3-Модуль. Корпустарды қолдану

Дәріс №11
Дәріс тақырыбы: Корпус менеджерлері
Сұрақтар:

1. Корпус іздеу жүйесі ретінде
2. Сұрау салу (сауал/сұрау) тілдері

1. Корпусты зерттеу процесі үш негізгі кезеңді қамтиды: корпусты құрастыру, аннотациялау және іздеу және алып шығу (аудармада іздеу деп аударғанымен, жай ғана іздеу емес, алып шығу) (Rayson 2008 қараңыз) - corpus compilation, annotation, and retrieval (see Rayson 2008). Корпус алдымен онлайн көздерден транскрипция, сканерлеу немесе іріктеу арқылы құрастырылуы керек. Содан кейін, екінші кезең мәтіндік және тілдік сипаттарды анықтайтын тегтерді, кодтарды және құжаттаманы қосу үшін қолмен және автоматты әдістердің кейбір комбинациясы арқылы аннотация болып табылады. Корпусты зерттеу процесінің бірінші және екінші кезеңдерін қолдайтын құралдар мен әдістер үшінші кезеңде көрініс табады. - The corpus research process involves three main stages: corpus compilation, annotation, and retrieval (see Rayson 2008). A corpus first needs to be compiled via transcription, scanning, or sampling from online sources. Then, the second stage is annotation, through some combination of manual and automatic methods to add tags, codes, and documentation that identify textual and linguistic characteristics. A snapshot of tools and methods that support the first and second stages of the corpus research process are described in Sections 2.1 and 2.2.

retrieval **the process of finding and bringing back something: https://dictionary.cambridge.org/dictionary/english/retrieval**

Міне осы үшінші кезең – алып шығу жүйесі - корпус менеджерімен тығыз байланысты ұғым. Корпус менеджері – корпусқа енгізілген мәтіндерді және сол мәтіндерге енгізілген деректерді басқару жүйесі. «Корпусный менеджер – это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме». (Захаров, Богданова). - Корпус менеджері арнайы іздеу жүйесі, ол іздеу белгілі бір бағдарламалар арқылы іске асады, сол арқылы статистикалық ақпарат алынады, корпусты пайдаланушы өзіне қажетті нәтижелерді ыңғайлы формада алуға мүмкіндігі болады.

Корпус менеджерінің істеуге міндетті әрекеттерін В.Захаров пен С.Богданова атап көрсетеді (қараңыздар).

Корпустан лингвистикалық ақпаратты алып шығу арнайы компьютерлік бағдарламалар көмегімен жүзеге асады. Ол бағдарламалардың біразы ақылы, біразы тегін. Корпустан белгілі бір ақпарат алынады, ол алынатын ақпараттың түрлері көп зерттеулерде айтылатын түрлері – простые списки (Кутузовта, «Самый простой формат отображения информации о корпусе — это простые списки. Эти списки могут быть разных типов — от простых глоссариев до конкордансов.), ағылшынша – key word lists. Алып шығу құралдары мен әдістері корпустарға негізделген линвгистикалық зерттеулер жүргізуге мүмкіндік береді. Олар мысалы, жиілік талдауы, конкорданстар, коллакациялар, кілт сөздері және n - граммдар - Retrieval tools and methods enable the actual linguistic investigations based on corpora: i.e. frequency analysis, concordances, collocations, keywords and n-grams.

Конкорданс әдісімен қатар корпус пайдаланушысының жұмысында негізгі ретінде тағы төрт әдіс пайда болды: жиілік тізімдері, кілт сөздер, n-граммдар және коллокациялар.- Alongside the concordance method, a further four methods have emerged as central to the work of the corpus user: frequency lists, keywords, n-grams, and collocations. (The Cambridge handbook of English corpus linguistics / edited by Douglas Biber and Randi Reppen Cambridge

University Press 2015 Tapay - Computational tools and methods for corpus compilation and analysis Paul Rayson)

Бұл әдістердің ішінде бастысы – конкорданс болып табылады, ол корпустан алынған және контексте көрсетілген белгілі бір тілдік ерекшеліктің барлық мысалдарын көрсететін, әдетте әр жолға бір мысал ретінде берілген, мысалдың сол және оң жағындағы қоршалған мәтіннің қысқаша бөлігімен сәйкестік болып табылады.- Central amongst these methods is the concordance, which displays all examples of a particular linguistic feature retrieved from the corpus and displayed in context, usually presented as one example per line, with a short section of surrounding text to the left and right of the example itself as shown in Figure 2.1.

Барлық конкорданс құралдары қарапайым сөз арқылы іздеуді қамтамасыз етеді, ал кейбір құралдар жұрнақтарды, бірнеше сөз тіркестерін, тұрақты тіркестерді, сөз бөлігінің тегтерін, корпусқа енгізілген басқа аннотацияларды немесе күрделі контекстік үлгілерді іздеуге мүмкіндік береді. - All concordance tools provide for searching by a simple word and some tools permit searching for suffixes, multiple word phrases, regular expressions, part-of-speech tags, other annotation embedded within the corpus, or more complex contextual patterns.

> Undoubtedly the single most important tool available to the corpus linguist is the concordancer. A concordancer allows us to search a corpus and retrieve from it a specific sequence of characters of any length – perhaps a word, part of a word, or a phrase. This is then displayed, typically in one-example-perline format, as an output where the context before and after each example can be clearly seen.
>
> We have avoided saying that a concordance shows *words* in their context – though this procedure is often called *key word in context* (KWIC) concordancing, KWIC need not be limited just to showing whole words. For example, in English we might want to produce a concordance of a common suffix in order to explore its context of use – Figure 2.3 shows an example of this, a concordance of words ending in the nominalising suffix *-ness*.

Соның қатарында аталатын корпус менеджері KWIC-ті де (Key Word In Context) және де толық конкорданс тізімдерін жасауы қажет. Контекстегі кілт сөз (KWIC) сәйкестік жолдарының ең көп таралған форматы болып табылады. KWIC терминін алғаш рет Ханс Питер Лун енгізген. Жүйе 1864 жылы Андреа Крестадороның Манчестер кітапханаларына алғаш ұсынған *атаулардағы кілт сөз* тұжырымдамасына негізделген. KWIC индексі мақаланың тақырыбындағы сөздерді сұрыптау және туралау арқылы жасалады, осылайша тақырыптардағы әр сөзді индексте алфавиттік ретпен іздеуге болады. (таныстыру таратпа материал LancsBox_3.0_KWIC.pdf)

> **KWiCFinder** A web-based system (a personal internet search agent) developed by William H. Fletcher of the United States Naval Academy that allows one to conduct **concordance** searches across the **World Wide Web**. Using the web as a corpus enables users to access very large, unordered, text collections in a range of languages. The system is Windows based and requires the use of Internet Explorer. The service is available free of charge at http://www.kwicfinder.com.

Мынадай белгілі әмбебап корпус менеджерлері бар: SARA, XAIRA (BNC), Manatee/Bonito, CQP, DDC. Корпус деректерін өңдеу үшін дерекқорды басқару жүйелері (систем управления базами данных - СУБД) немесе іздеу жүйелері негізінде менеджерлер әзірленуі мүмкін. Оған мысалы орыс тілінің ұлттық корпусының іздеу жүйесі. Мұнда іздеу Яndex.Server 3.8 Professional іздеу жүйесі арқылы іске асады.

The historical timeline of **corpus retrieval software** can be divided into **four generations**. **In the first generation** that developed alongside machine-readable corpora, software tools running on large mainframe computers simply provided concordance or key-word-in-context (KWIC) displays, and separate tools were created in order to prepare frequency lists, e.g. as used by Hofland and Johansson (1982). These tools were usually tied to a specific corpus. **In the second generation**, applications such as the Longman Mini-Concordancer, Micro-Concord, Wordcruncher, and OCP were developed to run on desktop computers. These were capable of dealing with multiple corpora and extra features to sort concordance lines by left and right context were added. Increased processing capabilities of PCs and laptops in the 1990s led to **the third generation** of retrieval software with systems such as WordSmith, MonoConc, AntConc, and Xaira being developed. They were able to deal with corpora of the order of tens of millions of words, containing languages other than English, and they included implementations of the other methods outlined above in one package rather than as separate tools. **The fourth generation** of corpus retrieval software has moved to web-based interfaces. This allows developers to exploit much more powerful server machines and database indexes, provide a userfriendly web interface and host corpora that cannot otherwise be distributed for copyright reasons. Most of the web-based interfaces only permit access to pre-existing corpora rather than texts that the users collect themselves. For example, Mark Davies' corpus.byu.edu interface permits access to very large corpora: 450 million words of the Corpus of Contemporary American English (COCA), 400 million words of the Corpus of Historical American English (COHA), 100 million words of the Time Magazine Corpus, and 100 million words of the British National Corpus. Other systems tend to rely on the Corpus Query Processor (CQP) server (part of the Open Corpus Workbench) or Manatee server. Their well-known web-facing front ends are BNCweb (providing access to the British National Corpus), SketchEngine (aimed at lexicographers), and CQPweb (based on the BNCweb design but suitable for use with other corpora). Other web-based tools in a similar mold are Intellitext (aimed at humanities scholars), Netspeak, and ANNIS. The web-based Wmatrix software (Rayson 2008) allows the user to perform retrieval operations but it also annotates uploaded English texts with two levels of corpus annotation: part-of-speech and semantic field. For further information on the four generations of corpus retrieval tools, a good survey can be found in McEnery and Hardie (2012: 37–48).

2. Сауал тілдері. Ақпараттық сұрау - бұл белгілі бір ақпарат қажеттілігінің ауызша көрінісі. Сұраулар пән және формальды мазмұны бойынша талданады және корпуспен жұмыс істейтін қолданбалы бағдарламаның сұраныс тілінің сөздік құрамы тұрғысынан сипатталады. Сұрау салуды сіздер сабаққа дайыналу барысында немесе басқа да мақсаттарда жиі пайдаланасыздар. Мысалы гуглдің сұрау салу жүйесі. Жалпы, сұрау тілі үлгісі келесі элементтерді қамтиды: 7 (оқыңыз)

Кітаптарыңда **Bonito/Manatee** сауал тілі жан жақты берілген. Manatee - корпус менеджері. Bonito - бұл Manatee корпус менеджерінің графикалық пайдаланушылық интерфейсі (GUI). Бұл жүйені Чехияда Масарика атындағы университеттің информатика факультетіндегі NLPlab (Natural Language Processing Laboratory) лабораториясы мен П. Рыхли жасаған.

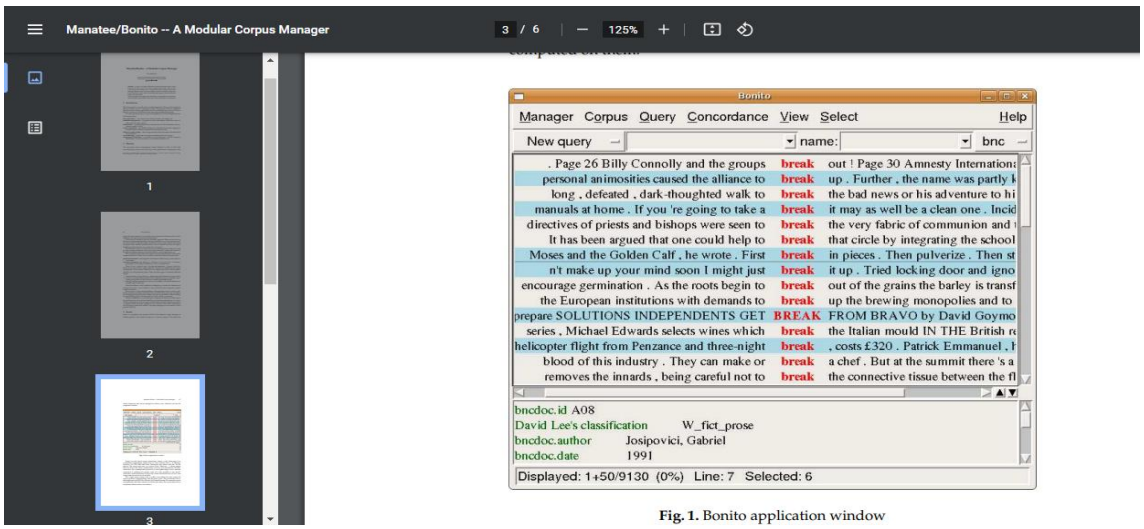

Fig. 1. Bonito application window

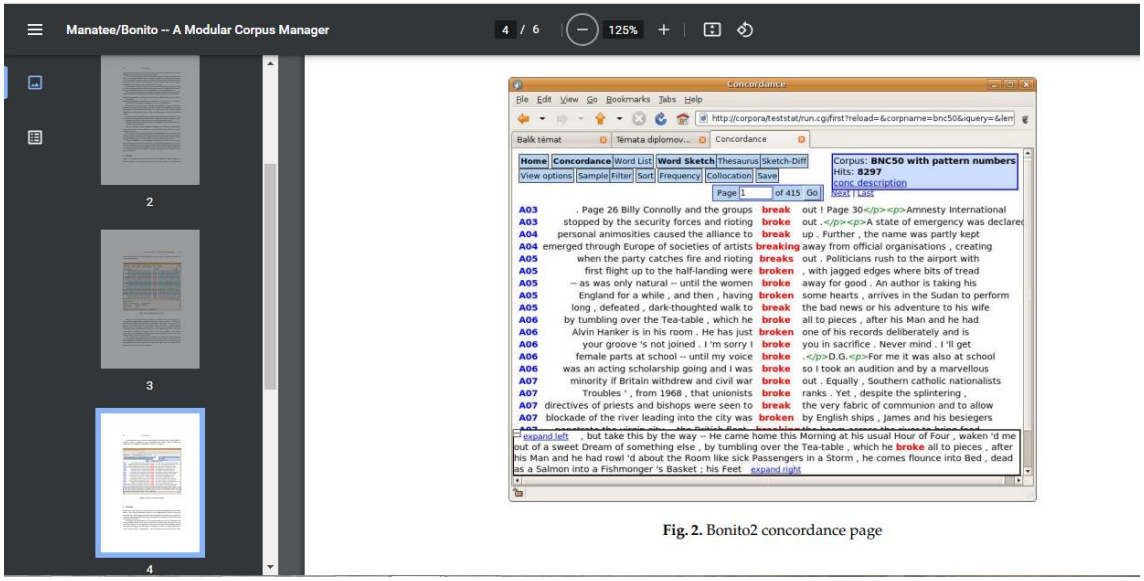

Fig. 2. Bonito2 concordance page

Мен сіздерге Скетч енжинді таныстырайын.

Sketch Engine is a corpus manager and analysis software has developed by Lexical Computing since 2003. This software consists of three main components which enable to search and build text corpora.

- **Bonito** – a graphical user interface to corpora maintained, see the changelog of Bonito
- **Manatee** – a corpus management tool including corpus building and indexing, fast querying and providing basic statistical measures
- **FinLib** – fast indexing library, see the changelog of FinLib

https://www.sketchengine.eu/documentation/manatee-changelog/

Bonito жүйесінің негізгі ерекшеліктері (кітаптан); Сауалдар (кітапта), сауал типтері (кітапта), шаблондар (кітапта); сауал мысалдары 7 (кітапта)