1-Модуль. Корпус лингвистикасына кіріспе

Дәріс №1
Дәріс тақырыбы: Корпус лингвистикасы туралы
Сұрақтар:

1. Корпус лингвистикасы пән ретінде
2. Корпус лингвистикасы дегеніміз не?

1. Корпус лингвистикасының пән ретінде қалыптасуының белгілі бір алғышарттары бар. Бастапқыда компьютер лингвистикасы пәнінің шеңберінде қалыптасып, кейіннен табиғи тілді машинада өңдеудің кейбір сұрақтары мен мәселелері бөлініп корпус лингвистикасы пәнінің аясына өтті.

Корпус лингвистикасы ауызша немесе жазбаша тілдің үлгілерін жинау арқылы табиғи пайда болатын тілдік құрылымды және қолдануды талдайтын зерттеу саласын айтады. В.Захаров пен С.Богданова 2011 жылғы «Корпусная лингвистика» (Учебник для студентов направления «Лингвистика») оқулығында корпус лингвистикасын былайша анықтаған: «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий» (7). Компьютерлік лингвистиканың бөлімі ме? Оқулық авторларының ойынша, КЛ-сы оны жеке пән ретінде танудың кем дегенде екі қырына ие дейді: «1) характер используемого словесного материала; 2) специфика инструментария ().

Корпус лингвистикасының басқа пәндерден ерекшелігі сонда, егер семантика, синтаксис немесе когнитивті лингвистика дербес пәндер ретінде даусыз танылса, КЛ - өте кең ұғым. Мысалы, А.Захаров пен С.Богданова КЛ «тілдік зерттеулердің көптеген аспектілеріне қолдануға болатын әдіснама болып табылады» дейді. Кейбір зерттеулерде ол лингвистикалық талдаулардың әдісі деп те сипатталады. Вольфганг Теуберт/Teubert (2005) КЛ когнитивті лингвистикамен салыстырады: «мен үшін, - дейді ғалым, - корпус лингвистикасы және когнитивтік лингвистика бір-бірін толықтыратын, бірақ сайып келгенде үйлеспейтін екі парадигма. Когнитивті Л сөйлеушілер мен тыңдаушылардың ойларында не бар соған қызығушылық танытады. Біз ойымызды сөйлеуге қалай айналдырамыз? Біз естігенімізді немесе оқығанымызды ойға қалай көшіреміз? т.т. Мұның бәрінің корпус лингвистикасына қатысы жоқ. КЛ тілге психологиялық емес, әлеуметтік перспективадан қарайды. Вербалды өзараәрекеттесу адамзаттың әлеуметтік топтарына біздің ең жақын туыстарымыздан - маймылдар тобынан да - күрделірек болуына мүмкіндік береді. Маймылдар не істесе де, олар мазмұнды жеткізіп жатқанын білмейді және олар жеткізіп жатқан мазмұнға ұжымдық түрде ойлана алмайды. Адамзат мазмұнды ұжымдық түрде келісе алатындығымен ерекше.

> For me, corpus linguistics and cognitive linguistics are two complementary, but ultimately irreconcilable paradigms. Cognitive linguistics is interested in the minds of speakers and hearers. How do we turn our thoughts into speech? How do we transform what we hear or read into thoughts? If we are all born with the same language faculty, how can we draw the line between what is common to all languages, and what is specific to individual languages? To what extent is our linguistic performance governed by universal laws underlying our language faculty, and to what extent is it governed by the cultural conventions of the language community in which we have grown up?
> All this does not concern corpus linguistics. Corpus linguistics looks at language not from a psychological, but from a social perspective. Verbal interaction is what allows human social groups to be infinitely more complex than even groups of the apes, our closest relatives. Whatever apes do, they are not aware that they are conveying content and they cannot collectively reflect on

the content they are conveying. Humans are unique in that they can can negotiate content collectively. (My version of corpus linguistics Wolfgang Teubert)

Ендеше КЛ дегеніміз не осыны жақсылап ұғынып алайық.

2. Корпус лингвистикасы дегеніміз не? Лингвистер арасында корпус линвгистикасы дегеніміз не немесе не болуы керек деген сұрақ көп талқыланған. Корпус лингвистикасы - бұл құрал, әдіс, әдістеме, әдістанымдық тәсіл, пән, теория, теориялық көзқарас, парадигма (теориялық немесе әдіснамалық) немесе осылардың жиынтығы теория ма, модель ме, немесе әдіс пе, не ол? Осы сұраққа бірізді жауап таба алмайсыз. Зерттеуші Шарлотта Тейлор: «Бұл сұраққа жауаптардың, корпус линвгистикасына берілген анықтамалар мен сипаттамалардың әртүрлілігі таңқаларлық», - дейді. «Корпус лингвистикасы дегеніміз не?» деген мақаласында Шарлотта Тейлор «бұл пән бе, әдістану ма (методология) ма, парадигма ма, осылардың ешқайсысы емес пе немесе осылардың барлығы ма?» деген сұраққа жауап іздейді. Бірақ анық жауаптар жоқ дейді. (Charlotte Taylor «What is corpus linguistics?" – is it a discipline, a methodology, a paradigm or none or all of these? – but does not attempt to offer any definitive answers). (2008, 179 What is corpus linguistics? What the data says ICAME Journal No. 32 2008 179-200)

Корпус лингвистикасының негізін салушылардың бірі Ян Аартс, Мэйдспен бірігіп жазған 1984 жылғы кітабында (Aarts and Meijs (1984), оның алдында 1982 жылы шыққан (Аартс и ван ден Хеувель /Aarts and van den Heuvel (1982) ван ден Хеувельмен бірігіп жазған зерттеулерінде корпус лингвистикасы терминін қолданған. Аартс бұл терминнің біраз дүдәмалдықпен туындағанын айтады, өйткені «бұл өте жақсы атау емес деп ойладық (және әлі де ойлаймын): бұл оның басты зерттеу құралымен және деректеркөзімен аталатын таңқаларлық пән (странная дисциплина). Мүмкін, бұл термин қазір өзінің пайдалылығынан айырылған шығар/Возможно, этот термин к настоящему времени изжил себя». Corpora Jul 1998 to -: Corpora: First use of the term 'corpus linguistics' (uib.no) Бұл айтылған ой корпус лингвистикасы туралы жиі айтылатын мәселелердің бірін тудырады және баламаларға басымдық беруі мүмкін.

Джеффри Лич «Directions in Corpus Linguistics, 'Corpora and theories of linguistic performance' (Mouton de Gruyter, 1992, p. 105) деген зерттеуінде (кітап тарауы болуы керек) былай деген: «Корпус лингвистикасы» термині Ян Аартс пен У.Мейстің редакторлығымен «Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research» (1984) атты кітабы шыққанға дейін анда-санда ғана пайда болған». Ян Аартстың айтуынша, терминді 1980 жылы зерттеу бағдарламасының атауы ретінде пайладаланған, алайда сонда ойланып (though with some hesitation) қолданғанын айтады.

On the Corpora List, Aarts is reported as commenting that the term was coined with some hesitation "because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source. Perhaps the term has outlived its usefulness by now". This raises one of the recurrent concerns over talking about corpus linguistics, and may account for the preference for alternatives.

In response to my query, Jan Aarts wrote that, to his knowledge, their book was the first entirely devoted to the subject but that he had used the term in 1980 as the title of a research programme, though with some hesitation "...because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source." He added that "Perhaps the term has outlived its usefulness by now. Its boundaries have shifted and have become rather vague - nowadays you hardly know whether you are talking about a subdiscipline of linguistics or about

language technology. The TOSCA [TOols For Syntactic Corpus Analysis] research programme, called 'corpus linguistics' for over 15 years, is now called 'Empirical description of language use'." http://korpus.uib.no/icame/corpora/1998-3/0006.html

Aarts, Jan and Willem Meijs (eds.). 1984. Corpus linguistics: Recent developments in the use of computer corpora in English language research. Amsterdam: Rodopi.

Aarts, Jan and Theo van den Heuvel. 1982. Grammars and intuitions in corpus linguistics. In S. Johansson (ed.). Computer corpora in English language research, 66–84. Bergen: Norwegian Computing Centre for Humanities.

Aarts, Jan. 2002. Does corpus linguistics exist? Some old and new issues. In L. E. Breivik and A. Hasselgren (eds.). From the COLT's mouth… and others': Language corpora studies in honour of Anna-Brita Stenström, 1–19. Amsterdam: Rodopi.

1992 жылы Лийч «компьютерлік корпус лингвистикасы тілді үйренудің жаңадан пайда болған әдістанымын ғана емес, сонымен қатар жаңа ғылыми-зерттеу кәсіпорнын және іс жүзінде бұл пәнге жаңа философиялық көзқарасты анықтайды» деп тұжырымдап, компьютерлік корпус лингвистикасының сипаттамаларын жаңа парадигма ретінде сипаттаған (Leech 1992: 106). Яғни Лийч компьютерлік корпус лингвистикасын жаңа философиялық көзқарас деп анықтаған. (Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Startvik (Ed.), Directions in Corpus Linguistics (pp. 105-122). Berlin: Mouton de Gruyter. (ed.). 105–122).

Стаббс (1993) корпус лингвистикасының әдістану ретіндегі анықтамасын жоққа шығарады және Синклердің 1991 ж. зерттеуіне пікір білдіре отырып, «пәннің осы көзқарасы бойынша корпус лингвистикалық талдау құралы ғана емес, лингвистикалық теориядағы маңызды ұғым» деп атап өткен (1993: 23–24). (Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair. In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). Text and technology: In honour of John Sinclair, 1–36. Amsterdam: John Benjamins.)

Вольфганг Теуберт (2005) те теориялық концептуализацияға баса назар аударады және корпус лингвистикасын «тілді зерттеудегі теориялық тәсілдеме» («corpus linguistics as a theoretical approach to the study of language») деп сипаттайды (2005: 2). (Teubert, Wolfgang. 2005. My version of corpus linguistics. International Journal of Corpus Linguistics 10(1): 1–13.). Ғалым КЛ 25 тезиспен сипаттап шығады (Corpus linguistics in 25 theses: 2-8).

Corpus linguistics in 25 theses:

1. The focus of corpus linguistics is on meaning. Meaning is what is being verbally communicated between the members of a discourse community. Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language. It claims no privileged knowledge of the workings of the mind or of an innate language faculty.	14. As with idioms, we can describe a complex lexical item holistically as a semantic unit whose meaning cannot be inferred from decomposing it into the smaller lexical items it consists of. It is, however, a matter of degree to what extent the meaning of a complex lexical item is independent of the meaning of the parts it is composed of. Sometimes it can be useful to describe a complex lexical item by assuming that its node is imbued with certain semantic (usually connotative) features inherent in the other elements that the complex lexical item consists of. This approach turns

	our attention to the phenomenon called semantic prosody (for connotative features) or semantic preference (for denotational features.)
2. In corpus linguistics, language study is always the study of written (or transcribed or quoted or otherwise recorded) texts or text pieces, i.e. language which can be reproduced, heard, read and interpreted repeatedly. What is not written or transcribed or quoted or recorded is lost, both for the discourse community and for linguistic investigation. The question of what is spoken and therefore transient, and what is written and therefore permanent, is rather a matter of perspective than of linguistic 'reality'.	15. Lexical items can be single words or complex units of meaning. They are, in principle, monosemous. This is what distinguishes the concept of a lexical item from the concept of a word. Most words, particulary the more frequent ones, are polysemous. Complex lexical items can be seen as a node word together with all those words in its context with which it forms a se□ mantic unit. As long as this unit is still ambiguous it is not yet complete, and more elements have to be added. It is complete as soon as it has, as a lexical item type, only one meaning. Once we replace the concept of the polysemous single word by the concept of the monosemous lexical item, the problem of ambiguity that has aggravated many linguists for a long time suddenly disappears.
3. Every text segment, word, multi-word unit, phrase etc., can be viewed under the aspect of form and the aspect of meaning. The form is what represents the meaning, and there is no meaning without the form by which it is represented. Text segments are symbolic; they always mean something to someone. Normally the members of the language community deal with text segments without being aware what these text segments mean, just as people are often jealous without being conscious of their jealousy. Unless there is some (potential) communication disorder, there is no need to discuss meaning within the discourse community.	16. If the same (monosemous) lexical item recurs in a discourse, then each occurrence is one instantiation of the same lexical item type. Each instance can thus be seen as a token of the type constituted by this lexical item.
4. Meaning is in the discourse. Once we ask what a text segment means, we will find the answer only in the discourse, in past text segments which help to interpret this segment, or in new contributions which respond to our question. Meaning does not concern the world outside the discourse. There	17. Many lexical item types often allow for a certain degree of variation within their instantiations. This variation often has no effect on meaning. If variation affects meaning, we should talk rather about a set of related lexical item types each of which has its own instantiations.

is no direct link between the discourse and the 'real world'. It is up to each individual to connect the text segment to their first-person experiences, i.e. to some discourse-external ideation or to the 'real world'. How such a connection works is outside the realm of the corpus linguist.	
5. For corpus linguistics, the meaning of a text or of a text segment is independent of the intentions of its speaker (its author). The dislocation of the speaker/author from his or her text distinguishes written (recorded) language from spoken language. In spoken language, the speaker is usually present, and if there is a communication disorder, we ask: "What do you mean?" and not: "What does this mean?".	18. To posit a lexical item is to make a general claim. However, corpus linguistics is also concerned with specific claims. Any text segment can be viewed in two ways: as an instantiation of one or more potential lexical item types, or as a unique occurrence whose meaning has to be interpreted through the intertextual clues which connect it to other texts of the discourse. Frequency is irrelevant when our goal is to interpret text segments as unique occurrences. This is the point at which the diachronic perspective takes over from the synchronic view.
6. Corpus linguistics is empirical. Its object is real language data. The discourse is the totality of all the texts that have been produced within a discourse community. Only those of which we have records (in form of written texts or transcripts) can be the object of linguistic investigation. However, just like the discourse community, the discourse is not an ontological reality; it is a construct, the object of research constructed by the linguist. The linguist's task is to define and delimit his or her object of research, to specify which language data he or she wants to analyse. Delimiters include linguistic, spatial, temporal, social, topical and medial parameters.	19. Meaning is paraphrase. Whenever lexical item tokens are the cause of a communication disorder, their meaning will be negotiated, described or explained, replaced by synonyms, and sometimes even 'defined' as in dictionaries or in encyclopaedias. What we find paraphrased in the discourse are text segments that are looked upon intuitively by the members of the discourse community as units of meaning. However, the same lexical item type can be paraphrased in an infinite variety of ways. Therefore, whenever a lexical item token is being paraphrased, we can view it from two perspectives: as an instantiation of the lexical item type, and as a unique occurrence. From a synchronic perspective, the meaning of a lexical item type is a generalisation on all the paraphrases we find for the instantiations of this lexical item. But paraphrases are also relevant for specific claims. From a diachronic perspective, it is the history of paraphrases of a recurrent text segment, as evidenced in its intertextual links, that tells us what it means as a unique occurrence.

7. Corpus linguistics makes general and specific claims about the discourse, based on the analysis of a suitably selected cross-section of it, i.e. the corpus. General claims have to do with rules or with probabilistic expectations. They fall within the field of grammar or variation or language change, and also into the field of lexical meaning insofar as a text segment occurring in a text can be viewed as an instantiation (a token) of a lexical item. Specific claims are interpretations of texts or text segments viewed as unique occurrences.	20. There is no true and no fixed meaning. Everyone can paraphrase a unit of meaning however they like, therefore the meaning of any lexical item type is always provisional. The next paraphrase may already lead to a revision. The members of the discourse community will continue to negotiate, among themselves, what a unit of meaning means. They may agree or not: the issue is not truth, but acceptance. An explanation, a paraphrase that is widely accepted and re-used, is more relevant than a paraphrase that is never repeated, just as texts which are constantly referred to are more relevant than texts that leave no traces in subsequent texts.
8. Each discourse has, by necessity, a diachronic dimension. What is said today is a reaction to what has been said before, an argument in a simultaneous debate and an anticipation of what we expect to be said tomorrow. If we look at language from this perspective, we want to make a specific claim. We want to know what makes a given text segment a unique occurrence rather than a token of a lexical item type. This will be determined by the unique position it maintains in the discourse as a whole, embedded in a context that is unique, and referring to a unique set of other texts. Unless we find the intertextual clues that relate this text segment to previous texts, and to relevant contemporaneous texts, we do not know what makes it unique. However, we can only find these intertextual links if our corpus has a diachronic structure.	21. The discourse is a self-referential system. Natural language is the only codification system in which the functions of its elements are determined not by ascription from outside but by discourse-internal negotiation. This sets natural languages apart from formal calculi, like the code of mathematics.
9. While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach.	22. When we speak, we do not refer to a discourse-external reality but to what has been said before. When we negotiate the meaning of a text segment, we do this within the discourse, not outside or on top of it. This autoreferentiality of the discourse makes it, by necessity, circular. It holds for the discourse as it is known to hold for any dictionary that we can never

	escape circularity in making sense of language. Each lexical item refers to other lexical items. Whenever a new lexical item(and each lexical item once has been a new item) is introduced into the discourse it has to be explained in terms of lexical items which are already available.
10. Corpus linguistics is not in itself a method: many different methods are used in processing and analysing corpus data. It is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus. However, one should be wary of using such data merely to find out more about what we know already, since what (we think) we know is often derived from pre-corpus study. Corpus data provide insights of a type which has not previously been available. Concepts and categories derived from introspective language study or from models taken from other fields (e.g. computation) may not be appropriate for describing real language data.	23. The discourse contains only testimony, provided by the members of the discourse community. It does not contain first-person experiences. The discourse-external reality can enter the discourse only as testimony. The link between the discourse and the discourse-external reality is, as said before, not part of the language system; it has to be established by each member of the discourse community individually.
11. Corpus linguistics does not have its starting point in language universals if we understand universals as ontological features (and not as theoretical concepts). Little is reliably known about the language faculty all human beings share. The study of this language faculty is outside the remit of corpus linguistics. Rather, corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions. It is primarily concerned with the contingencies of language use. Normally, we become aware of language only if there is a communication disorder. These disorders have their origins in the variation we find within and between discourses. They can be analysed in terms of the differences we observe between one language use and another.	24. Corpus linguistics does not distinguish between lexical meaning and encyclopaedic meaning. The meaning of the unit *lemon* is everything that has been said about lemons. Lexical items and what they stand for are discourse objects (and not objects of the 'real world'), constructed through the contributions of the members of the discourse community. As discourse objects, unicorns are as real as lemons. It is up to each member to decide for themselves whether unicorns or lemons are part of their first-person experiences.
12. The (single) word is not privileged in terms of meaning. The corpus linguist posits endocentric entities, formally held together by some local	25. Linguistics is not a science like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of

grammar, and calls these entities (complex) lexical items or, alternatively, units of meaning. Lexical items can be single words, compounds, multiword units, phrases, and even idioms. Just like single words, (complex) lexical items tend to recur in a discourse. This is why statistical procedures can be used for detecting them in a reasonably large corpus, as significant co-occurrences of the same entities.	the endeavour to make sense of the human condition. Interpretation, and not verification, is the proper response to the quest for meaning. There is no true meaning. The corpus linguist is not privileged as an 'expert' to pass judgment on what is permissible and what not. He or she is part of a discourse community, not outside of it. Corpus linguists have to submit their findings to their discourse community and argue for their acceptance. The discourse community is, in principle, a democratic community. Every member has the right to contribute to the discourse, and to discuss, modify or reject what other members say. The discourse organises itself. All regimentation from the outside strangles the creativity of the discourse community.
13. Frequency is an important parameter for detecting recurrent patterns defined by the co-occurrence of words. Frequency is thus an essential feature for making general claims about the discourse. However, statistical 'significance' is never enough. Lexical items also have to be semantically relevant.	

Корпус лингвистикасы қазіргі кезде машинада оқылатын мәтіндердің үлкен коллекциялары – корпустар (*корпоралар)* - арқылы лингвистикалық құбылысты зерттеу ретінде қарастырылады. Корпус лингвистикасы, дегенмен, компьютерлерді қолдану арқылы тілдік деректерге қол жеткізумен бірдей емес. Корпус лингвистикасы - бұл корпустан алынған деректерді зерттеу және талдау. Корпус лингвисінің негізгі міндеті - деректерді табу емес, оларды талдау. Компьютерлер бұл процесте қолданылатын пайдалы, кейде таптырмайтын құралдар.

Корпус лингвистикасы қолданбалы лингвистика саласына кіреді. Қазақ зерттеулерінде жоғарыда айтқан шет тілді зерттеушілер арасындағы даулы мәселе жоқ. олар бір анықтама берумен шектеледі, көбіне басқа зерттеушілер ұсынған анықтаманы пайдаланады. Мысалы, «Компьютерлік лингвистика» оқулығын жазған ҚАЗҰУ ғалымдары ..... «Корпустық лингвистика – лингвистикалық корпустарды компьютердің көмегімен қолдану және жалпы құру қағидаларын құрастырумен айналысатын қолданбалы лингвистиканың бөлігі [1, 3]» - деген анықтаманы пайдаланған. «Бұл анықтамаға сай, корпустық лингвистика екі аспектіні қамтиды:

1) автоматты құралдар қолданатын мәтіндер корпусын құрастыру;

2) әртүрлі типтегі корпустар базасында тілдің әртүрлі деңгейлерінде эксперименталды зерттеулер тәсілін құрастыру (Г. Б. Мәдиева, С. Б. Бектемірова, Ж.Ж. Күзембекова. Компьютерлік лингвистика: оқу құралы / Г.Б. Мәдиева, С.Б. Бектемірова, Ж.Ж. Күзембекова. – Алматы: Қазақ университеті, 2016. – 164 б.)

Толдова С.Ю., Архипов А.В., Логинова Е.А., Попова Д.П. Корпусная лингвистика // Фонд знаний «Ломоносов». – М., 2011. www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article (қаралған күн: 28.02.2012).