисципліни «Прикладна лінгвістика», зокрема у цьому виданні за темами робочої програми дисципліни подано завдання для самостійної роботи, інструкції до користування автоматизованою системою синтаксичного аналізу Корпусу української мови, список літератури до курсу – як основної, так і додаткової, критерії оцінювання, питання до іспиту.

ЗМІСТ

ПЕРЕДМОВА.....	4
НАВЧАЛЬНО-ТЕМАТИЧНИЙ ПЛАН.....	8
Тема 1. Чинники створення АГАТ-синтаксису. Теоретичний синтаксис і автоматичний синтаксичний аналіз.....	9
Тема 2. Роль дистрибутивного підходу у створенні АСА. Метод моделювання..	10
Тема 3. Підходи до опису формального синтаксису. Граматика безпосередніх складників. Граматика залежностей.....	11
Тема 4. Підходи до опису формального синтаксису. Граматика конструкцій «GxC». «Конструктикон» Ч. Філлмора.....	14
Тема 5. Автоматичний аналіз АГАТ з виокремлення словосполучень. Граматики валентностей частин мови.....	15
Тема 6. Граматика залежностей як спосіб представлення синтаксичної структури речення. Зображення присудка і підмета в ГЗ. Зображення структур з однорідністю.....	17
Тема 7. Дерева залежностей у системах автоматичного синтаксичного аналізу. Загальний алгоритм автоматичного синтаксичного аналізу...	20
Тема 8. Автоматичне конструювання частотного словника сполучуваності української мови.....	22
ОРІЄНТОВНІ ТЕМИ ДО ІСПИТУ.....	24
СПЕЦІАЛЬНЕ МЕТОДИЧНЕ (ІТ) ЗАБЕЗПЕЧЕННЯ ДЛЯ ЛАБОРАТОРНИХ РОБІТ.....	25
РЕКОМЕНДОВАНА ЛІТЕРАТУРА	25

ПЕРЕДМОВА

Навчальна дисципліна «Автоматичний синтаксичний аналіз» спрямована на поглиблене ознайомлення студентів із лінгвістичними засадами створення автоматичного синтаксичного аналізу (АСА): місцем автоматичного синтаксичного аналізу у процесі автоматичного опрацювання текстової інформації, основними поняттями синтаксису: речення, типи зв'язків, види семантико-синтаксичних відношень; роллю дистрибутивного методу і методу безпосередніх складників у розробці АСА; граматику залежностей як способом представлення синтаксичної структури речення; деревами залежностей у системах АСА. Об'єктом вивчення у цьому курсі є структура речення, способи вираження синтаксичних граматичних значень. Предмет вивчення дисципліни вивчення законів побудови зв'язного мовлення, які постають в реченнєвих конструкціях. Це відображається в сукупності операцій, які виконуються над послідовностями інформації морфологічного характеру (результатом роботи автоматичного морфологічного аналізу - АМА), що представляють вхідний текст, для встановлення синтаксичних зв'язків між текстовими одиницями у вигляді дерева залежностей. Вихідним текстом для автоматичного синтаксичного аналізу є редукована після АМА послідовність інформації до словоформ. Дисципліна спрямована навчити студентів основам і правилам побудови дерева залежностей речення і створенню системи автоматичного синтаксичного аналізу речення.

Навчальна дисципліна «Автоматичний синтаксичний аналіз» викладається на другому курсі (4 семестр), навчальні заняття проводяться у формі лекцій – 14 год., семінарських занять – 8 год. та лабораторних занять – 8 год., самостійна робота становить 30 год. Дисципліна входить до багатосеместрової комплексної дисципліни «Прикладна лінгвістика» обов'язкової частини ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова». Цей курс є шостою структурною частиною, що викладається після курсів "Корпусна лінгвістика", "Автоматичний морфемно-словотвірний аналіз", "Автоматичний морфологічний аналіз» і читається після теоретичних нормативних курсів «Парадигмологія української мови», «Синтаксис української мови». Для успішного оволодіння знаннями та вміннями у межах дисципліни «Автоматичний синтаксичний аналіз» студент повинен успішно опанувати зазначені попередні курси. Необхідною умовою доступу до дисципліни є такі компетентності студента:

знати категорійно-поняттєвий апарат теоретичної лінгвістики (аспекти аналізу речення в сучасній синтаксичній науці; основні теоретичні поняття синтаксису: одиниці мови/мовлення; граматичне синтаксичне значення,

синтаксичні зв'язки та відношення, що формують речення як цілісну, синтаксично неперервну одиницю; типологію речення, яка охоплює такі його ознаки: будова; здатність членуватися на функціональні компоненти; склад граматичної основи; наявність другорядних членів речення; наявність структурно й семантично необхідних компонентів; мета висловлювання; емоційне забарвлення; спосіб оформлення змісту; наявність ускладнювальних конструкцій) та комп'ютерної лінгвістики (техніку методу моделювання лінгвістичних явищ); основні напрями сучасної прикладної лінгвістики, співвідношення прикладної лінгвістики з теоретичною, структурною і математичною лінгвістикою – фундаментом сучасного прикладного мовознавства;

уміти ефективно працювати з інформацією: шукати і добирати для вирішення професійних завдань необхідні інформаційно-довідникові ресурси з урахуванням сучасних досягнень науки і техніки, критично аналізувати, інтерпретувати, впорядковувати, класифікувати й систематизувати інформацію;

володіти елементарними навичками наукового дослідження та управління інформацією; критичного ставлення до аналізованих явищ; використання іншомовних фахових інформативних джерел; роботи з різного роду текстами; продукування складних усних і письмових повідомлень; взаємодії і співробітництва у навчанні у ситуаціях пошукового характеру.

Навчальна дисципліна «Автоматичний семантичний аналіз» спрямована на формування таких загальних та фахових компетентностей студента за описом ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»:

ЗК 4. Здатність бути критичним і самокритичним.

ЗК 10. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК 11. Здатність застосовувати знання у практичних ситуаціях.

ЗК 12. Навички використання інформаційних і комунікаційних технологій.

ФК 1. Усвідомлення структури філологічної науки та її теоретичних основ, а також набуття базових математичних знань та знань у галузі інформаційних технологій, що визначають прикладну (комп'ютерну) лінгвістику як окрему філологічну спеціалізацію у взаємозв'язку цілісної системи міждисциплінарних знань.

ФК 2. Здатність використовувати в професійній діяльності знання про мову як особливу знакову систему, її природу, функції, рівні та про

структурно-системну організацію писемного та усного тексту як механізму вербальної комунікації.

ФК 7. Здатність до збирання й аналізу, систематизації (створення системоорієнтованих та текстоорієнтованих лінгвістичних баз даних) та інтерпретації мовних, літературних, фольклорних фактів, інтерпретації та перекладу англomовного тексту.

ФК 8. Здатність вільно оперувати спеціальною термінологією для розв'язання професійних завдань у галузі комп'ютерної лінгвістики.

ФК 10. Здатність здійснювати лінгвістичний та літературознавчий аналіз текстів різних стилів та жанрів, а також застосовувати методи структурної, математичної, комп'ютерної лінгвістики у традиційному лінгвістичному аналізі та в комп'ютерному моделюванні процесів аналізу і синтезу лінгвістичних об'єктів і явищ.

ФК 14. Здатність ідентифікувати, класифікувати та формулювати вимоги до програмного забезпечення, включаючи проведення моделювання (формальний опис) його структури, поведінки та процесів функціонування.

ФК 22. Здатність використовувати базові знання фундаментальних розділів математики у завданнях математичної, структурної, комп'ютерної лінгвістики та інженерії програмного забезпечення.

Навчальна дисципліна «Автоматичний синтаксичний аналіз» спрямована на досягнення студентом таких результатів навчання:

Знати:

- категорійно-поняттєвий апарат традиційного і структурного синтаксису; об'єкт, предмет, мету, завдання автоматичного семантичного аналізу;
- основні витoki граматики залежностей – структурний синтаксис Л.Теньєра;
- основні принципи граматики залежностей;
- підходи до розв'язання завдання автоматичного синтаксичного аналізу українськомовного тексту.

Вміти:

- будувати лінгвістичний алгоритм для програмування автоматичного синтаксичного аналізу тексту;
- розробляти метамову опису семантичних структур, одиниці семантичного аналізу;
- конструювати тезаурус як концептуальний опис семантики лексичної системи мови та тексту;
- використовувати інформаційні технології та комунікативні стратегії в наукових дискусіях з питань традиційної та комп'ютерної семасіології;

налагоджувати успішну комунікацію у колективних завданнях із розроблення комп'ютерних програм та проведення автоматичного аналізу на великих лінгвістичних даних;

- обґрунтовувати власний погляд на дискусивні проблеми автоматичного семантичного аналізу.

Навчальна дисципліна «Автоматичний семантичний аналіз» спрямована на досягнення таких програмних результатів навчання за описом ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»:

ПРН 6. Використовувати інформаційні й комунікаційні технології для вирішення складних спеціалізованих задач і проблем професійної діяльності.

ПРН 7. Розуміти основні проблеми філології та підходи до їх розв'язання із застосуванням доцільних методів, зокрема інноваційних міждисциплінарних підходів прикладної (комп'ютерної) лінгвістики та інформаційних технологій, і пояснювати їх взаємозв'язок у цілісній системі міждисциплінарних знань.

ПРН 12. Аналізувати мовні одиниці, явища і процеси методами структурної, математичної та комп'ютерної лінгвістики, представляти в алгоритмічному вигляді процеси аналізу і синтезу лінгвістичних об'єктів.

ПРН 17. Збирати, аналізувати, систематизувати (створювати системоорієнтовані та текстоорієнтовані лінгвістичні бази даних) й інтерпретувати факти мови та мовлення й використовувати їх для розв'язання складних задач і проблем автоматичного лінгвістичного аналізу у спеціалізованих сферах професійної діяльності та/або навчання.

ПРН 25. Проводити передпроектне обстеження предметної лінгвістичної галузі, системний аналіз лінгвістичного об'єкта програмного проєктування.

З метою досягнення поставлених завдань курсу у програмі поряд з українськомовною основною, подано додаткову російськомовну та англійськомовну наукову літературу, що відображає засадничі положення та новітню методологію автоматичного семантичного аналізу як новітньої галузі вузькофахового спрямування. Пропонована російськомовна література не містить інформації антиукраїнського спрямування і не становить загрози державним інтересам України. Іншомовна література може використовуватися за згодою студентів, але при цьому викладач зобов'язується ознайомити здобувачів освіти зі змістом цих матеріалів українською мовою.

НАВЧАЛЬНО-ТЕМАТИЧНИЙ ПЛАН КУРСУ

№ п/п	Назва теми	Кількість годин			
		лекції	семінарські зан.	лабораторні зан.	самостійна робота
Модуль1. Синтаксис словосполучення					
1	Ознайомлення з програмою курсу, формами та методами оцінювання. Теоретичний синтаксис і автоматичний синтаксичний аналіз. Чинники, що були причиною створення АГАТ-синтаксису. Чотири аспекти аналізу речення в СУМі. Об'єкт, предмет, мета АСА	2	1		3
2	Роль дистрибутивного підходу у створенні АСА. Метод моделювання і його роль в АСА.	2	1		3
3	Основні підходи до опису формального синтаксису (граматика безпосередніх складників, граматики залежностей і комбіновані теорії, теорія синтаксичних груп.).	2	2		8
4	Граматики конструкцій «GxC»; «Конструктикон» Ч. Філмора; ймовірнісні контекстно-вільні граматики.	2	1		3
5	Принципи і процедури автоматичного аналізу АГАТ: виокремлення словосполучень. Граматики валентностей всіх частин мови.	2	1		3
Модуль2. Синтаксис речення					
6	Граматики залежностей як спосіб представлення синтаксичної структури речення. Зображення присудка і підмета в ГЗ. Зображення структур з однорідністю. Ознайомлення із автоматизованою системою синтаксичного аналізу в корпусі української мови.	2	1	2	4
7	Дерева залежностей у системах АСА. Загальний алгоритм АСА. Автоматичне / автоматизоване конструювання дерев залежності в корпусі української мови	2		4	2
8	Автоматичне конструювання частотного словника сполучуваності словоформ української мови. ЧС в корпусі української мови.	2	1	2	4

Тема 1. Теоретичний синтаксис і автоматичний синтаксичний аналіз. Чинники створення АГАТ-синтаксису.

Анотація змісту лекції. Мета, завдання, прийоми автоматичного синтаксичного аналізу. Автоматичний синтаксичний аналіз як логічне продовження етапу автоматичного морфологічного аналізу. Взаємозв'язок і взаємозалежність цих етапів. Необхідність практичного дослідження сполучуваності синтаксичних одиниць на значних обсягах тексту. Прикладні аспекти як чинники створення автоматичного синтаксичного аналізу (автоматизація лінгвістичних досліджень, автоматичне визначення меж словосполучень, автоматичне реферування тексту тощо). Особливості АГАТ-синтаксису на базі корпусу української мови (лінгвістичний портал Mova.info) порівняно з іншими автоматичними системами синтаксичного розмічування текстів. Послідовність синтаксичного аналізу: виокремлення словосполучень, із яких автоматично будуються дерева залежності.

Теоретичний синтаксис як вершина у системно-структурній піраміді мови. Речення – основна одиниця синтаксису. Аспекти розгляду речення у синтаксичній науці. Формально-граматичний аспект, який передбачає виокремлення членів речення – компонентів поверхневої синтаксичної структури та встановлення всіх синтаксичних зв'язків між ними (базис для автоматичного синтаксичного аналізу). Семантико-синтаксичний та функціональний аспекти – вирізнення одиниць семантичної природи (синтаксем), що формують глибинну структуру речення. Комунікативна природа речення і явище актуалізації та темо-ремного членування повідомлень.



Питання до семінарського заняття

1. Дати визначення автоматичному синтаксичному аналізу тексту.
2. Яка мета автоматичного синтаксичного аналізу?
3. Перерахувати чинники розвитку автоматичного семантичного аналізу.
4. Які перешкоди на шляху машинної реалізації синтаксичного аналізу?
5. Назвати напрями перетворення формальної структури тексту у смислову.
6. Які аспекти розгляду речення використовуються в автоматичному синтаксичному аналізі?

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К.ВПЦ, 2008. – С. 145–168.
2. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
3. Сучасна українська літературна мова. Морфологія. Синтаксис. – К., 2010.
4. Тестелец Я.Г. Введение в общий синтаксис. М.:РГГУ, 2001, 798 с.

Додаткова:

5. Кудряшова И.М. Анализ лексических валентностей – первый этап семантического анализа ФРАП // Разработка формальной модели естественного языка. – Новосибирск, 1981.
6. Кунце Ю. Введение семантических критериев в синтаксические правила. НГИ /серия 2/, 1981, № 6, с.30-34.
7. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М. Изд-во. Московского Университета, 2011. – 512 с.
8. Мальковский М.Г., Старостин А.С. Модель синтаксиса в системе морфосинтаксического анализа «TREETON» // Труды международной конференции Диалог. М., 2006. С. 481–492.

Тема 2. Роль дистрибутивного підходу у створенні АСА.

Метод моделювання

Анотація змісту лекції. Основними поняттями формального синтаксису є поняття члена речення і типу синтаксичних зв'язків та відношень. При традиційному аналізі виникають труднощі у розмежуванні як головних, так і другорядних членів речення Традиційно виділяються типи синтаксичних відношень: предикативні, непередикативні, сурядні й підрядні. Усі вони нерівнозначні, а сурядні й підрядні взагалі різні, оскільки сурядність – це логічна зв'язка, а підрядність – бінарний зв'язок. При визначенні типів підрядних відношень виникає неоднозначність, яка впливає на результати синтаксичного аналізу речення. Таке семантизоване знання про синтаксичну структуру речення передати комп'ютеру неможливо, мають бути детальні алгоритмічно виконувані правила. Щоб їх створити, треба виявити формальні закономірності, що діють у процесі побудови речення, які б стосувалися форми, а не значення. Дистрибутивний аналіз як вивчення явищ в оточенні інших явищ і у взаємодії з ними. Індуктивний принцип у вивченні структури мови (на вході сума речень, на виході класи мовних одиниць та правила їх сполучуваності). Дедуктивний принцип вивчення структури мови (на вході певні одиниці і певний набір операцій над ними, на виході – породжувані цією системою тексти). Метод моделювання – основний метод у комп'ютерній лінгвістиці. Лінгвістичний опис в дескриптивній лінгвістиці як набір процедур опрацювання тексту, виконання яких приведе до створення моделі мовленнєвої діяльності. Моделі, що імітують дослідницьку діяльність лінгвіста. Увага декриптивістики до зовнішньої мовної форми.



Питання до семінарського заняття

1. Основні поняття формального синтаксису.
2. Чим відрізняється машинний синтаксичний аналіз від традиційного?
3. Дедуктивний та індуктивний принципи вивчення структури мови.
4. Роль дескриптивістики в розвитку автоматичного аналізу тексту.
5. Метод моделювання як основний метод у комп'ютерній лінгвістиці..

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К.ВПЦ, 2008. – С. 145–168.
2. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
3. Сучасна українська літературна мова. Морфологія. Синтаксис. – К., 2010.

Додаткова

4. Pike, Kenneth L. 1954. Language in relation to a unified theory of the structure of human behavior. Glendale, CA: Summer Institute of Linguistics. 170 p.
5. Синтаксический анализ научного текста на ЭВМ. – Киев, 1999. – С. 40–127.
6. Bloomfield L. A set of postulates for the science of language. Language 2 (3): 153–164. 1926
7. Глисон Г. Введение в дескриптивную лингвистику. М., 1959.
8. Атаян Э. Р. Предмет и основные понятия структурного синтаксиса. Ереван. 115. гос. ун-т. – Ереван : Митк, 1968. – 280 с.

ТЕМА 3. Підходи до опису формального синтаксису.

Грамматика безпосередніх складників. Грамматика залежностей

Анотація змісту лекції. Метод аналізу за безпосередніми складниками став подальшим розвитком дистрибутивного методу на синтаксичному рівні. Метод базується на таких припущеннях: істотну роль відіграє тільки одне відношення – відношення підрядності; речення аналізується шляхом послідовного поєднання одне з одним дрібних його складників: спочатку слова поєднуються одне з одним, утворюючи певний складник, потім цей складник поєднується з іншим складником або окремим словом, й утворюється новий складник і т.д. Останній акт побудови речення – це поєднання груп підмета і присудка. В результаті все речення буде представлене у вигляді єдиного блока, а оскільки відношення підрядності бінарне, то кожний блок складається не більше, ніж із двох частин. Усі групи, які послідовно виникають у ході аналізу, є складниками. Речення – максимальний складник; мінімальні складники – окремі слова. Група підмета і група присудка – це безпосередні складники речення; групи, які складаються більше ніж з одного слова, мають свої безпосередні складники. Увесь процес аналізу може бути зображений трьома способами: у вигляді правил підстановки, дужкового запису та у вигляді дерева.

Граматика безпосередніх складників відображає будову речення і пояснює спосіб його породження (деривацію). Ряд правил «перепишування» дозволяють перетворювати абстрактні позначення (іменна група, дієслово) у дрібніші лінійні ланцюжки, доки не будуть складатися з кінцевих (термінальних) позначень – конкретних словоформ (або морфем). Процедурі синтаксичного розбору можна звести до пошуку всіх можливих деривацій аналізованого висловлювання. До числа таких алгоритмів відносяться низхідні (Top-Down), висхідні (Bottom-Up) і комбіновані (Left-Corner) алгоритми.



Питання до семінарського заняття

1. Принципи аналізу речення за методом безпосередніх складників.
2. Обмеження та труднощі використання методу в автоматичних аналізаторах.
3. Дериваційний характер граматики безпосередніх складників.
4. Правила «перепишування» ланцюжків, послідовність етапів поєднання складників та груп у речення.

Завдання для самостійної роботи

Дібрати 10 простих речень (українською або англійською мовами). Проаналізувати їх за методом безпосередніх складників трьома способами: 1) за правилами підстановки; 2) дужкового запису; 3) у вигляді дерева, використавши коди: V – дієслово; А – атрибутивний клас; N – іменник; D – прислівник для класів повнозначних слів і підрядкові символи: n – наз. відм., g – род. відм., d – дав. відм., o – знах. відм., i – ор. відм., p – місц. відм. Прийменники і частки розглядати як нульові елементи, не виділяючи їх в окремий складник. Укласти список синтагм та їх перетворень.

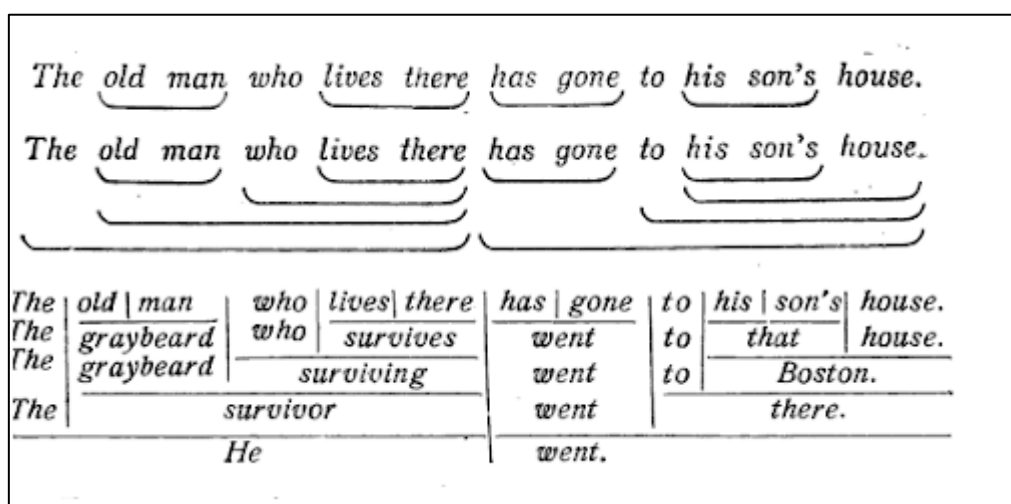


Рис. 1. Схема аналізу за безпосередніми складниками (Глісон, «Вступ до дескриптивної лінгвістики»)



Рис. 2. Схема аналізу за безпосередніми складниками
(Н.Дарчук. Комп'ютерне анотування українського тексту...)

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К. ВПЦ, 2008. – С. 145–168.
2. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
3. Глисон Г. Введение в дескриптивную лингвистику. М., 1959.

Додаткова

4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. Учеб. пособие / Большакова Е.И. И др. — М.: МИЭМ, 2011. — 272 с.
5. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. — М.: Наука, 1985. — 144 с.
6. Ермаков А. Е. Неполный синтаксический анализ текста в информационно ... Language technologies, 2001. – Morristown, NJ, USA, 2001. – P. 1–9.
7. Лангенбах М. О. Автоматизація стилістичних досліджень українських текстів ... автореф. дис. на здобуття наук. ступеня канд. філол. наук : спец. 10.01.08
8. Croft, W., & Cruise, A. D.. Cognitive Linguistics. Cambridge: Cambridge University Press. 2004
9. Fillmore C J & Kay P. Construction grammar cour-. sebook. 1993
10. Bachman F. Performance on cloze tests with fixed-ratio and rational deletions. TESOL Quarterly,. 19(3):535–556. 1985.

11. Pike, Kenneth L. 1954. Language in relation to a unified theory of the structure of human behavior. Glendale, CA: Summer Institute of Linguistics. 170 p.

ТЕМА 4. Підходи до опису формального синтаксису.

Грамматика конструкцій «GxC». «Конструктикон» Ч. Філлмора

Анотація змісту лекції. Один із підходів до опису структури речення – виявлення його трансформаційної структури і встановлення відношення між реченням і певною кількістю його перетворень. Цей спосіб аналізу речень базується на еквівалентності структур вихідних речень і трансформів, тому виходить за рамки автоматизації. Завдання з автоматичного виокремлення синтагм (аналог словосполучень) може вирішуватися статистичним шляхом. Для кожної словоформи на основі великого корпусу текстів підраховується частота його вживання і частота вживання пар словоформ. Для пар з високою ймовірністю вживання робиться припущення, що вони є словосполученнями (так звані «n-грами»).

Грамматика конструкцій «GxC» була започаткована у 1970-х роках. Конструкції складаються з «батьківських» і «дочірніх» елементів; відношення між ними відрізняються за ступенем жорсткості; за конструкціями можна визначати не тільки їх синтаксичні, а й лексичні, семантичні, прагматичні параметри; конструкції можуть бути ідіоматичними. Множина конструкцій є неоднорідною, вони розрізняються ступенем жорсткості і типом ідіоматичності.

Варіант граматики конструкцій – «Конструктикон» Ч. Філлмора, у якому розвиваються принципи проекту FrameNet на матеріалі корпусу речень із розміткою конструкцій. Ймовірнісні оцінки можуть бути одержані лише на підставі статистичних даних, а статистичні дані залежатимуть від жанрово-стилістичного типу.

В основі сучасних статистичних алгоритмів – формалізм ймовірнісних контекстно-вільних грамастик. До кожного правила переписування додається ймовірність застосування цього правила.

Формально-граматичні синтаксичні моделі не здатні охопити всю різноманітність мовної системи. Застосування різного роду формалізмів для структурування речення призводить до втрати правильного синтаксичного представлення через незняту омонімію, довжину зв'язку між словоформами, наявність сурядних конструкцій, які ускладнюють структуру речення і порушують його формалізований аналіз.



Питання до семінарського заняття

1. Що таке ймовірнісний парсинг?
2. Охарактеризуйте граматику конструкцій «GxC» та систему її елементів.
3. «Конструктикон» Ч. Філлмора як розвиток граматики конструкцій.
4. Трактують термінів «автоматичний словник» / «резидентний словник».
5. Переваги та недоліки граматики конструкцій.

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К. ВПЦ, 2008. – С. 145–168.
2. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
3. Лангенбах М. О. Автоматизація стилістичних досліджень українських текстів ... автореф. дис. на здобуття наук. ступеня канд. філол. наук : спец. 10.01.08

Додаткова

4. Сучасна українська літературна мова. Морфологія. Синтаксис. – К., 2010
5. Загнітко А.П. Основи українського теоретичного синтаксису. – Ч. 1. – Горлівка: ГДПШМ, 2004. – 228 с.; ч. 2. – 254 с.; Ч. 3. – 266 с.
6. Вихованець І. Р. Теоретична морфологія української мови: Академ. граматики укр. мови. К.: Унів. вид-во «Пульсари», 2004. 400 с.
7. Баранов А.Н. Идиоматичность и идиомы / А.Н. Баранов, Д.О. Добровольский // Вопросы языкознания. – 1996. – № 5. – С. 51 – 64.
8. Борисова Е. Г. Принципы описания коллокаций (на материале русского языка): Дис. на соиск. уч. степ, доктора филол. наук. М., 1996. 364 с. 21

ТЕМА 5. Автоматичний аналіз АГАТ: виокремлення словосполучень.

Граматики валентностей частин мови

Анотація змісту лекції. Одноцільовий парсер – евристичний алгоритм синтаксичного аналізатора, який використовує конкретне правило для конкретної словоформи. Різні частини мови вимагають за своєю валентністю відповідних актантів. Валентність слова – властивість слова певним чином реалізовуватися в реченні і вступати в певні комбінації з іншими словами. Л. Теньєр розробляв поняття валентності як здатності дієслова керувати певним числом актантів (суб'єкт, обставина, додаток). Послідовники Теньєра акцентували якісний аспект валентності: морфологічні, конструктивні, семантичні особливості головного і залежного компонентів. Поштовхом для подальшого розвитку цієї теорії була розробка алгоритмів автоматичного перекладу, які будуються за принципом пошуку елементів, які «насичують

валентність» головного слова. Поняття валентності почали тлумачити як потенційний зв'язок одного мовного елемента (фонема, морфема, слово, словосполучення) з іншими мовними елементами, а сполучуваність – як реальний зв'язок мовної одиниці з іншими одиницями у мовленні.

У системі автоматичного синтаксичного аналізу АГАТ виокремлення словосполучень базується на граматиці валентностей. Побудовані комп'ютерні підграматики валентностей іменника, прикметника, займенника, числівника, дієслова та прислівника за принципом: вказано лексему, прийменник, який бере участь у керуванні, та відмінок іменникової словоформи у вигляді двобуквенного коду.



Питання до семінарського заняття

1. Що таке валентність слова?
2. Як валентність використовується в автоматичному синтаксичному аналізі?
3. Валентність дієслова.
4. Валентність іменника та прикметника.
5. Принципи виокремлення словосполучень у граматиці АГАТ.

Рекомендована література:

Основна:

9. Дарчук Н.П. Комп'ютерна лінгвістика. К.ВПЦ, 2008. – С. 145–168.
10. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
11. Лангенбах М. О. Автоматизація стилістичних досліджень українських текстів ... автореф. дис. на здобуття наук. ступеня канд. філол. наук : спец. 10.01.08

Додаткова

12. Сучасна українська літературна мова. Морфологія. Синтаксис. – К., 2010
13. Загнітко А.П. Основи українського теоретичного синтаксису. – Ч. 1. – Горлівка: ГДПШМ, 2004. – 228 с.; ч. 2. – 254 с.; Ч. 3. – 266 с.
14. Вихованець І. Р. Теоретична морфологія української мови: Академ. грамика укр. мови. К.: Унів. вид-во «Пульсари», 2004. 400 с.
15. Баранов А.Н. Идиоматичность и идиомы / А.Н. Баранов, Д.О. Добровольский // Вопросы языкознания. – 1996. – № 5. – С. 51 – 64.
16. Борисова Е. Г. Принципы описания коллокаций (на материале русского языка): Дис. на соиск. уч. степ, доктора филол. наук. М., 1996. 364 с. 21

Тема 6. Граматика залежностей як спосіб представлення синтаксичної структури речення.

Анотація змісту лекції. Монографія Л.Теньєра «Основи структурного синтаксису» (1959) стала базою для розробки граматики залежностей, теорії валентності, семантичного синтаксису, відмінкової граматики. Теньєр розрізняє структурний порядок у реченні (відношення між словами, їх взаємозалежність) і лінійний порядок (розташування слів у мовленнєвому ланцюжку). Лінійний порядок одновимірний, структурний – багатовимірний.

Концепція Теньєра відображає мовленнєву діяльність мовця: говорити означає перетворювати структурний порядок на лінійний, розуміти мовлення – здійснювати протилежне. У структурі речень взаємодіють три чинники: форма вираження (морфологія, формальний порядок слів, які складають «одяг думки»); форма змісту (структурний синтаксис, який відображає ієрархічні відношення між словами); субстанція змісту – власне семантика. Перші два утворюють структуру мови, семантика залишається за її межами і має вивчатися також логікою та психологією.

Статичний синтаксис, за Л. Теньєром, охоплює парадигми слів, які існують у свідомості мовця. Динамічний – це впорядкування статичних елементів у свідомості разом із структурними моделями речень. Динамічний синтаксис базується на трьох поняттях: синтаксичний зв'язок (коннексія), який визначає будову реального речення; юнкція; трансляція. Синтаксичний зв'язок може бути вертикальним – при об'єднанні керованого і керувального членів, і горизонтальним – при об'єднанні супідрядних компонентів (прикладка, однорідні члени). Дієслово в реченні має абсолютну домінуючу позицію. Йому підпорядковуються субстантивні елементи – актанти (суб'єкт, прямий і непрямий об'єкти) і сирконстанти (обставини).

Праця І. Севбо «Графічне представлення синтаксичних структур і стилістична діагностика» як розвиток ідей граматики залежності. Видозміна та переосмислення традиційної концепції конструювання речень зі словосполучень і словосполучень зі слів. Основні принципи граматики залежностей за І.Севбо.

Коренем дерева вважається присудок, а підмет розглядається як заповнювач валентного місця поряд з іншими залежними елементами. При цьому вважається, що: а) корінь дерева не залежить від інших вузлів речення; б) від кореня прямо або непрямі залежать усі інші вузли речення. Однорідність присудків (і підметів) в АСА трактується як синтаксична однотипність сурядних членів.



Питання до семінарського заняття

1. Мовленнєва діяльність мовця за концепцією Теньєра.
2. Статичний і динамічний синтаксис за Теньєром..
3. Ядро, актанти, сирконстанти.
4. Ієрархічність синтаксичних структур у граматиці залежностей.
5. Що означає вербоцентричність структури речення?
6. Основні принципи граматики залежностей за І.Севбо.

Завдання для самостійної роботи

Порівняння дерев залежності.

Розгляньте наведені приклади синтаксичних дерев (англійського підкорпусу в Корпусі чеської мови, проєкту Universal Dependencies, Корпусу української мови на порталі Mova.info).

Виходячи зі своїх спостережень вкажіть на спільні типологічні риси синтаксичних дерев та на характерні риси, притаманні кожному із варіантів представлених зображень.

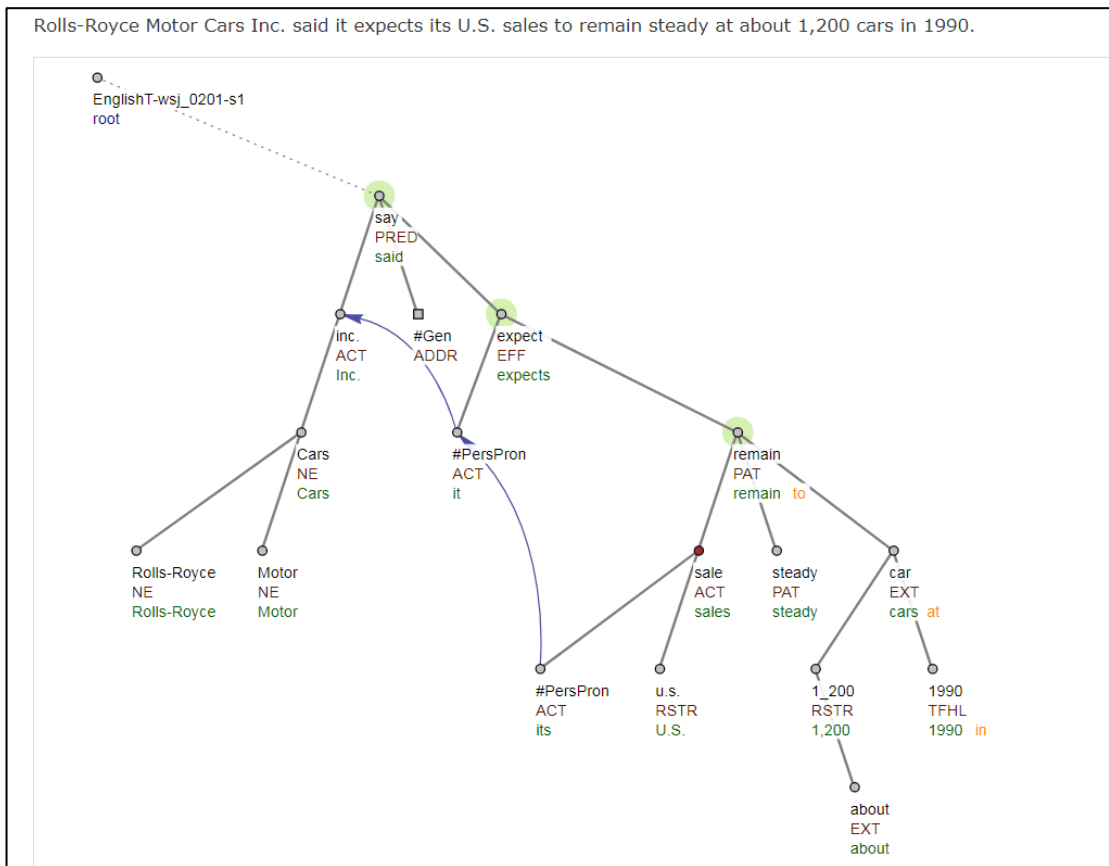


Рис. 3. Дерево залежності англійського підкорпусу в Корпусі чеської мови

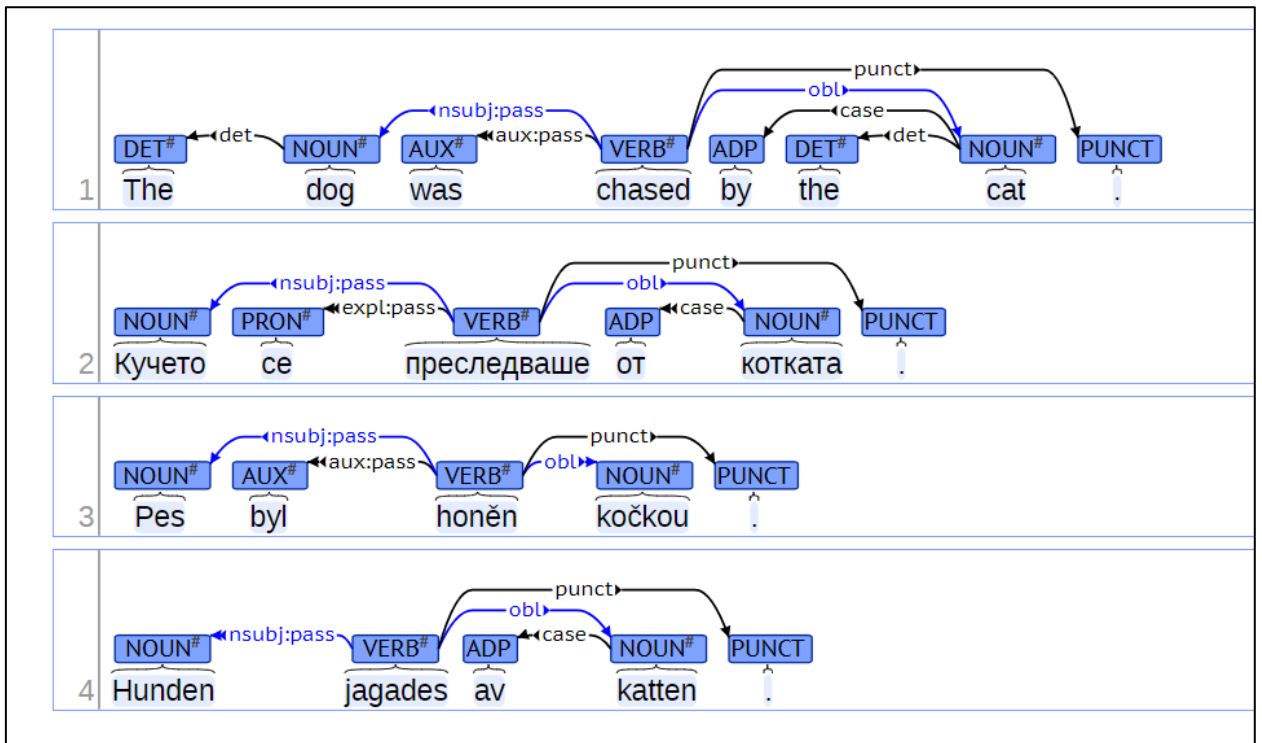


Рис. 4. Дерево залежності проєкту Universal Dependencies

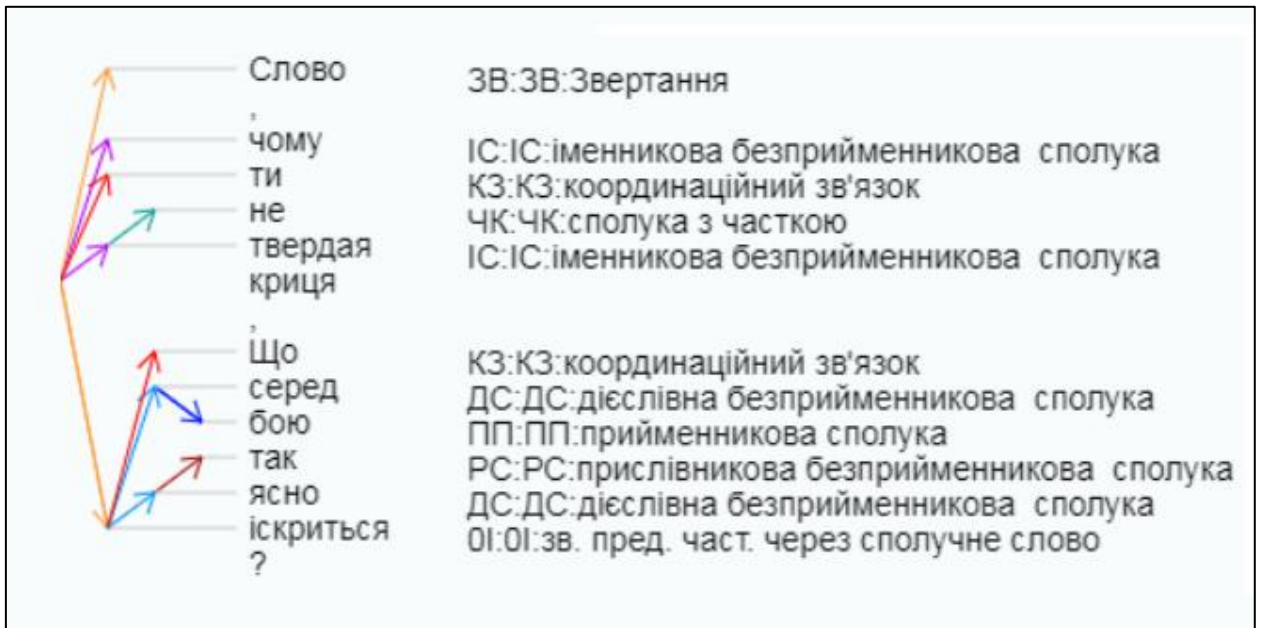


Рис. 5. Дерево залежності Корпусу української мови на порталі Mova.info

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К:ВГЦ, 2008.
2. Дарчук Н.П. Комп'ютерне анотування українського тексту. Результати і перспективи – К.: Освіта України. – 2013.
3. Карпіловська Є.А. Вступ до комп'ютерної лінгвістики. – Донецьк: Юго-Восток – 2003.
4. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. - Киев : Наук. думка, 1981. - 192 с.
5. Теньер Л. Основы структурного синтаксиса. / Пер. с франц. Вступ. ст. и общ. ред. В. Г. Гака. — М.: Прогресс, 1988. — 656 с.

Додаткова

6. Кучеренко І.К. Актуальні проблеми граматики. Львів: Світ, 2003. 226 с ...
Львів: Світ, 2003. 226 с.
7. Падучева, Е.В. О способах представления синтаксической структуры предложения // Вопросы языкознания. 1964, N 2. С. 99-113.
8. Фитиалов С.Я. О моделировании синтаксиса в структурной лингвистике.- М., 1962.- С. 100-114.
9. Холодович А. А. К вопросу о доминанте предложения.— Ленинград, 1979.— С. 293-298.

Тема 7. Деревя залежностей у системах автоматичного синтаксичного аналізу. Загальний алгоритм автоматичного синтаксичного аналізу

Анотація змісту лекції. В АГАТ-синтаксисі використовувалися послідовний і передбачувальний аналіз (виражаються у наявності граматики сполучуваності кожної частини мови і кожної лексеми), а також методика опорних точок (при побудові алгоритмів і програми ходу синтаксичного аналізу). Автоматичний синтаксичний аналіз першого етапу здійснюється за допомогою підграматик, які описують дієслівні, іменні, ад'єктивні словосполучення і зв'язки залежності між членами сполук. Алгоритми II–III етапів побудовано у вигляді наближених моделей: спочатку на базі певних понять і недостатніх фактичних даних будується діюча, але недосконала модель сегментації тексту. На IV етапі відбувається остаточний поділ складного речення на предикативні частини і встановлення зв'язків між ними.

Формальні засоби мови для позначення синтаксичного зв'язку між словами: морфологічна інформація; інформація про сполучуваність граматичних класів у межах певних синтаксичних конструкцій; дані про пунктуаційні знаки, використані для структурування речення; дані про позиційні умови реалізації певних синтаксичних зв'язків; лексична (списки слів, які можуть бути допоміжними словами у складі присудка або у підмета). Перші чотири види інформації належать до граматичного рівня і є

основними, а п'ятий є додатковим і звертання до нього обмежене в межах описуваної системи АГАТ.



Питання до семінарського заняття

1. Що таке граматики залежностей?
2. Що таке дерево залежностей?
3. Послідовність побудови дерева залежностей.
4. Труднощі і особливості методу дерев залежностей.
5. Представлення підмета і присудка у дереві залежностей.
6. Представлення однорідних членів речення у дереві залежностей.

Завдання для лабораторної роботи

Побудова дерев залежності в корпусі української мови

Ознайомтеся з інструкціями щодо побудови дерев залежності в Корпусі української мови на порталі Mova.info (відеозапис інструкцій Наталії Петрівни Дарчук додається). Побудуйте дерева залежності до текстів, внесених вами до корпусу зі знятою граматичною омонімією.

MOVA.info
ЛІНГВІСТИЧНИЙ ПОРТАЛ

СЛОВНИКИ КОРПУС ▾ ПРОЕКТИ РІЗНЕ ▾

<< 6 >>

На (ПП) білих (АЯ) вікнах (ЛЯ) змерзли (ГН) міражі (ЙА) . ()

На	вікнах	прийменникова сполука
вікнах	білих	іменникова безприйменникова сполука
змерзли	міражі	координаційний зв'язок
змерзли	На	дієслівна прийменникова сполука

Зберегти

New

stats статистичні параметри речення

Edit text params

Тестування АСА

Словосполучення на базі синтаксису

Виключення

змерзли [КЗ] міражі

змерзли [ЛП] На [ПП] вікнах [ІС] білих

На
білих
вікнах
змерзли
міражі

ДПЛ: ДП: дієслівна прийменникова сполука
ІС: ІС: іменникова безприйменникова сполука
ПП: ПП: прийменникова сполука
КЗ: КЗ: координаційний зв'язок

Рис. 6. Інтерфейс роботи з деревами залежності у Корпусі української мови на порталі Mova.info

Відеоінструкції Наталії Петрівни Дарчук з побудови дерев залежності в корпусі:

<https://www.youtube.com/watch?v=t-CkQoi17K8&authuser=0>

<https://www.youtube.com/watch?v=52TnWjcRGF4&authuser=0>

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К:ВГЦ, 2008.
2. Дарчук Наталія. Комп'ютерне анотування українського тексту. Результати і перспективи – К.: Освіта України. – 2013.
3. Карпіловська Є.А. Вступ до комп'ютерної лінгвістики. – Донецьк: Юго-Восток – 2003.

Додаткова

4. Городенська К.Г. Граматичний словник української мови : Сполучники. – Херсон : Вид-во ХДУ, 2007. – 340 с.
5. Критская В.И. Знаки препинания как текстообразующие единицы: Автореф. дис. канд. филол.наук.- Киев,1991. 20 с. 219.
6. Синтаксический анализ научного текста на ЭВМ. – Киев, 1999. – С. 40–127.

Тема 8. Автоматичне конструювання частотного словника сполучуваності української мови

Анотація змісту лекції. Словосполучення – важлива одиниця в породженні мовлення, відіграє конструктивну роль при побудові речення. Реципієнт сприймає мовленнєвий потік як окремі смислові комплекси, адекватні словосполученням. В квантитативній лінгвістиці елементи словника мовної системи – це одиниці, які зберігаються і відтворюються більшістю носіїв мови у цілісному вигляді (слова, словосполучення). Матеріалом для частотного словника словосполучень була перша тисяча найчастотніших повнозначних слів, відібраних з частотного словника української мови Корпусу української мови на порталі Mova.info. До списку включено тільки повнозначні слова: іменники, прикметники, дієслова, прислівники. Прийменники, сполучники, частки, вигуки, числівники, допоміжні і модальні дієслова не включені. Сполучуваність обстежується за текстами трьох функціональних стилів: белетристичного; наукового; публіцистичного. Робота з укладання словника складається з таких етапів: статистичне опрацювання матеріалу; лінгвістичне опрацювання матеріалу; укладення словникової статті.

Мінімальною достатньою для статистичного дослідження вважалася підвибірка з тисячі слововживань. Автоматично було поділено матеріал на підвибірки і обраховано абсолютну частоту слів, а потім середню (ділення абсолютної частоти на кількість мінімальних підвбірок). Це дало змогу порівнювати дані про вживаність певного слова у різних стилях і жанрах. Кожне джерело аналізується повністю. Граничним контекстом було обрано речення.



Питання до семінарського заняття

1. Що таке словосполучення.
2. Виокремлення словосполучень у автоматичному синтаксичному аналізі?
3. Що таке сполучуваність?
4. Що таке n-грами?
5. Етапи створення словника сполучуваності.
6. Статистичні принципи створення словника сполучуваності.

Рекомендована література:

Основна:

1. Дарчук Н.П. Комп'ютерна лінгвістика. К:ВГЦ, 2008.
2. Дарчук Н.П. Комп'ютерне анотування українського тексту. Результати і перспективи – К.: Освіта України. – 2013.
3. Дарчук Н.П. До питання про створення автоматичного словника словосполучень української мови / Н.П.Дарчук // «Science and Education. A New Dimension». Philology. V(42), Issue 149, 2018/ С. 15–19.

Додаткова

4. Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов / Э. С. Клышинский, Н. А. Кочеткова, М. И. Литвинов, В. Ю. Максимов // Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог», 26-30 мая 2010 г., Бекасово. – М.: Изд-во РГГУ, 2010. – 9 (16). – С. 181-185.
5. Марчук Ю.Н. Компьютерная лингвистика. – М., 2007.

ОРІЄНТОВНІ ТЕМИ ДО ІСПИТУ

1. Валентність і сполучуваність.
2. Граматика безпосередніх складників, відмінність від стандартної граматики. Недоліки граматики БС.
3. Граматика залежностей як спосіб представлення синтаксичної структури речення.
4. Графічні способи представлення синтаксичних структур у традиційній граматиці, граматиці Л. Теньєра, граматиці залежностей, граматиці безпосередніх складників.
5. Дескриптивна лінгвістика в дослідницькій діяльності лінгвіста.
6. Деякі поняття теорії графів (граф, направлений граф, вузол графа, орієнтований граф, проєктивне дерево).
7. Загальний алгоритм автоматичного синтаксичного аналізу тексту.
8. Зображення присудка, підмета у дереві залежностей.
9. Зображення структур з однорідністю у дереві залежностей. Зауваження про складне речення.
10. Лінгвістичні засади алгоритмів синтаксичного аналізу.
11. Передумови автоматизації синтаксичного аналізу. Деякі поняття теорії графів (граф, направлений граф, вузол графа, орієнтований граф, проєктивне дерево).
12. Питання вивчення синтаксичної структури тексту за допомогою ЕОМ.
13. Підходи до визначення синтаксичної структури тексту.
14. Підходи до визначення синтаксичної структури тексту: формально-граматичний, семантико-синтаксичний, функціональний, комунікативний.
15. Проблема слова й напрямок зв'язку у дереві залежностей.
16. Роль дистрибутивного методу у створенні АСА.
17. Роль і місце синтаксичного аналізу в процесі автоматичного опрацювання текстової інформації.
18. Теоретична синтаксична концепція Л.Теньєра.
19. Типи алгоритмів автоматичного синтаксичного аналізу (АСА).
20. Традиційний синтаксис й автоматичний синтаксичний аналіз тексту.
21. Що таке словосполучення? Принципи виокремлення словосполучень із тексту в системі АСА. Підрядні, предикативні і сурядні словосполучення.

**СПЕЦІАЛЬНЕ МЕТОДИЧНЕ (ІТ) ЗАБЕЗПЕЧЕННЯ
ДЛЯ ЛАБОРАТОРНИХ РОБІТ
навчально-наукової лабораторії комп'ютерної лінгвістики
<https://philology.knu.ua/struktura-if/laboratoriyi/laboratoriya-kompyuternoyi-linhvisty/> :**

1. Синтаксичний модуль / Корпус української мови
<http://www.mova.info/login.aspx>;
2. Корпусрайтер: синтаксичний модуль
https://www.youtube.com/watch?v=fM_Jzg8kDRA,
<https://www.youtube.com/watch?v=52TnWjcRGF4>.
3. Частотні словники / Корпус української мови
<http://www.mova.info/carticle.aspx?l1=210&DID=5215>.
4. Електронний словник мови Т. Шевченка. – Режим доступу:
<http://www.mova.info/cfqsh.aspx>.
5. Електронний словник мови Лесі Українки – Режим доступу
http://www.mova.info/corpus_msint.aspx?tid=2455.
6. Корпус української мови: Синтаксис
http://www.mova.info/syntaxis_search.aspx.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Основна

1. Вихованець І.Р. Граматика української мови. К.:Либідь, 1993. – 348 с.
2. Дарчук Н. Комп'ютерна лінгвістика. – К: ВПЦ, 2008 (розділ 2).
3. Дарчук Н. Комп'ютерне анотування українського тексту. Результати і перспективи. – К.: Освіта України, 2015. (розділ 3).
4. Дарчук Н.П. Дерево залежностей як параметр авторського стилю. // Монографія «У пошуках гармонії мови». Упорядники: Є.А Карпіловська, Л.П.Кислюк, Ю.В.Романюк, А.О.Савенко. – К.: «Видавничий дім Дмитра Бураго». 2020. – С.232 – 249.
5. Севбо І.П. Графическое представление синтаксических структур и стилистическая диагностика. – К.: Наукова думка, 1981.
6. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.:Наука., 1985. 144 с.
7. Леонтьева Н.Н. Автоматическое понимание текстов. Системы, модели, ресурсы.: Academia, 2006.303 с.Тетьер Л. Основы структурного синтаксиса. М.:Прогресс, 1988, 656 с.
8. Самотюк О.Л. Сучасні технології лінгвістичних досліджень. Львів: Вид-во Львівської політехніки. 2011. – 152 с.
9. Сучасна українська літературна мова: Морфологія. Синтаксис: підручник / А.К.Мойсієнко, І.М.Арібжанова, В.В.Коломийцева та ін. – К.:знання, 2010. – 373 с.
- 10.Тестелец Я.Г. Введение в общий синтаксис. М.:РГГУ, 2001, 798 с.
CohenS.B.,Satta G., Collins M. Approximate PCFG Parsing Using Tensor Decomposition //Proc of NAACL 2015

Додаткова

1. Дарчук Н. «Золотий корпус» творів Лесі Українки. // SCIENCE AND EDUCATION A NEW DIMENSION, p-ISSN 2308-5258 e-ISSN 2308-1996 VIII(69), Issue 234, 2020 Sept. <https://doi.org/10.31174/SEND-Ph2020-234VIII69> С. 7-11.
2. Дарчук Н.П. АГАТ-синтаксис як компонент комп'ютерної граматики української мови / Українська мова. 2020. № 4 (76). С. 89–102.
3. Діалог-95. Труды международного семинара по компьютерной лингвистике и ее приложениям. Казань, 31 мая – 4 июня 1995.
4. Зарубина Н.Д. Текст: лингвистический и методический аспекты. – М., 1981.
5. Звегинцев В.А. Предложение и его отношение к языку и речи. – М., 1976.
6. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. – М., 2007.
7. Лангенбах, М. "Автоматичний синтаксичний аналіз речення за принципами граматики залежностей." Науковий вісник Східноєвропейського національного університету імені Лесі Українки. Філологічні науки. Мовознавство 3 (2015): 249-254. Марчук Ю.Н. Компьютерная лингвистика. – М., 2007.
8. Москальская О.И. Грамматика текста. – М., 1981.
9. Леонтьева Н.Н., Вишнякова С.М. Опыт автоматического индексирования со смысловым сжатием. // Науч.-техн. информ. – М., 1977. – Сер. 2. – № 7. – С. 24–30.
10. Партико З.В. Прикладна і комп'ютерна лінгвістика. – Львів – 2008.
11. Перебийніс В.І. Математична лінгвістика. – К., 2014. Прикладная и компьютерная лингвистика. - Санкт-Петербург, 2015, стор.60 - 95

Електронні джерела:

1. Система Этап-3: <http://proling.iitp.ru/etap3>
2. Stanford NLP <http://nlp.stanford.edu:8080/corenlp/process>
3. OpenNLP: <https://opennlp.apache.org>