Scouts Association; Eng. *MBA* is a form of training a specialist in the field of business.

• Exclamations: Eng. *Oh my God!* – Ukr. *О, Боже мій*! as an expression of impatience, discontent.

• The names of movies: Eng. *Hot fuzz* – Ukr. *Типу круті легаві*, Eng. *21 jump street* – Ukr. *Мачо і Ботан*, Eng. *Hangover* – in Russian *Мальчишник в Вегасе*, Eng. *Identity Thief* – Ukr. Спіймай *товстуху, якщо зможеш*.

Conclusion. In the process of the given research, devoted to the Ukrainian – English and English – Ukrainian translation of "Shadows of Forgotten Ancestors" and "The Lord of the Rings", we come to the following conclusions:

1) Lexical units of the original language, which are not in the language of translation of the corresponding lexical equivalent, are called culture specific vocabulary; 2) The categories of non-equivalent vocabulary include: realities and historicism, terms, authorial neologisms, semantic gaps, complex words, phrases, abbreviations, volumes and exclamations; 3) The following methods of translation of lexical units are known in translation studies: vocabulary translation, transcoding, transcription, contextual replacement, developmental arithmetic, formal negatiation and descriptive translation; 4) The most common way of translation is the contextual change (about 70%).

References

1. Коцюбинський М. М. Тіні забутих предків / М. М. Коцюбинский. – Харків: Фоліо, 2017. – 154 с.

Kotsjubinsky M. M. Tini zabutykh predkiv [Shadows of Forgotten ancestors] / M. M. Kotsjubinsky. – Kharkiv: Folio, 2017. – 154 s.

2. Кочерган М. П. Загальне мовознавство. – К.: Академія , 2003. – 464 с.

Kochergan M. P. Zagal'ne movoznavstvo [General linguistics] / M. Kochergan. – Kyiv: Academia, 2003. – 464 s.

3. Bogatkina N. The internal form of the word as a manifestation of the peculiarity of the national language picture of the world // Scientific Notes. – Issue XXVI. – Series: Philological Sciences (Linguistics). – Kirovograd: RVC KDPU them. V. Vynnychenko, 2000.

4. Tolkien J. R. R. The Lord of the Rings: Two Towers / J. R. R. Tolkien, 1954. – Retrieved from: http://ae-lib.org.ua/texts-c/tolkien__the_lord_of_the_rings_2__en.htm

5. Ukrainian Academic Press for the Canadian Institute of Ukrainian Studies, 1981. – 33 p.

Alina Teletska
Vasyl' Stus Donetsk National University
Vinnytsia
Research Supervisor: V. I. Kalinichenko, PhD in Philology, Ass. Prof.
Language Advisor: V. I. Kalinichenko, PhD in Philology, Ass. Prof.

TECHNOLOGIES OF INFORMATION PROCESSING IN TEXT CORPORA

Introduction. Today text corpus is considered an important resource for finding information. It carries out researches of internal and external data and obtains statistic information of linguistic phenomena. There are a number of programs and tools for creating text corpus and analyzing its textual information.

Review of recent publications. This issue was deeply investigated by D. Jurafsky and J. Martin in their book "Speech and Language Processing". The paper of Anjali G. Jivani "A Comparative Study of Stemming Algorithms" was also devoted to this topic. Zhukovska V. V. in "Introduction to Corpus linguistics" and N. Darchuk in "Corpus linguistics: problems, methods, perspectives" deeply addressed to this research subject arear as well. In our paper we rely upon the three mentioned works.

Objective of the paper is to consider the effects of the linguistic operation during processing text corpora data and, in particular, the technologies of information processing in text corpora.

Results of the research. In general **text corpus** is "an electronic collection of texts, marked for quick search of words and word constructions with given grammatical and other characteristics that are interesting for linguists" [1: 8].

An important role in creating the text corpus takes **part-of-speech tagging**. The text of the natural language becomes formal with it. It provides further analysis of the language data and eventually the use of the body. The process of part-of-speech tagging consists in attributing texts and their components in special marks that have information about certain language phenomena.

There are also procedures and programs which take part in the formation of the text corpora: tokenization, parsing, lemmatization, stemming.

Tokenization is a process of lexical analysis of a text that divides sentences or strings of words into ***tokens*** – separate verbal units [2: 88]. Let us see an example in the sentence: *"Look, Miss Leefolt, she dress up nice ever day".* The tokenization process will follow the principle:

- at first, the program finds and selects certain symbols (. ; : ; () [] {} * - + = and others); in our case, these are commas ",": *"See , Miss Leefolt , she dress up nice ever day*". These symbols according to regular expressions can indicate the beginning or end of a token.
- next stage – selection of the tokens by the system of spaces and punctuation marks, that is the program can recognize 9 tokens in this sentence. In the next sentence 10 tokens will be recognized, including number "1924", because it is separated by spaces from the general string of words: "*It was 1924 and I'd just turned fifteen years old* ".

Parsing is a process of syntax analysis of the text that performs the analysis of the input data of the natural language according to the given formal grammar. A syntactic tree (tree of dependencies) is a result of the parsing analysis (figure 1) [2: 90]. Syntactic analysis is an important intermediate stage for further semantic text recognition. As a result, each verbal unit belongs to a certain part of speech.

One of the most commonly used parsing algorithms is the Earley algorithm, which is based on dynamic programming. Dynamic programming consists in making

optimal decision after performing some steps. There are also other algorithms for parsing analysis: the Cocke-Younger-Kasami algorithm (CYK) and the Graham-Harrison-Ruzzo (GHR) algorithm.

The advantage of the Earley algorithm is that it is capable of solving the problem of polysemy by applying two strategies of analysis – *top-down* and *bottom-down*. The top-down process of syntax analysis performs the analysis of the sentence from general to specific. There are following actions:

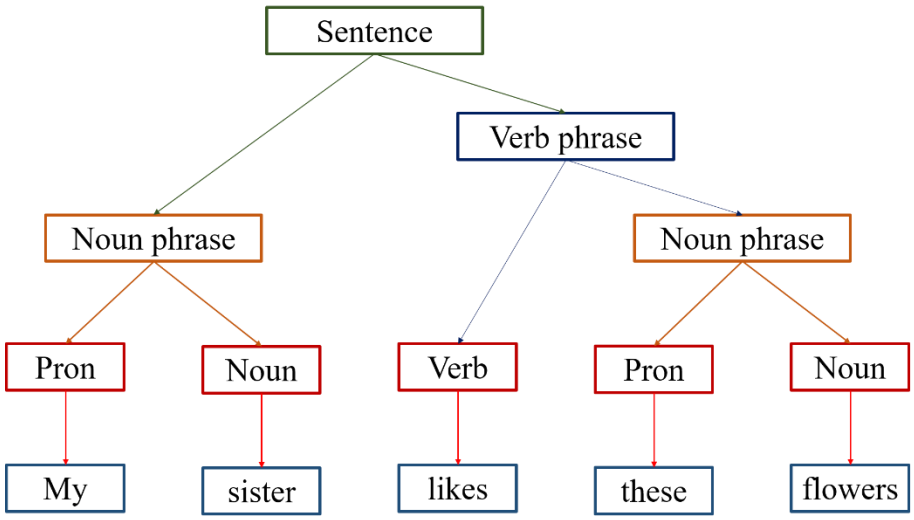

Figure 1. Parsing of the sentence

- the sign of the *Sentence (S)* is assigned to the input information;
- searching the corresponding lower levels of the tree from the left to the right according to the grammar rules.

The bottom-down algorithm creates a dependency tree, starting with the input words. The end point must be a completed sentence. This algorithm solves the problem of polysemy, considering all possible cases of using a certain word and choosing one – the most accurate. The top-down and bottom-down analysis strategies complement each other and make the process of parsing more qualitative [3: 353-359].

Lemmatization is the process of word transforming to its basic form (***lemma***), this means that the system creates a reference of various inflection forms of a word to its base form. Thus, lemmatization can group all forms of a particular word and, during the analysis of the text, consider them as one lexeme. For example, the word *sing* unites such inflection forms: *singing, sang, sung, sings;* and the lexeme *play* has such a paradigm: *playing, plays, played* [2: 88-89].

Stemming is the process of transforming common-root words to their common base (***stem***). During stemming suffixes and flexions are rejected, some algorithms can also cut off prefixes [4: 258-259]. One of the most accurate algorithms for the process of stem selection is the Porter algorithm (3: 83). In the example of the word *act*, we can see the result of the stemmer work: the word *act, acts, actor, actress, acted, acting, activ, activated* will be transformed to the stem *act*. Realization of this process takes place without using a context, this means that a result depends on the construction of the algorithm for the procedure implementation, consequently, there

is a probability that the conclusion will not be absolutely accurate. For example, such derivative forms differ from their base forms *twelve – twelfth, four – forty*.

Lemmatization and stemming are very similar processes, nevertheless there is a difference between them. Lemma is the base form of the word to which other forms of it are transformed. They have a common basis, they belong to the same part of the speech and have the same meaning: *formalizes, formalizing, formalized = formalize.* During stemming all word forms are transforming to a common base, despite being referred to different parts of speech: *formalizes, formalizing, formalized, formalization = formaliz.* Stemming is also not carried out for suppletive forms, since these lexemes do not have a common stem *(I – me, go – went, two - second)* [5: 1930-1931].

Conclusion. Part-of-speech tagging is performed using such processes in text corpora: parsing, tokenization, lemmatization and stemming. The separation and creation of connections among the linguistic units, the transformation of their forms and stems occurr with the help of these phenomena. They are important data for modern studies that are based on the text corpora.

References

1. Дарчук Н. П. Корпусна лінгвістика: проблеми, методи, перспективи: [Навчальна програма для аспірантів (спеціальність 10.02.01 – українська мова)] / Дарчук Н.П. – Київ, 2013. – С. 1–10.

Darchuk N. P. Korpusna lingvistyka: problemy, metody, perpektyvy: [Navchalna prohrama dlia aspirantiv (spetsialnist 10.02.01 – ukrainska mova] [Corpus linguistics: problems, methods, perspectives: [Postgraduate curriculum (specialty 10.02.01 – Ukrainian language)] / N.P. Darchuk. – Kyiv, 2013. – P. 1-10. [in Ukrainian]

2. Жуковська В. В. Вступ до корпусної лінгвістики: [Навчальний посібник] / В. В. Жуковська. — Житомир: Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.

Zhukovska V. V. Vstup do korpusnoi lingvistyky [Navchalnyi posibnyk] [Introduction to Corpus Linguistics]: [Tutorial] / V. V. Zhukovska. – Zhytomyr: ZhDU them. I. Franco, 2013. – 142 p. [in Ukrainian]

3. Jurafsky D. Speech and Language Processing / D. Jurafsky, James H. Martin. – Retrieved from: https://web.stanford.edu/~jurafsky/slp3/

4. Бісікало О.В. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту / О. В. Бісікало, В. А. Висоцька // Вісник Національного університету "Львівська політехніка". Серія: Інформаційні системи та мережі: збірник наукових праць. – Л.: ЛНУ, 2015. – № 829. – С. 255–272.

Bisikalo O. V. Eksperymentalne doslidzhennia poshuku znachushchyh kliuchovyh sliv ukrainomovnoho kontentu [Experimental research of searching meaningful key words of Ukrainian-language content] / O. V. Bisikalo, V. A. Vysotskaya // Bulletin of the National University "Lviv Polytechnic". – Seriia: Informatsiini systemy ta merezhi [Series: Information systems and networks: a collection of scientific works]. – L.: LNU, 2015. – No. 829. – P. 255-272. [in Ukrainian]

5. Anjali G. Jivani. A Comparative Study of Stemming Algorithms / Anjali Jivani Ganesh // Department of Computer Science & Engineering. – India, Baroda: The Maharaja Sayajirao University of Baroda, 2016. – P. 1930-1928.

Yelyzaveta Timchenko
Vasyl' Stus Donetsk National University
Vinnytsia
Research Supervisor: I. H. Danyluk, PhD in Philology, Ass. Prof.
Language Advisor: V. I. Kalinichenko, PhD in Philology, Ass. Prof.

MACHINE TRANSLATION WITH THE USE OF DEEP LEARNING

Introduction. Deep learning is one of many methods of machine learning that is based on training the data characteristics. Nowadays when considering deep learning we can distinguish three of its historical stages. The first one is related to cybernetics and dates back to the 1940–1960's, when the theory of biological education was developed and the first models, including the perceptron, which allowed one neuron to be trained, were implemented. The second phase of the 1980–1990's period is associated with the connectionist approach when the backpropagation method was applied to train a neural network with one or two hidden layers. The third stage – deep learning – started around 2006.

Objective of the paper is to investigate the concept of machine translation and to discuss the translation procedure under consideration in the plane of the deep learning use.

Results of the research. Artificial neural networks are mathematical models built on the principle of the organization and operation of biological neural networks. Today there is a large number of varieties of artificial neural networks, aimed at solving various problems.

The tasks performed by artificial neural networks include:

1. Recognition (classification) of samples;
2. Approximation of multidimensional functions;
3. Clustering of samples;
4. Restoration of samples;
5. Reducing the dimension of data;
6. Forecast;
7. Filtration;
8. Management;
9. Identification of model parameters.

A recurrent neural network (RNN) is one of the architectures of deep learning. Such networks are successfully used in various fields, including machine translation.

Today RNNs called Long-Short-Term Memory (LSTM) networks, are often used. Such network was suggested by S. Hochreiter and J. Schmidhuber in 1997. It is a system that, unlike traditional recurrent neural networks, does not have problems with the 'disappearance of the gradient'. This is a complication that arises during the