Дәріс №8
Дәріс тақырыбы: Табиғи тілді өңдеудің негізгі рәсімдері
Сұрақтар:

1. Токен және токендеу
2. Лемма және леммалау
3. Стемминг
4. Парсинг

1**.** Шын мәнінде, корпус өзінің қазіргі мағынасында компьютерлік деректер қоры, ал оны құру барысында арнайы процедуралар мен бағдарламаларды қолдану заңды.

**Токен -** бір тілдік бірлік, көбінесе сөз, дегенмен қолданылатын кодтау жүйесіне байланысты бір сөз бірнеше токендерге бөліне алады. мысалы *he's* (*he + 's*).

> **token** A single linguistic unit, most often a word, although depending on the **encoding** system being used, a single word can be split into more than one token, for example *he's* (*he + 's*).
>
> Paul Baker, Andrew Hardie and Tony McEnery. (2006). A Glossary of Corpus Linguistics. Edinburgh University Press

- Кейбір еңбектерде сөз тұлғалары: Токены (словоформы) (А.Захаров, С.Богданова)

**токендеу/изация** мәтіннің барлығын жекелеген токендерге көшірудің/аударудың автоматты процесі, мысалы, *he's* сияқты біріккен сөздерді бөлу, тыныс белгілерін (үтір және нүкте сияқты) сөздерден бөлу және бас әріптерді алып тастау. Токендеу әдетте леммалаудың немесе сөз таптарын тегтеудің алғашқы қадамы болып табылады.

> **tokenisation** The automatic process of converting all of a text into separate **token**s, for example, by splitting conjoined words like *he's*, separating punctuation (such as commas and full stops) from words and removing capitalisation. Tokenisation is usually the first stage in **lemmatisation** or **part-of-speech tagging**.
>
> A GLOSSARY OF CORPUS LINGUISTICS 159

**лемма** Сөздің канондық түрі (дұрыс көпше түрі грекше – леммата (lemmata), дегенмен кейбір адамдар көпше түрін леммалар (lemmas) ретінде жазады және лемматаны сәл педантикалық деп санауы мүмкін). У.Фрэнсис пен Кучера (1982: 1) лемманы «бірдей стемы бар және бір негізгі сөз табына жататын, тек флексия және/немесе емлесі бойынша ғана ажыратылатын лексикалық тұлғалардың жиынтығы» деп анықтайды. Лемматизацияланған тұлғалар кейде кішкентай бас әріппен жазылады, мысалы, walk етістік леммасы walk, walked, walking and walks сөздерінен тұрады. Корпус зерттеулерінде сөздер жиілігі кейде түрлермен емес, леммалармен есептеледі; сөздерге леммаландыру деп аталатын аннотация түрі де берілуі мүмкін.

> **lemma** The canonical form of a word (the correct Greek plural is lemmata, although some people write the plural as lemmas and may consider lemmata to be somewhat pedantic). Francis and Kuc_era (1982: 1) define it as a 'set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling'. Lemmatised forms are sometimes written as small capitals, for example the verb lemma walk consists of the words walk, walked, walking and walks. In corpus studies, word frequencies are sometimes calculated on lemmata rather than types; words can also be given a form of annotation known as lemmatisation.

**Леммаландыру/лемматизация** Сөз таптарын анықтаумен тығыз байланысты және корпустағы сөздерді сәйкес лексемаларға келтіруді көздейтін автоматты аннотацияның түрі. Лемматизация зерттеушіге белгілі бір лексеманың барлық нұсқаларын барлық мүмкін

нұсқаларды енгізбей-ақ бөліп алып, зерттеуге, сондай-ақ лексеманың жиілігі мен таралуы туралы ақпарат алуға мүмкіндік береді. Келесі тізімде сөздердің екінші бағаны лемматизацияланған. Қараңыз Beal (1987).

> **lemmatisation** A form of automatic annotation that is closely allied to the identification of parts-of-speech and involves the reduction of the words in a corpus to their respective lexemes. Lemmatisation allows the researcher to extract and examine all the variants of a particular lexeme without having to input all the possible variants, and to produce frequency and distribution information for the lexeme. In the following list the second column of words have been lemmatised. See Beale (1987).
> He he
> studied study
> the the
> problem problem
> for for
> a a
> few few
> seconds second
> and and
> thought think
> of of
> a a
> means means
> by by
> which which
> it it
> might may
> be be
> solved solve

**stem** Сөздің флекциялық аффикстер жалғанатын бөлігі; және керісінше, бұл аффикстерді алып тастағаннан кейін қалатын бөлік. Мысалы, *walk - walks, walked* and *walking* сөздерінің стемы. Сөздің стемын шығарудың лемматизация үшін маңызы зор және көбінесе морфологиялық анализатордың көмегімен жүзеге асады.

> **stem** The part of a word to which inflectional affixes are added; conversely, it is the part that remains when affixes are removed. For instance, walk is the stem of walks, walked and walking. Isolating the stem of a word is important for lemmatisation, and is often done by a morphological analyser. (See also lemma.)

Ниже приведены примеры стемминга и лемматизации. Дано следующее предложение:
*[The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs].*
Один из наиболее популярных стеммеров, Snowball Analyzer, выдает следующие стемы:
*[quick] [brown] [fox] [jump] [over] [lazy] [dog].*
Леммы слов данного предложения будут следующими:
*[the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog]. (Захаров и Богданова)*

Рассмотрим пример работы морфологического анализатора с английским предложением *"All women were walking in the streets"*. Токены (словоформы) представлены слева в скобках <>, звездочка '*' показывает, что слово в тексте начинается с заглавной буквы. Под каждым токеном располагается лемма (лексема) и приводится морфологический разбор. Мысалы "were" токені "be« леммасына қатысты және оның морфологиялық сипаттамасы – етістік, өткен шақ, жіктелетін; "streets" токені "street" леммасына жатады және оның морфологиялық сипаттамасы – зат есім, жалпы есім, жекелік және т.б.

Например, токен "were" относится к лемме "be", и его морфологические характеристики – глагол, прошедшее время, спрягаемый; токен "streets" относится к лемме "street", и его морфологические характеристики – существительное, нарицательное, ед. числа и т.д.
"<*all>" (токен)
"all" <*> <Quant> DET PRE SG/PL (лемма)
"<women>"
"woman" N NOM PL
"<were>"
"be" <SV> <SVC/N> <SVC/A> V PAST VFIN
"<walking>"
"walk" <SV> <SVO> PCP1
"<in>"
"in" PREP
"<the>"
"the" <Def> DET CENTRAL ART SG/PL
"<streets>"
"street" N NOM PL
"<$.>"

Дәл осы үлгімен Қазақ тілінің ұлттық корпусы мен Алматы корпусын ашайық, токен лемма стем талдаулары бар ма зерттейік.

**парсер** Мәтінге талдау тегтерін автоматты түрде қосатын компьютерлік бағдарлама: мысалдарға Minipar және Link Grammar Parser кіреді.

**parser** A computer program that adds parsing tags to a text automatically: examples include Minipar and the Link Grammar Parser.

**парсинг** Мәтін парсингтелген кезде, оның синтаксистік құрылымын көрсету үшін оған тегтер қосылады. Мысалы, зат есім, етістікті тіркестер және бағыныңқы сөйлемдер сияқты бірліктердің бастапқы және соңғы нүктелері парсинг тегтері арқылы көрсетіледі. Парсинг синтаксистік бірліктер бір-бірімен қалай байланысады сол туралы ақпаратты да қоса алады. Парсингтелген корпустар мысалдарына Lancaster–Leeds Treebank, the Penn Treebank, the Gothenburg Corpus and the CHRISTINE Corpus жатады.

**parsing** When a text is parsed, tags are added to it in order to indicate its syntactic structure. For instance, the start and end points of units such as noun phrases, verb phrases, and clauses would be indicated by parsing tags. The parse might also add information about how the syntactic units relate to one another. Examples of parsed corpora include the Lancaster–Leeds Treebank, the Penn Treebank, the Gothenburg Corpus and the CHRISTINE Corpus. (See also phrase structure, treebank, skeleton parsing.)

парсинг жайлы видео

**Parsing Explained - Computerphile**
https://www.youtube.com/watch?v=bxpc9Pp5pZM

**Lemmatization vs Stemming in NLP**
https://www.youtube.com/watch?v=7klD678xYxE

**Stemming And Lemmatization Tutorial | Natural Language Processing (NLP) With Python | Edureka**
https://www.youtube.com/watch?v=p1ccbR2P_xA

**Парсинг** – тілдің лексемаларының (сөздерінің, лексемаларының) сызықтық тізбегін оның формальды грамматикасымен сәйкестендіру процесі. Нәтиже әдетте тәуелділік ағашы (синтаксис ағашы) болып табылады. Ірі корпустар үшін автоматты синтаксистік анализаторларды (парсерлерді) құру – компьютер лингвистикасының маңызды бағыттарының бірі. (В.Захаров т.т.) В.Захаров, С.Богданова. (2011). Корпусная лингвистика.

- Иркутск: ИГЛУ

Парсинг – это процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой. Результатом обычно является дерево зависимостей (синтаксическое дерево). Построение автоматических синтаксических анализаторов (парсеров) для больших корпусов является одной из самых важных областей компьютерной лингвистики.