Дәріс №9
Дәріс тақырыбы: Корпустарды белгілеу құралдары
Сұрақтар:

1. Белгіленім ұғымы туралы
2. Лингвистикалық белгіленім және оның түрлері
   2.1 Морфологиялық белгіленім
   2.2 Синтаксистік белгіленім

1. Табиғи өңдеуге арналған арнайы бағдарламалардың арасында автоматты белгілеу бағдарламалары ерекше орын алады. markup - белгілеу/белгіленім бірнеше баламалар жарыса қолданылып жүр (мақаламда айтқанмын). Корпусқа аннотация жасау үшін қолданылатын арнайы кодтарға арналған термин; немесе осы кодтарды мәтінге қосу процесі.

> markup A term for the special codes used to annotate a corpus; or, the process of adding these codes to a text. (A glossary of Corpus linguistics. *Paul Baker, Andrew Hardie and Tony McEnery 2006)*

Корпусты белгілеу (тегтеу, аннотация) әсіресе қазіргі корпустардың өлшемін ескерсек, көп еңбекті қажет ететін операция. Белгілеудің кейбір түрлері үшін, атап айтқанда, анафоралық, просодиялық, автоматты жүйелерді құру әлі де айтарлықтай қиын және жұмыстың негізгі бөлігі қолмен жасалса, морфологиялық және синтаксистік талдау үшін әртүрлі бағдарламалық құралдар бар, олар әдетте теггерлер (taggers) және парсерлер (parsers) деп аталады. Автоматты морфологиялық талдау (теггерлер) бағдарламаларының жұмысының нәтижесінде әрбір лексикалық бірлікке грамматикалық сипаттамалар, оның ішінде сөз табы, лемма және граммемалар жиынтығы (мысалы, тек, сан, септік, жанды/жансыз және т.б. – орыс тілінде) беріледі. Автоматты синтаксистік талдау бағдарламаларының жұмысының нәтижесінде сөздер мен сөз тіркестерінің арасындағы синтаксистік байланыстар бекітіліп, синтаксистік бірліктерге сәйкес белгілер (сөйлем түрі, сөз тіркесінің синтаксистік қызметі т.б.) беріледі.

Дегенмен, табиғи тілді автоматты талдау қатесіз болмайды және көпмәнді болады, ол әдетте бір лексикалық бірлікке (сөздер, сөз тіркестері, сөйлемдер) бірнеше талдау нұсқаларын береді. Бұл **грамматикалық омонимия** мәселесі деп аталады. Корпусты құру кезіндегі екіұштылықты/көпмәнділікті/неоднозначности шешу үшін автоматты және қолмен әдістер қолданылады. Жаңа буын корпусы жүздеген миллион сөздерді қамтиды, сондықтан адамның араласуын барынша азайтатын жүйелерді әзірлеу принциптері алға қойылған. Морфологиялық немесе синтаксистік екіұштылықты автоматты шешу, әдетте, статистикалық әдістерді пайдалана отырып, жоғары деңгейдегі ақпаратты (синтаксистік, семантикалық) пайдаланумен жүзеге асады.

Бір мысал: кейбір тұлғалар бірнеше категорияға тиесілі болуы мүмкін. Бұл мәселе «морфологиялық екіұштылық (ambiguity) мәселесі» деп аталады. Мысалы, бір сөз бірнеше сөз табына жатады. Айталық, аш, қаз, қара т.т. Ағылшын тілінде де бар: *words, forms, can, use, present* және *process* зат есім де, етістік те бола алады. Әрине, контексте (яғни, нақты қолданыста) сөз формасы тек бір категорияға жатады. Сондықтан ағылшын корпусының дәл белгілеуіне контексті талдау немесе одан жоғары деңгейді - морфологиялық белгілеу үшін синтаксистік талдау, синтаксистік белгілеу үшін семантикалық - талдау арқылы қол жеткізуге болады.

> Конечно, в контексте (т.е. в действительном использовании) словоформа принадлежит только одной категории. Следовательно, достичь точной разметки английского корпуса можно путем анализа контекста или анализа более высокого

уровня: синтаксического анализа для морфологической разметки, семантического – для синтаксической.

2. Лингвистикалық белгілеу түрлеріне мыналар жатады: морфологиялық, синтаксистік, семантикалық, анафоралық, просодиялық, дискурстық т.б. Олардың барлығы мынадай принциптерге сәйкес жүзеге асырылады:
1) описание (обоснование) схемы разметки;
2) общепринятая система лингвистических понятий;
3) известная для пользователя схема анализа;
4) мотивированность введения параметров;
5) теоретически нейтральная (традиционная) схема разметки;
6) следование международным стандартам.

**Морфологиялық белгі.** Ең жиі кездесетін түрі – сөз таптарын белгілеу. Ағылшын тіліндегі термині - POS-tagging. Сөзбе сөз аударсақ, сөз таптарын тегтеу. Шындығында, морфологиялық белгілерге тек сөйлем мүшесінің ерекшелігі ғана емес, сонымен бірге берілген сөйлем мүшесіне тән грамматикалық категориялардың белгілері де жатады.

Браун корпусының морфологиялық белгілеуі:
the_AT jury_NN further_RB said_VBD in_IN term-end_NN
presentments_NNS that_CS the_AT *city_NP *executive_NP *committee_NP
,_, which_WDT had_HVD over-all_JJ charge_NN of_IN the_AT election_NN
,_, deserves_VBZ the_AT praise_NN and_CC thanks_NNS of_IN the_AT
*city_NP of_NP *atlanta_NP for_IN the_AT manner_NN in_IN which_WDT
the_AT election_NN was_BEDZ conducted_VBN | **(осы тегтеуден сөйлемді алып шығыңыз)**

В.Захаров пен С.Богданова келтірген орыс тіліндегі мәтін үзіндісіне қойылған морфологиялық белгілеуді көрейік. Белгілеу XML-форматында АОТ белгілеуішімен қойылған. (*рис. 1*).

«*Звонили к вечерне. Торжественный гул колоколов*»

В представленной записи использованы тэги <text> – текст, <p> – абзац, <s> – предложение, <w> – словоупотребление, <pun> – знак пунктуации. Тэг <w> содержит вложенный тэг <ana> с атрибутами <lemma> – лемма, <pos> – часть речи, <gram> – набор граммем. Значения граммем приводятся в Приложении 3. (**қараңыздар** )

```
<?xml version="1.0" encoding="windows-1251" ?> <text> <p>
<s>
<w>Звонили<ana lemma="ЗВОНИТЬ" pos="Г" gram="мн,нс,нп,дст,прш,"
/></w>
<w>к<ana lemma="К" pos="ПРЕДЛ" gram="" /></w>
<w>вечерне
<ana lemma="ВЕЧЕРНЯ" pos="С" gram="жр,ед,дт,пр,но," />
<ana lemma="ВЕЧЕРНИЙ" pos="П" gram="ср,ед,кр," /></w>
<pun>.</pun> </s>
<s><w>Торжественный<ana lemma="ТОРЖЕСТВЕННЫЙ" pos="П"
gram="мр,ед,им,вн," /></w>
<w>гул<ana lemma="ГУЛ" pos="С" gram="мр,ед,им,вн,но," /></w>
<w>колоколов
<ana lemma="КОЛОКОЛ" pos="С" gram="мр,мн,рд,но," />
<ana lemma="КОЛОКОЛОВ" pos="С" gram="мр,фам,ед,им,од," /></w>
……<pun>.</pun> </s></p></text>
```

*Рис. 1.* Пример морфологической разметки текста на русском языке (список граммем см. Приложение 3)

Синтаксистік белгілеу.

Синтаксистік белгілеу морфологиялық талдау деректері негізінде жасалған парсингтің нәтижесі болып табылады. Белгілеудің бұл түрі лексикалық бірліктер мен

әртүрлі синтаксистік құрылымдар арасындағы синтаксистік байланыстарды сипаттайды (мысалы, бағыныңқы сөйлем, етістікті сөз тіркесі, т.б.).

Орыс тілі үшін синтаксистік талдау көбінесе тәуелдік құрылымдармен беріледі. 2-суретте тәуелділік ағашының (дерева зависимостей) визуализациясының мысалы көрсетілген.

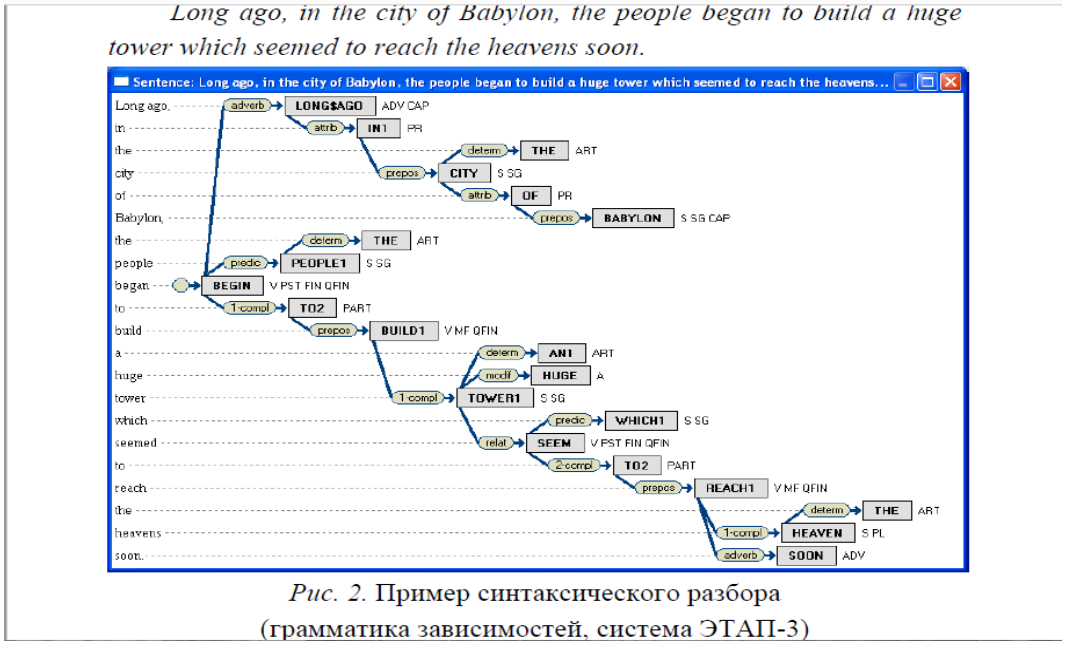


*Рис. 2.* Пример синтаксического разбора
(грамматика зависимостей, система ЭТАП-3)

http://corpora.lancs.ac.uk/clmtp/1-annot.php

**CLAWS** tagger (below). Ланкастер университетінде UCREL жасаған теггер. Бұл BNC-ті белгілеу үшін пайдаланылған бағдарламалық жасақтама болып табылады. Оны екі тегтер жинағының кез келгенін пайдалану үшін конфигурациялауға болады: стандартты C7 және күрделі емес C5.

**CLAWS** tagger (below). This tagger, created by UCREL at Lancaster University, is the software that was used to tag the BNC. It can be set to use either of two tagsets, the standard C7 and the less-complex C5.

**Өзіміз тегтеу жасап көрейік. CLAWS** tagger http://ucrel-api.lancaster.ac.uk/claws/free.html

A more complex form of grammatical annotation is **parsing**. One easy way to try out parsing is to use the online Stanford Parser.

**Өзіміз парсинг жасап көрейік.** Stanford Parser

**http://nlp.stanford.edu:8080/parser/index.jsp**