Дәріс №10
Дәріс тақырыбы: Лингвистикалық белгіленім және оның түрлері
Сұрақтар:

1. Семантикалық белгіленім
2. Анафоралық белгіленім және просодикалық белгіленім
3. Экстралингвистикалық белгіленім

1. Семантикалық белгілеу. «Семантикалық тегтер көбінесе берілген сөз немесе сөз тіркесі жататын семантикалық категорияларды және оның мағынасын ерекшелейтін ішкі категорияларды (более узкие подкатегории) белгілейді», - дейді Захаров пен Богданова. «Корпусты семантикалық белгілеу сөздердің мағынасын нақтылауды, омонимия мен синонимияны шешуді, сөздерді (категорияларды) жіктеуді, тақырыптық класстарды бөлуді, себепті байланыс белгілерін, бағалау және туынды сипаттамаларды және т.б.» «Семантическая разметка корпусов предусматривает спецификацию значения слов, разрешение омонимии и синонимии, категоризацию слов (разряды), выделение тематических классов, признаков каузативности, оценочных и деривационных характеристик и т.д.» (Захаров, Богданова)

Семантикалық аннотациялауды автоматтандыру грамматикалық ақпаратты аннотациялауды автоматтандыруға қарағанда қиынырақ; сондықтан ол көбінесе қолмен немесе жартылай қолмен орындалады. Семантикалық аннотацияланған корпустар аз, ал шын мәнінде көптеген аннотация жүйелері бағдарламалық ұсыныстар болып табылады (CORPUS MARKUP AND ANNOTATION).

Сөз сезімталдығы. (Word senses) Корпустармен жұмыс істегенде кездесетін мәселенің бірі – бір сөз формасы әртүрлі сөздерді білдіруі (омонимия) және сөздің бірнеше байланысты мағыналары болуы (полисемия) құбылысы. Ықтимал шешім әрбір сөзді леммасымен (ол көрсететін сөздің негізгі формасы) аннотациялау және ол қолданылған нақты мағынаның кодын аннотациялау болып табылады. (as in the following example, where this code (the value of the lexsn attribute) refers to the computerized dictionary Wordnet: (мен берген параққа қараңыздар: CORPUS MARKUP AND ANNOTATION) http://wordnetweb.princeton.edu/perl/webwn

Semantic roles. semantic annotation systems may go beyond word senses to encode semantic properties the clause. For example, the FrameNet annotation system annotates nominal and clausal arguments for the role that they play in the semantic frame associated with verbs or deverbal nouns:

With this, [El Cid Agent] at once avenged [the death of his son Injury]. [Hook Agent] tries to avenge [himself Injured_party] [on Peter Pan Victim] [by becoming a second and better father Response_action].

This system may be extended to include information about phrase type and grammatical function:

[Managers (Speaker, NP, Ext)] claim [there was no radiological hazard to staff or the public (Message, Sfin, Comp)]. [His (Speaker, Poss, Gen)] claims [to have more energy (Message, VPto, Comp)] are simply laughable.

Metaphor Corpora may be annotated for even more abstract semantic properties. As an example, consider metaphor. In metaphorical expressions, there is always one element (the target), which is treated linguistically as though it were an element from a different semantic domain (the source domain). For example, in the sentence *Stocks are very sensitive creatures*, stocks are treated

as though they were an organism. This may be annotated as follows (a system suggested by Trausan-Matu et al.): (қараңыздар мысал)

In this system, the attribute what refers to the target-domain item, the attribute how refers to the source-domain concept, and why refers to the property of the source domain concept that motivates the use of the metaphor. A simpler system might simply use a Framenet-type annotation scheme: (қараңыздар мысал)

Орыс тілінің ұлттық корпусына семантикалық белгілеу жасалған (қараңыз https://ruscorpora.ru/page/instruction-semantic/). Мәтіндегі кез келген сөзге берілген лексика-семантикалық ақпарат үш топтық белгілеулерден тұрады:

1) разряд (имя собственное, возвратное местоимение и т.д.);
2) лексико-семантические характеристики (тематический класс лексемы, признаки каузативности, оценки и т.д.);
3) деривационные характеристики («диминутив», «отадъективное наречие» и т.д.).

Лексико-семантикалық тэгтер былайша топтастырылған:

• таксономия (тематический класс лексемы) – для имен существительных, прилагательных, глаголов и наречий;
• мереология (указание на отношения «часть – целое», «элемент – множество») – для предметных и непредметных имен;
• топология (топологический статус обозначаемого объекта) – для предметных имен;
• каузация – для глаголов;
• служебный статус – для глаголов;
• оценка – для предметных и непредметных имен, прилагательных и наречий. (әрі қарай қараңыз https://ruscorpora.ru/page/instruction-semantic/)

Орыс тілінің ұлттық корпусынан:

При такой разметке большинству слов в тексте приписывается один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество', 'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'диминутив', 'отглагольное имя' и т. п. Используется фасетная классификация, при которой одно слово может попадать в несколько классов. На первом этапе поиск осуществляется по части имеющихся в словаре признаков. https://ruscorpora.ru/page/instruction-semantic/

Разметка текстов осуществляется автоматически с помощью программы Semmarkup (автор А. Е. Поляков) в соответствии с Семантическим словарем Корпуса. Поскольку ручная обработка семантически размеченных текстов очень трудоемка, семантическая омонимия в Корпусе не снимается: многозначным словам приписывается несколько альтернативных наборов семантических признаков.

Шетел корпустарындағы семантикалық белгілеудің ерекшеліктерін қарайық. Ланкастер университеті семантикалық белгілеудің USAS деп аталатын семантикалық тегтеу жүйесін дамытқан. Кестеге қараңыздар. (мен берген парақты қараңыздар). - *Corpus Mark-up and Annotation*

Semantic tagging

Beyond grammatical annotations, semantic annotation is an obvious next step. For example, semantic word-tagging can be designed with the limited (though ambitious enough) goal of distinguishing the lexicographic senses of same word: a procedure also known as `sense resolution'.

The ACASD semantic tagging system (Wilson and Rayson, 1993) accepts as input text which has been tagged for part of speech using the CLAWS POS tagging system. The tagged text is fed into the main semantic analysis program (SEMTAG), which assigns semantic tags representing the general sense field of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. *as a rule*), which are updated as new texts are analyzed. (Items not contained in the lexicon or idiom list are assigned a special tag, Z99, to assist in updating and manual postediting.) The tags for each entry in the lexicon and idiom list are arranged in general rank frequency order for the language. The text is manually pre-scanned to determine which semantic domains are dominant; the codes for these major domains are entered into a file called the `disam' file and are promoted to maximum frequency in the tag lists for each word where present. This combination of general frequency data and promotion by domain, together with heuristics for identifying auxiliary verbs, considerably reduces mistagging of ambiguous words. (Further work will attempt to develop more sophisticated probabilistic methods for disambiguation.) After automatic tag assignment has been carried out, manual postediting takes place, if desired, to ensure that each word and idiom carries the correct semantic classification (SEMEDIT). A program (MATRIX) then marks key lexical relations (e.g. negation, modifier + adjective, and adjective + noun combinations). The following is an example of semantic word-tagging, taken from the automatic content analysis project at Lancaster:

EXAMPLE OF SEMANTIC TAGGING

```
PPIS1    I          Z8
VV0      like       E2+
AT1      a          Z5
JJ       particular A4.2+
NN1      shade      O4.3
IO       of         Z5
NN1      lipstick   B4
```

In this fragment, the text is read downwards, with the grammatical tags on the left, and the semantic tags on the right. The semantic tags are composed of:

- an upper case letter indicating general discourse field
- a digit indicating a first subdivision of the field
- (optionally) a decimal point followed by a further digit to indicate a finer subdivision
- (optionally) one or more `pluses' or `minuses' to indicate a positive or negative position on a semantic scale

For example, A4.2+ indicates a word in the category `general and abstract words' (A), the subcategory `classification' (A4), the sub-subcategory `particular and general' (A4.2), and `particular' as opposed to `general' (A4.2+). Likewise, E2+ belongs to the category `emotional states, actions, events and processes' (E), subcategory `liking and disliking' (E2), and refers to `liking' rather than `disliking' (E2+).

The semantic annotation is designed to apply to open-class or `content' words. Words belonging to closed classes, as well as proper nouns, are marked by a tag with an initial Z, and set aside from the statistical analysis.

For more information, see USAS web page, Thomas and Wilson (1996) and Garside and Rayson (1997).

https://ucrel.lancs.ac.uk/annotation.html#POS

2. Анафоралық белгілеу. Ол референттік байланыстарды белгілейді, мысалы, есімдіктерді;

АНАФОРА (греч. anapherein 'относить назад, возводить <к чему-л.>, возвращать'), использование языковых выражений, которые могут быть проинтерпретированы лишь с учетом другого, как правило предшествующего, фрагмента текста. В первую очередь, понятие анафоры, или анафорической отсылки используется в лингвистике применительно к анафорическим местоимениям – ср., например, следующий микротекст: (1) Космонавті вернулся на борт станции. Оні сообщил, что чувствует себя нормально. Анафорическое выражение (или анафóр) он во втором предложении может быть понято адресатом (и использовано автором) такого текста лишь на том основании, что соответствующий референт «космонавт» уже введен в предыдущем предложении. В анафорической функции выступают местоимения 3 лица, а также другие типы местоимений, в частности указательные, возвратные, относительные (последние два типа имеют ярко выраженную синтаксическую специфику). Однако не все местоимения являются анафорическими.

Anaphoric annotation

The UCREL anaphoric annotation scheme co-indexes pronouns and noun phrases within the broad framework of cohesion such as is described by Halliday and Hasan (1976). The software used for this annotation (XANADU) is an X-windows interactive editor written by Roger Garside. This allows the user to move around a block of text, displaying around 20 lines at a time. The user can use a mouse to highlight any segment of the text which s/he wishes to annotate. A further window displays a set of command keys, mainly listing the different types of anaphora. Clicking on one of the buttons sets up the insertion routine for that anaphor type. Another window lists the items which have previously been highlighted, and at some point in the insertion routine the annotator is required to click on the appropriate one as the antecedent of the anaphor (Fligelstone, 1992).

An example of this form of annotation is as follows:

ANAPHORIC ANNOTATION OF AP NEWSWIRE

S.1 (0) The state Supreme Court has refused to release
{1 [2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on
bail .
S.2 (1 The fighter 1) is serving 30-40 years for a 1975 armed
robbery conviction .
S.3 (1 Scott 1) had asked for freedom while <1 he waits for an
appeal decision .
S.4 Meanwhile , [3 <1 his promoter 3] , {{3 Murad Muhammed 3} ,
said Wednesday <3 he netted only $15,250 for (4 [1 Scott 1] 's
nationally televised light heavyweight fight against {5 ranking
contender 5}} (5 Yaqui Lopez 5) last Saturday 4) .
S.5 (4 The fight , in which [1 Scott 1] won a unanimous
decision over (5 Lopez 5) 4) , grossed $135,000 for [6
[3 Muhammed 3] 's firm 6], {{6 Triangle Productions of
Newark 6} , <3 he said .

Key: The use of the same index 1, 2, ... *n* binds one syntactic constituent to another to which it is coreferential or semantically equivalent. In the following list, i represents an arbitrary index:

(i i) OR
[i...] enclose a constituent (normally a noun phrase) entering into an equivalence `chain'
<i indicates a pronoun with a preceding antecedent
>i indicates a pronoun with a following antecedent
{{i i} enclose a noun phrase entering into a copular relationship with a preceding noun phrase
{i i}} enclose a noun phrase entering into a copular relationship with a following noun phrase
(0) represents an anaphoric barrier, in effect, the beginning of a new text.

Such annotations have potential use for studying and testing mechanisms like pronoun resolution, which are important for text understanding and machine translation.

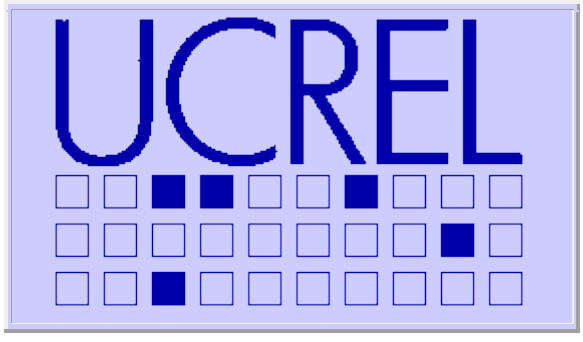
https://ucrel.lancs.ac.uk/annotation.html#POS

Просодикалық белгілеу. Просодикалық корпустарда екпін мен интонацияны белгілейтін тегтер пайдаланады. Ауызекі сөйлеу корпустарында просодикалық белгілеу көбінесе үзілістерді, қайталауларды, ескертпелер () және т.б. белгілеуге қызмет ететін дискурстық белгілеумен қоса жүреді.

Prosodic annotation

Prosodic annotation aims to indicate patterns of intonation, stress and pauses in speech. It is a much more difficult type of annotation to achieve than the types discussed above: it cannot be done automatically and requires careful listening by a trained ear. The Lancaster/IBM Spoken English Corpus is the only prosodically transcribed corpus in which Lancaster has been involved to date. The prosodic annotation of the SEC was carried out by two phoneticians (Gerry Knowles and Briony Williams). A set of 14 special characters was used to represent prosodic features. Stressed syllables were marked with a symbol indicating the direction of the pitch movement. Syllables which were felt to be stressed but with no independent pitch movement were marked with a circle (or bullet in the printed version). Unstressed syllables, whose pitch is predictable from the tone marks of surrounding accented syllables, were left unmarked.

Prosody is considerably more impressionistic than other linguistic levels in corpus annotation. Thus to check on the consistency of transcriptions, some sections (approx. 9% of the corpus) were independently transcribed by both transcribers. There are considerable differences in these transcriptions (cf. Wilson, 1989; Knowles, 1991), but the resulting correlations could have important implications for future research (cf. Wichmann, 1991).



Lancaster university

Prosody

Prosody refers to all aspects of the sound system above the level of segmental sounds e.g. stress, intonation and rhythm. The annotations in prosodically annotated corpora typically follow widely accepted descriptive frameworks for prosody such as that of O'Connor and Arnold (1961). Usually, only the most prominent intonations are annotated, rather than the intonation of every syllable. The example below is taken from the London-Lund corpus:

```
1 8 14 1470 1 1 A 11 ^what a_bout a cigar\ette# .                    /
1 8 15 1480 1 1 A 20 *((4 sylls))*                                   /
1 8 14 1490 1 1 B 11 *I ^w\on't have one th/anks#* - - -     /
1 8 14 1500 1 1 A 11 ^aren't you .going to sit d/own# -      /
1 8 14 1510 1 1 B 11 ^[/\m]# -                                                  /
1 8 14 1520 1 1 A 11 ^have my _coffee in p=eace# - - -                /
1 8 14 1530 1 1 B 11 ^quite a nice .room to !s\it in ((actually))# /
1 8 14 1540 1 1 B 11 *^\isn't* it#                                   /
1 5 15 1550 1 1 A 11 *^y/\es#* - - -                                 /
```

The codes used in this example are:

end of tone group
^ onset
/ rising nuclear tone
\ falling nuclear tone
/\ rise-fall nuclear tone
_ level nuclear tone
[] enclose partial words and phonetic symbols
. normal stress
! booster: higher pitch than preceding prominent syllable
= booster: continuance
(( )) unclear
* * simultaneous speech
- pause of one stress unit

Problems of Prosodic Corpora

1. **Judgements are inherently of an impressionistic nature.** For example, the level of a tone movement is a difficult matter to agree upon. Some listeners may perceive a fall in pitch, while others may perceive a slight rise after the fall. This leads to our second point:
2. **Consistency is difficult to maintain**, especially if more than one person annotates the corpus. (This can be alleviated to some degree by having two people both annotate a small part of the corpus.)
3. **Recoverability is difficult** (see Leech's 1st Maxim) since prosodic features are carried by syllables rather than whole words - annotations appear within the words themselves making it difficult for software to retrieve the raw corpus.
4. Sometimes special graphics characters are used to indicate prosodic phenomena. However, **not all computers and printers can handle such characters**. TEI guidelines for text encoding will hopefully alleviate these difficulties.

3. **Экстралингвистикалық белгілеу** метадеректер деп те аталады. Метадеректер – мәтіннің өзі туралы бірдеңе айтатын ақпарат – мысалы, метадеректер мәтінді кім жазғанын және оның қашан жарияланғанын айта алады. Метадеректер корпус мәтінінде кодталуы немесе бөлек құжатта немесе дерекқорда сақталуы мүмкін. Метадеректер әдетте мәтіндегі сөйлеушілерді анықтайды және олардың әрқайсысы туралы кейбір пайдалы ақпарат береді, мысалы, жасы мен жынысы сияқты. Мысалы, BNC-де әрбір айтылым белгіленген және белгілі бір сөйлеушіге арналған метадеректермен байланыстырылған. Әрбір сөйлеуші үшін келесі метадеректер сақталады:

- Name (anonymised)
- Sex
- Age
- Social class
- Education
- First language
- Dialect/Accent
- Occupation

Біз бұл метадеректерді BNC-тегі іздеулерді шектеу үшін пайдалана аламыз — мысалы, 35 пен 44 жас аралығындағы әйелдер сөйлеген сөздің барлық мысалдарын алып шығу үшін.

Экстралингвистикалық белгілеу «сыртқы», «интеллектуалдық» белгілеуді (библиографиялық сипаттамалар, типологиялық сипаттамалар, тақырыптық сипаттамалар, әлеуметтік сипаттамалар), «формальды» құрылымдық белгілеуді (мәтін, бөлім, тарау, бөлік, абзац, сөйлем), сонымен қатар техникалық және технологиялық белгілеу (кодтау, өңдеу мерзімі, орындаушылар, электронды нұсқаның көзі) кіреді.

«Сыртқы», «интеллектуалды» белгілеу, біріншіден, тіл мен оның өмір сүру жағдайларының арасындағы байланысты анықтау үшін қажет; екіншіден, тілдің жеке ішкі жиынтығын зерттеу. Мәтіндердің тіліне әсер ететін факторлардың екі класы бар:
 сыртқы, экстралингвистикалық факторлар (E – external сыртқы);
 ішкі факторлар (I – internal ішкі).

Дж.Синклер E-факторлардың үш тобын ажыратады:
 E1 (origin) – автордың мәтін құруына байланысты факторлар;
 E2 (state) – мәтіннің сыртқы ерекшеліктеріне байланысты факторлар (оның ішінде ауызша немесе жазбаша сөйлеуді қоса);
 E3 (aims) – мәтіннің жасалу себептеріне және оның аудиторияға әсеріне байланысты факторлар
және I-факторлардың екі тобы:
 I1 (topic)– мәтіннің тақырыптық аймағы;
 I2 (style) – стильдік ерекшеліктер (стиль, жанр) (*Sinclair J.M.* Preliminary recommendations on text typology. 1996. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html)

Орыс тілінің ұлттық корпусында мәтінге мынадай метадеректер (оны метаразметка текстов деп атаған) қолданылған: https://ruscorpora.ru/page/instruction-parameter/
Первый блок:
1) *автор текста*: имя, пол, дата рождения (или примерный возраст);
2) *название текста*;
3) *время и место создания текста* (может указываться точно или приблизительно);

4) *объем текста*: для художественных произведений принято, что обычная длина рассказа – менее 5 тыс. слов; обычная длина повести – от 5 до 15 тыс. слов; обычная длина романа – более 15 тыс. слов.

Второй блок: параметры метаописания трех основных *массивов* текстов корпуса – художественных текстов; нехудожественных текстов; драматургических произведений. Например, для художественных текстов в НКРЯ указывается:
1) жанр текста: нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатира;
2) тип текста: автобиографическая проза, анекдот, ассоциативная проза, боевик, детектив, очерк, литературное письмо, повесть, притча, пьеса, рассказ, роман, сказка, триллер, эпопея, эссе и др.;
3) хронотоп текста: приблизительное указание на место и время описываемых в тексте событий [27].
Реально предлагается следующее: ревний Восток; Россия XVII век; Россия XIX век; Россия/СССР: советский период в целом; Россия, советский период – Германия 1920-1940-е годы; Россия/СССР – Европа 1960-1980-е годы; Россия/СССР: перестройка;
Россия/СССР: советский и постсоветский период; Америка: современная жизнь; Израиль: современная жизнь; Средняя Азия: современная жизнь; ирреальный мир и др. Также может встретиться тэг «хронотоп не определен».
Метадеерктер туралы мына сілтемеден оқисыздар:
https://ruscorpora.ru/new/corpora-parameter.html

Академиялық ағылшын сөйлеу тілінің Мичиган корпусын ашып, осы метадеректер не үшін қажет болады соны түсініп көрейік. https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple