https://www.sworldjournal.com/index.php/swj/article/view/swj14-01-020
DOI: 10.30888/2663-5712.2022-14-01-020

THE ALGORITHM OF ANALYTHING ENGLISH TEXTS WITH THE HELP OF PARSERS

АЛГОРИТМ АНАЛІЗУ АНГЛОМОВНОГО ТЕКСТУ ЗА ДОПОМОГОЮ ПАРСЕРІВ

Parnus K.I. / Парнус К.І.
ORCID: 0000-0003-1811-0980
викладач кафедри іноземних мов / lecturer of the department of foreign languages
Svysiuk O. V. / Свисюк О.В.
ст. викладач кафедри іноземних мов / senior lecturer of the department of foreign languages
Zhytomyr Polytechnic State University, Zhytomyr, Zhytomyr region, Ukraine, 10003
Державний університет «Житомирська Політехніка»,
Житомир, Житомирська область, Україна,10003

Анотація: *Потреба в науковому дослідженні алгоритму обробки англомовного тексту за допомогою парсерів актуальна для вивчення граматичних структур студентами в англомовних текстах сучасного соціуму. Метою дослідження є обґрунтування методології концептуального підходу до пасерування англомовних текстів. Необхідність вивчення сполучуваності лексичних одиниць зумовлена нерозробленістю широкого кола як теоретичних, так і прикладних проблем. Теоретичні аспекти, які потребують вивчення, – це, зокрема, граматична і лексична валентність слів, типова сполучуваність, синонімія словосполучень різних структурних типів, лексична і граматична валентність як критерій синонімічності, закони комбінаторики словосполучень різних типів і розрядів, лексична валентність як критерій розмежування вільних і фразеологічних словосполучень, взаємодія стійкості й ідіоматичності тощо. До прикладних проблем можна віднести автоматизацію лінгвістичних досліджень, автоматичне визначення меж словосполучень, установлення критеріїв членування фрази на синтагми, автоматичний синтаксичний аналіз речення, автоматичне реферування й анотування тексту на основі сполучувальнісних критеріїв тощо. Об'єктами досліджень є: власне природні мови; використання комп'ютерних програм при аналізі англомовних текстових матеріалів*

Предметом дослідження *сучасні англомовні тексти.*

Об'єктом дослідження *є граматичні структури сучасних англомовних текстів.*

Мета дослідження *− обґрунтування методології концептуального підходу до пасерування текстів.*

Методи дослідження*. Пропоноване дослідження проведене з використанням загальнонаукових методів аналізу, синтезу і узагальнення, а також бібліографічно-описового методу з метою залучення та систематизації англомовних текстових матеріалів.*

Ключові слова: *парсинг, парсер, граббер, синтаксичний аналізатор.*

Парсер або синтаксичний аналізатор – частина програми, яка перетворює вхідні данні, як правило текст структурованого формату. Парсер виконує синтаксичний аналіз тексту. Найчастіше зустрічаються такі типи парсерів: черга класифікованих лексем, абстрактне дерево, ієрархічні структури, таблиці даних[25].

Парсинг (від. *англ. Parse*) – процес розбору певного контенту на складові за допомогою спеціальних програм або скриптів. Наприклад, у SEO цим контентом є html-код сторінок сайтів. **Парсинг** (від англ. Parse) – процес аналізу або розбору певного контенту на складові за допомогою роботів – парсерів (спеціальних програм або скриптів). У SEO цим контентом є html-код сторінок сайтів [28].

Парсинг — це процес збору деяких даних і складання з них бази. Наприклад, можна зібрати базу гостьових книг. Або базу каталогів сайтів. Ось навіщо це робити і робити взагалі — кожен вирішує сам. Найчастіше такі бази використовуються спамерами. Але не обов'язково.

Найвідоміші парсери в мережі Інтернет – це пошукові роботи, які аналізують сторінки, зберігають дані аналізу у себе в базі і потім при пошуку видають релевантні та актуальні документи.

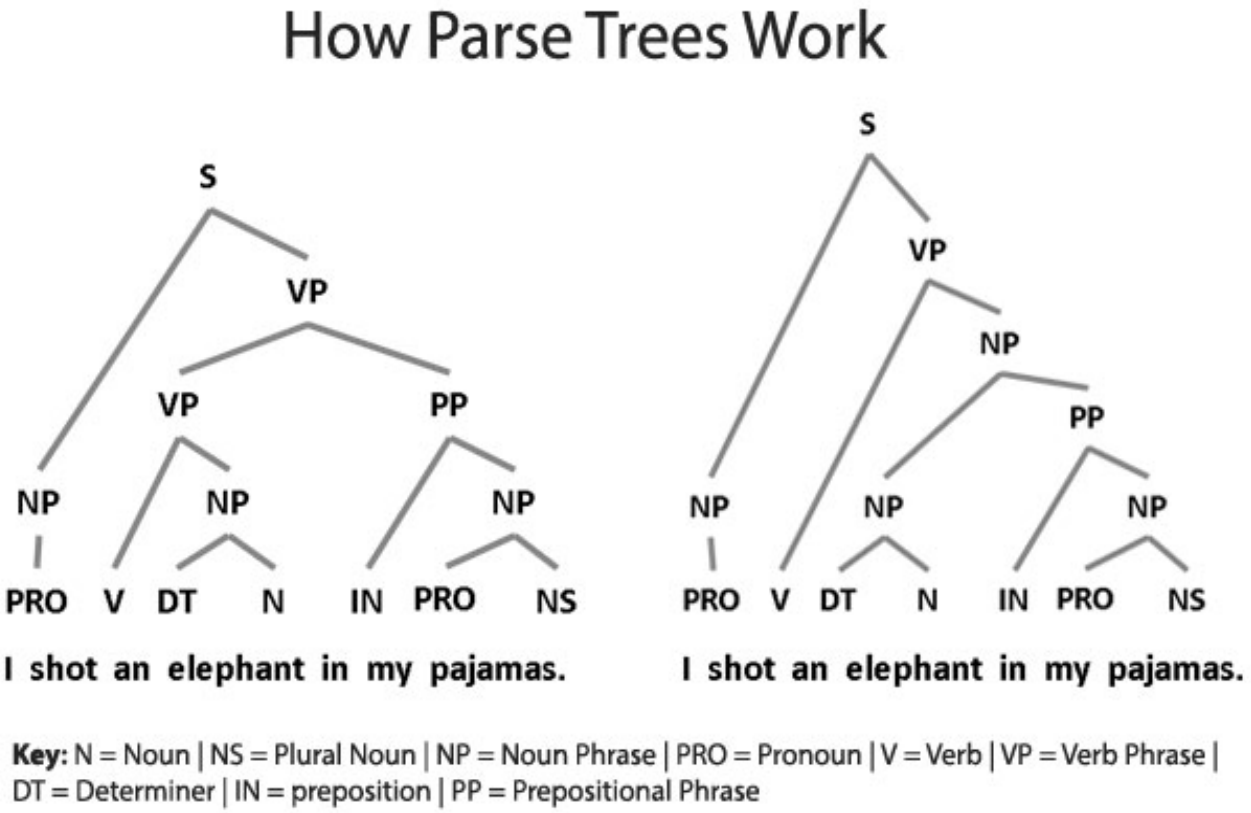
Часто ' *парсинг*' плутають з граббінгом. Це близькі поняття, але все ж мають різні значення.

Граббер – дозволяє скачувати інформацію з мережі (html-сторінки, rss-стрічки, xml-документи) в свою базу, а парсер дозволяє виявити з цієї купи корисну інформацію і обробити її, залежно від поставлених завдань [52].

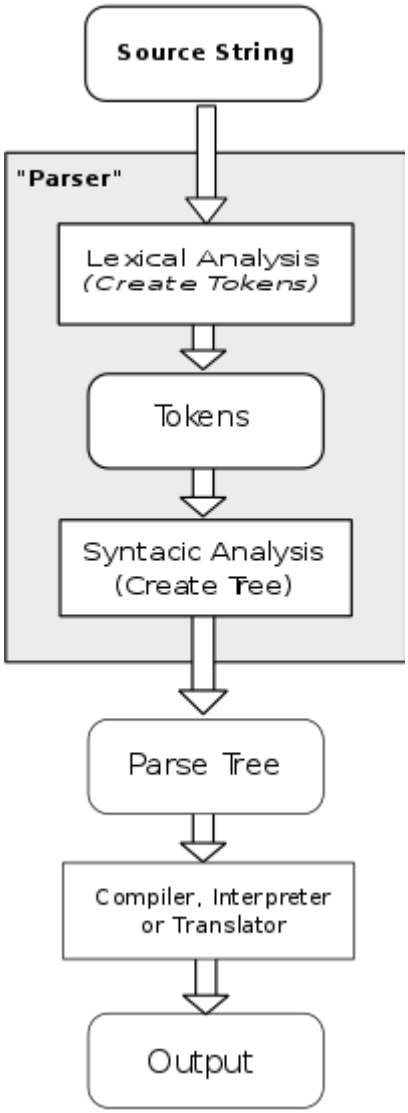
В компютерній лінгвістиці нас цікавить саме синтаксичний аналіз текстів, робота парсерів та яким чином парсери можна використовувати у навчанні іноземних мов. В інформатиці – це процесс аналізу вхідної послідовності символів, з метою розбору граматичної структури згідно із заданими задачами формальної грамматики.

Під час синтаксичного аналізу текст оформлюється у структуру даних, зазвичай — в дерево, яке відповідає синтаксичній структурі вхідної послідовності, і добре підходить для подальшої обробки. Зазвичай синтаксичні аналізатори працюють в два етапи: на першому ідентифікуються осмислені токени (виконується лексичний аналіз), на другому створюється дерево розбору.

Наприклад, дуже відомий випадок автоматичного синтаксичного розбору:



Мал. 1 Зразок відомого синтаксичного розбору речення ' How parse trees work'



Мал.2 Схема роботи синтаксичного аналізатора.

Парсер працює в командному рядку. В якості вхідних даних він отримує текстовий файл у utf-8. Результати роботи зберігаються в форматі XML.

Зразок виклику парсера:

« parser.exe – verbose –tagger 0 –parser 0 –lemmatizer 0 –emit_morph 0 –
dictionary. xml input. txt –o output.txt »

Параметри:

- **verbose** друк інформації про хід парсинга в консоль.
- **tagger 0** для виконання морфологічного аналізу (Part-Of-Speech tagging) використовується базова модель російської або англійської морфології.
- **parser 0** для виконання синтаксичного аналізу (побудова dependency tree) використовується базовий shift – reduce парсер.
- **lemmatizer 0** лематизація виконується з використанням ймовірнісної моделі російської мови і з урахуванням контексту слова.
- **emit_morph 0** не видавати в XML файл результатів списки морфологічних тегів слів; режим 1 приведе до значного зростання обсягу результатів.
- **d dictionary.xml** шлях до конфігураційного файла словникової бази.

- **o output.txt** назва створюваного файлу з результатами.
- **fuzzy_wordrecog 1** включається нечіткий пошук словоформ в лексиконі, який дозволяє редагувати деякі орфографічні помилки і опечатки (за замовчуванням режим виключений) [14].

Якщо вхідний файл містить текст, вже розбитий на речення, так що кожне речення знаходиться на окремому рядку і відокремлено символом '\ n', то можна вказати параметр – eol, і парсер не виконуватиме сегментацію тексту на речення на основі своїх евристик.

Окрім того розрізняють ще режим роботи парсера за допомогою серверів. Парсер спроектований для максимально швидкого розбору великих обсягів текстів (десятки кілобайт або сотні мегабайт). Для отримання максимальної продуктивності парсер при запуску завантажує в оперативну пам'ять всю словникову базу. Завантаження бази займає кілька секунд. У зв'язку з цим нераціонально використовувати парсер для розбору окремих речень.

Для роботи та використання парсера необхідна ліцензія та дистрибутив. В дистрибутив входить всі необхідні словникові бази та моделі ймовірності. За замовчуванням конфігурація парсера запрограмована так, що ним можна користуватися одразу після відкриття, не вносячи змін в параметри. Парсер можна використовувати не лише в лінгвістичних цілях, так як він містить інформацію необмежену для використання.

Парсер має як позитивні аспекти використання так і негативні. Дуже часто при використанні парсера, він не приймає вхідний текст, тоді необхідно ввести додаткову інформацію, яка більш точно описує проблему.

Парсер генерує інформацію про помилку, яка описує позицію, чому парсер не зміг продовжити синтаксичний аналіз. Якщо було багато помилок, то парсер дає повідомлення про останню. Повідомлення про помилки, можуть виявитися бажаними при парсингу альтернатив. [68,330с.]

def value: Parser[Any] = numerclit| “true” | “false”

Наприклад, якщо дано правило і парсер не зміг знайти жодної альтернативи, тоді повідомлення “false” вводить в оману. В данній ситуації, можна надати додаткову інформацію failure з текстом про помилку :

def value: Parser[Any] = numerclit| “true” | “false” |
failure (“Not a valid value”)

Якщо парсер зазнає невдачі, метод parselAll поверне результат типу Failure, він повертатиме користовачу інформацію з помилкою.

Література

1. Борисова Н. В., Кочуєва З. А., Оліфенко І. В. Метод автоматизованої лематизації дієслів німецької мови. – [Електронний ресурс] /Н.В. Борисова - Режим доступу: http://vlp.com.ua/node/16722

2. Буніятова І. Р. Еволюція гіпотаксису в германських мовах (IV – XIII ст.) : монографія / І.Р. Буніятова. – К. : Вид. центр КНЛУ, 2003. – 327 с.

3. Гирин О.В. Автоматичний синтаксичний аналіз англійської мови: застосування та перспективи – [Електронний ресурс] / О.В.Гирин – Режим доступу: http://nniif.org.ua/File/17govasa.pdf

4. Гузеева К. А. Инфинитив / К. А. Гузеева, С. И. Костыгина //Грамматика английского языка. – С-Пб : Союз. 2000. – 219 с.

5. Дарчук Н. П. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н. П. Дарчук //Українське мовознавство. – КНУ ім. Т. Шевченка, 2013. – № 43. – С. 11–19.

6. Загнітко А. П. Основи українського теоретичного синтаксису. Частина 1 / А. П. Загнітко. – Горлівка, 2004.

7. Енциклопедія сучасної україни – [Електнонний ресурс]/ режим доступу: http://esu.com.ua/search_articles.php?id=4396

8. Карамишева І. Д. Структурні та функціональні особливості вторинної предикації в сучасній англійській мові (досвід формально-граматичного моделювання) : дис. … канд. філол. наук : спец. 10.02.04 «Германські мови» / Карамишева Ірина Дамірівна. – Київ, 2005. – 320 с.

9. Комп'ютерна лінгвістика – [Електронний ресурс]/ режим доступу: http://computernalingvistska.blogspot.com

10. Комп'ютерна лінгвістика - [Електронний ресурс]/ режим доступу: http://kikref.ru/iref_589382.html

11. Сучасна українська літературна мова. Морфологія. Синаксис. – К., 2010. 6. Русская граматика: в 2 т. Т. 2. Синтаксис. – М., 1980. – С. 21.

12. Чейлитко Н. Г. Корпусне дослідження зон зв'язків словоформ в українському реченні / Н. Г. Чейлитко // Лінгвістичні студії : [зб. наук. праць]. – Донецьк, Вид-во ДонНУ, 2009. – Вип. 18. – С. 268–275.

13. Методи вивчення змісту медіа – комунікації – [Електронний ресурс]/ режим доступу: http://www.gpedia.com/uk/gpedia/Методи_вивчення_змісту_медіа-комунікацій

14. Моховик В.В. Реферат - [Електронний ресурс]/ В.В.Моховик/ режим доступу: http://masters.donntu.org/2014/fknt/mokhovykh/diss/indexu.htm

15. Напрямки комп'ютерної лінгвістики/ режим доступу: https://otherreferats.allbest.ru/programming/00397420_0.html

16. Основні методи обробки природної мови – [Електронний ресурс]/ режим доступу: http://studall.info/all2-40857.html

17. Полховська М. В. Аналіз англійських медіальних конструкцій з позиції генеративної граматики / М. В. Полховська // Studia philologica. – 2013. – Вип. 2. – С. 32-36.

18. Полховська М. В. Критерії розрізнення медіальних та ергативних конструкцій в англійській мові / М. В. Полховська // Наукові записки [Національного університету "Острозька академія"]. Сер. : Філологічна. – 2012. – Вип. 26.– С. 277-280.

19. Проблеми компютерної лінгвістики – [Електронний ресурс]/ режим доступу: http://kumlk.kpi.ua/node/844

20. Раєвська М. Н. Present-day English Syntax / М. Н. Раєвська //Синтаксис сучасної англійської мови : підруч. – К. : Вища школа,1970. – 179 с.

21. Снісаренко І. Є. Інфінітивна конструкція з прийменником *for* у середньоанглійській мові : семантика та функціонування : дис …канд. філол. наук : 10.02.15 / Снісаренко Ірина Євгенівна // КНЛУ. – К., 2001. – 190 с.

22. Снісаренко І.Є. Позиційні характеристики інфінітивного ад'юнкта мети (на матеріали пам'яток середньоанглійського періоду) / І.Є. Снісаренко // Вісник Харківського національного університету ім. В. Н. Каразіна. – Харків : Видавництво ХНУ ім. В. Н. Каразіна. – № 636. – 2004. – С.186–187.

23. Українське мовознавство – [Електронний ресурс]/ режим доступу: http://philology.knu.ua/files/library/ukrmov/45/45.pdf

24. Чейлитко Н. Г. Корпусне дослідження зон зв'язків словоформ в українському реченні / Н. Г. Чейлитко // Лінгвістичні студії : [зб. наук. праць]. – Донецьк, Вид-во ДонНУ, 2009. – Вип. 18. – С. 268–275.

25. Стемінг – [Електронний ресурс]/ режим доступу: https://lookup-api.apple.com/uk.wikipedia.org/wiki/Стемінг

Abstract: *The need for a scientific study of the algorithm for processing English-language text using parsers is relevant for the study of grammatical structures by students in English-language texts of modern society. The purpose of the study is to justify the methodology of the conceptual approach to the passage of English-language texts. The need to study the conjugation of lexical units is due to the lack of development of a wide range of both theoretical and applied problems. Theoretical aspects that need to be studied are, in particular, the grammatical and lexical valence of words, typical conjunctiveness, synonymy of phrases of different structural types, lexical and grammatical valence as a criterion of synonymy, laws of combinatorics of phrases of various types and degrees, lexical valence as a criterion for distinguishing between free and of phraseological phrases, interaction of stability and idiomaticity, etc. Applied problems include the automation of linguistic research, the automatic determination of phrase boundaries, the establishment of criteria for dividing a phrase into syntagms, automatic syntactic analysis of a sentence, automatic abstracting and annotation of text based on conjunctive criteria, etc. The objects of research are: actual natural languages; the use of computer programs in the analysis of English-language textual materials.*

Key words: *parsing, grabber,parser,syntax analizator.*