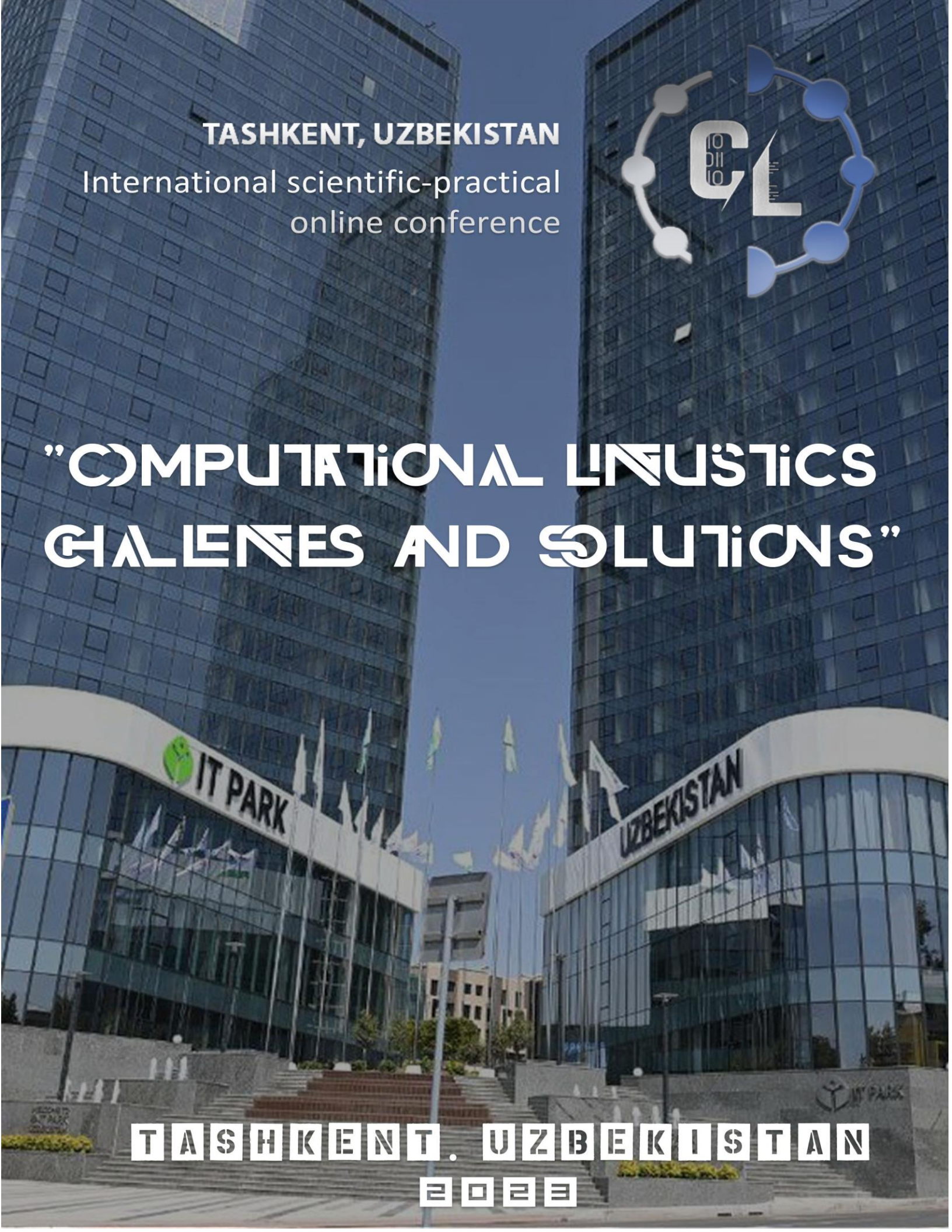


TASHKENT, UZBEKISTAN

International scientific-practical online conference

"COMPUTATIONAL LINGUISTICS CHALLENGES AND SOLUTIONS"

TASHKENT, UZBEKISTAN

2023

O‘ZBEKISTON RESPUBLIKASI
RAQAMLI TEXNOLOGIYALAR VAZIRLIGI

IT PARK

OOO “UNIVER-CLASS SYSTEM”

Kompyuter lingvistikasi: muammo va yechimlar (Компьютерная лингвистика: проблемы и решения, Computational linguistics and solutions) mavzusidagi xalqaro an’anaviy onlayn ilmiy-amaliy konferensiya materiallari to‘plami. Toshkent, 23-may, 2023-y.

TOSHKENT – 2023

 DOI: 10.36078/1688391447

MUNDARIJA

DOI: 10.36078/1688391447

DOI: 10.36078/1688391447

KORPUS BIRLIKLARINI TEGLASH ZARURATI VA AHAMIYATI

Manzura ABJALOVA
Filologiya fanlari doktori (DSc), dotsent
Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi
Toshkent davlat o‘zbek tili va adabiyoti universiteti
abjalovamanzura@navoiy-uni.uz
ORCID: 0000-0002-1927-2669

Annotatsiya. Lingvistik korpuslarni matnlar arxivi, elektron ensiklopediya va lug‘atlar tizimi, elektron kutubxona kabi matnli tizimlardan farqlaydigan xususiyatlardan biri — lingvistik izoh yoxud teg hisoblanadi. Korpus birliklarini teglash natijasida bir qancha qulay imkoniyatlar yuzaga keltiriladi. Maqolada shu haqida so‘z boradi.

Kalit so‘zlar: teg; annotatsiya; lingvistik izoh; korpus; metama’lumot.

Abstract. One of the characteristics that distinguishes linguistic corpora from text systems, such as an archive of texts, a system of electronic encyclopedias and dictionaries, an electronic library, is a linguistic annotation or tag. Tagging of corpus units provides a number of convenient features. This will be discussed in the article.

Keywords: tag; annotation; linguistic comment; Corpus; metadata.

Turli lingvistik topshiriq / amalni bajarish uchun matnga lingvistik va ekstralingvistik qo‘shimcha ma’lumot bilan ishlov berilgan bo‘lishi lozim. Buning uchun mavjud matnning komponentlariga maxsus izoh (masalan, so‘z turkumi haqida ma’lumot) berilishi zarur hisoblanadi. Bu izoh matn ***tegi*** (belgi, ishora) yoki ***annotatsiyasi*** (rus. *razmetka*, ing. *tag*) deb ataladi. Matn birliklarini izohlash esa *annotatsiyalash, teglash* yoki oddiygina, *lingvistik izohlash* (*разметка, tagging, annotation, mark-up*) deyiladi. Bunday ma’lumotlarning eng oddiy misoli — so‘z turkumlari tegi hisoblanadi. Bu shunday ko‘rinishi mumkin:

Samiya xalqaro tanlovga yaqinda boradi

Izohlaymiz:

Samiya $^{(ot)}$ *xalqaro* $^{(sifat)}$ *tanlovga* $^{(ot)}$ *yaqinda* $^{(ravish)}$ *boradi* $^{(fe'l)}$.

Men o‘g‘lim bilan faxrlanaman.

Izohlaymiz:

Men $^{(olmosh)}$ *o‘g‘lim* $^{(ot)}$ *bilan* $^{(ko‘makchi)}$ *faxrlanaman* $^{(fe'l)}$.

Tarixga nazar. 80-yillarda SGML (Standard Generalized Markup Language) deb nomlangan elektron matnlarni belgilash standarti qabul qilingan. U tipografiya sanoatida

DOI: 10.36078/1688391447

ishlab chiqilgan, ammo tez fursatda boshqa sohalarga tarqaldi. SGMLning maqsadi shundaki, turli matn protsessorlarida yozilgan hujjatlarni tahrirlash, tahlil qilish va ularning istalgancha o‘zgartirish mumkin bo‘ladi.

SGML teglar tushunchasini kiritdi. Teglar (ing. tags) bu – matndagi xizmat belgilari, matnning o‘zi haqidagi ma’lumotni o‘z ichiga oladi. Har bir holat maxsus teglarni belgilash va shu bilan SGML tilining dialektlarini yaratishi mumkin. SGML belgilash tili – tillarning “konstruktori”. U juda murakkab til sanaladi va juda kam ishlatiladi. Ammo uning asosida HTML va XML kabi taniqli belgilash tillari yaratildi.

Matnli ma’lumotlarni (korpuslarni) teglash uchun bir necha universitetlar matnlarning qaysi parametrlarini teglash kerakligini tavsiflovchi tizimni maxsus ishlab chiqdi. Ushbu tizim XMLdan foydalanadi va *Text Encoding Initiative Guidelines* (*TEI Guidelines*)[1] deb nomlanadi. Bu kodlash, teglash va indeksirlash mumkin bo‘lgan matnlarning turli xil xususiyatlarining ro‘yxati hisoblanadi. Masalan, tizim matndagi turli xil tuzatishlar, iqtiboslar, qisqartmalar, atoqli otlar, initsial, akronimlar, chet el so‘zlari va boshqalarni sanab o‘tadi. Hozirgi vaqtda korpuslarni yaratish bo‘yicha deyarli barcha loyihalar (shu jumladan, Britaniya Milliy korpusi) TEI tavsiyalariga u yoki bu tarzda amal qilishga harakat qilmoqda [Кутузов А.Б.].

An’anaga muvofiq, teglar burchakli qavslarda juft, ya’ni ochish va yopish holatida bo‘ladi. Masalan, <a> ochuvchi teg, yopuvchi teg. Yopish tegi ochish tegida berilgan xabarning tugaganligini bildiradi. Fikrga yuqoridagi gap bilan misol keltiramiz:

<pron>Men</pron></N>o‘g‘lim<N><prep>bilan</prep>
<V>faxrlanaman</V>.

Ayon bo‘lganidek, gap boshidagi LB *<pron>Men</pron>* olmoshi ekanligi haqida belgi berilgan.

Yoki yana

<ds>Samimiyatni o‘zingizga bezak qilib oling</ds> – deydilar onajonim.

Ushbu gapdagi “*Samimiyatni o‘zingizga bezak qilib oling*” qismi *<ds>* va *</ds>* teglarida berilgan, bu teg ko‘chirma gap (direct speech – ds) ni anglatadi.

Og‘zaki nutq korpuslarida *<pause>* tegi toq holda qo‘llanilishi mumkin. Uning ochuvchi yoki yopuvchi ekanligi ahamiyatsiz bo‘ladi. Bu teg qo‘yilgan o‘rinda to‘xtam bo‘lganligini anglatadi.

Teglar qisqa belgi yoki ramzlardan iborat bo‘ladi. Masalan, *sifat – sif, fe’l – f, noun – N, verb – v* tarzida. Teglar foydalanuvchiga ko‘rinmaydi. Matnni *annotatsiyalangan*

[1] http://www.teic.org/Guidelines/index.xml.

DOI: 10.36078/1688391447

koʻrsatadigan dastur teglarni oʻziga xos qoidalarga muvofiq izohlaydi va foydalanuvchiga shu qoidalarga muvofiq shakllantirilgan matnni taqdim etadi.

Avtomatik annotatsiyalash / teglash. Katta hajmli korpuslarni teglash koʻp vaqt va mablagʻni talab qiladi. Shu bois XX arsning 70-yillarida annotatsiyalashni kompyuter orqali qilish loyihalari paydo boʻla boshladi. Shunda TAGGIT dasturi Braun korpusining 77 % soʻzining turkumlarini teglagan. Qolgan 23 % esa oʻn yil davomida qoʻlda teglangan. 80-yillarda CLAWS (Constituent Likelihood Automatic Word-tagging System) tizimining teglash koʻrsatkichi 95 % ga chiqdi. Unda ehtimollik nazariyasi tatbiq qilingan. Bu haqida quyida ma'lumot berildi. Bugungi kunda asosiy Yevropa tillarining soʻz turkumlarini avtomatik teglash (morfologik tahlil, word-class tagging) va gap boʻlaklarini avtomatik teglash (sintaktik tahlil, parsing) tizimlari ishlab chiqilgan. Bu imkoniyatlar Internet qidiruvi va mashina tarjimasida ham zarur sanaladi. Shuningdek, "Matnlarni avtomatik qayta ishlash" (http://www.aot.ru) nomi bilan rus tilini kompyuter texnologiyalari yordamida qayta ishlash imkoniyatlari yaratildi. Rossiya davlat gumanitar universiteti Lingvistika fakultetining bir guruh mutaxassislari ushbu tizimda rus, nemis va ingliz tillari uchun:

- grafematik (soʻzlarning chegarasini aniqlashtirish);
- morfologik (soʻz turkumlarini aniqlash);
- sintaktik (gap boʻlaklarini belgilash);
- semantik (soʻzlardagi semantik munosabatni aniqlash) modullar ishlab chiqilgan (Abjalova, 2020).

Umuman, matnlarni teglashning ikki turi mavjud: **metama'lumot** yozish va **lingvistik izoh**, ya'ni belgi biriktirish.

1. *Metama'lumot* (*metadata, metaizoh, metalingvistik ma'lumot, ekstralingvistik ma'lumot*) korpusga kiritilgan manba nomi, muallifi, yaratilgan vaqti, voqea joyi, uslubi, janri, shoir yashagan davr tilshunosligi nuqtayi nazaridan (tilning ma'lum davrga xosligi) ilmiy yondashishni: shoirning badiiy soʻz qoʻllash uslubi va mahorati (xalq ogʻzaki ijodi namunalari: maqol, matal, topishmoq, turli qochirimlar, hikmatli soʻzlar, iboralar)ni chuqur oʻrganish, janrning ijtimoiy-tarbiyaviy ahamiyatini yoritish, janrlarni tahlil qilish jarayonida mazmunan qaysi auditoriya yoshiga mosligini aniqlash mumkin.

2. *Lingvistik izoh (lingvistik belgi)*da matndagi birliklarni lingvistik xususiyatlariga koʻra tasniflashda soʻz shaklning grammatik ma'no (*fonetik, leksik, morfologik va sintaktik*) birliklarining nutqni shakllantiruvchi umumiy ma'nolari, qoʻshimchalar (*prefiks, suffiks, kompozitsiya),* unda har bir soʻzning turkumi va shu turkumga tegishli kategoriyalari (fe'l, ot, sifatlar va boshqalar), arxaizm, istorizm soʻzlar haqidagi ma'lumotlar *teglar* koʻrinishida korpusga qoʻshilishi korpusning oʻqishliligini oshirish

DOI: 10.36078/1688391447

bilan birga foydalanuvchilarga korpusda maxsus qidiruvni ham amalga oshirishda ulkan ma’lumotlar xazinasi vazifasini o‘taydi.

Lingvistik izohlash tiplari:

1) ***morfologik*** (part-of-speech tagging yoki POS-tagging) — so‘z turkumlarini teglash;

2) ***sintaktik tahlil yoki parsing*** — leksik birliklar va turli sintaktik tuzilmalar orasidagi sintaktik munosabatlarni tavsiflash;

3) ***semantik*** — berilgan so‘z yoki ibora tegishli bo‘lgan semantik toifalarga va uning ma’nosini aniqlaydigan kichikroq toifalarga ko‘ra tavsiflash;

4) ***anaforik*** — referent aloqalarini, masalan, olmoshlar bilan bog‘lanishni izohlash;

5) ***prosodik*** — urg‘u va intonatsiyani tavsiflaydigan teglardan foydalanadi;

6) ***diskurs*** — og‘zaki nutq korpusida pauza, takrorlash, eslatma va hokazolarni ko‘rsatish uchun matn maxsus izohlanadi;

7) ***stilistik*** — leksik birlikning uslubiy xoslanishini ko‘rsatadi.

Ushbu lingvistik izohlarni amalga oshirishda quyidagi asosiy prinsiplarga rioya qilish maqsadga muvofiq:

✓ Nazariy jihatdan neytral (an’anaviy) izohlash sxemasi — har bir korpusga o‘ziga xos izohlash sxemalari, ya’ni elementlaridan foydalanib, chigalliklarni yuzaga keltirgandan ko‘ra, yirik lingvistik korpuslarni teglash spesifikasidagi elementlarni umumfoydalanish uchun asos qilib olinishi o‘zbek tili korpuslarining jahon talabidagi zamonaviy korpus deya e’tirof etilishiga asos bo‘ladi va o‘z o‘rnida bunday korpus standart korpus vazifasini o‘taydi. Modomiki, ko‘pchilikka ma’lum izohlash sxemasidan foydalanilmasdan mualliflik nazariyasi yaratib olinsa, korpusdan foydalanuvchi izohlash tizimini chuqurroq o‘rganib chiqishga majbur bo‘ladi. Tabiiyki, bunday ortiqcha izlanish foydalanuvchiga ma’qul kelmaydi.

✓ Lingvistik tushunchalarning umumiy qabul qilingan tizimi — lingvistik korpuslarning dunyo miqyosida ahamiyatga ega bo‘lishi uchun teglar xalqaro belgi va ramzlardan foydalanish o‘rinli bo‘ladi. Bu, asosan, lingvodidaktikada ahamiyatli hisoblanib, til o‘rganish va o‘rgatish jarayonini yanada qulaylashtiradi.

✓ Parametrlarni samarali kiritish — juda katta miqdordagi lingvistik birliklarni to‘g‘ri teglashda inson omili va yarim avtomat jarayoni ishonchli hisoblanadi. Buning uchun fidoyi mutaxassislar jamoasining sermahsul mehnati talab qilinadi.

✓ Xalqaro standartlarga rioya qilish — teglashning TEI xalqaro standartiga rioya etish ulkan tajribaga tayanish hisoblanadi.

Quyida lingvistik izohlashning morfologik turiga kengroq to‘xtalamiz.

 DOI: 10.36078/1688391447

Morfologik belgilashning asosiy birligi bu – belgilar zanjiri sifatida tushuniladigan va odatda, oddiy so‘z shaklga teng bo‘lgan matniy shakl yoki *token*. Bunday ramziy holat kompyuter dasturining ishi uchun zarur hisoblanadi. Matndagi tokenlarni alohidalash jarayoni *tokenizatsiya* deyiladi. Ayrim adabiyotlarda *grafematik tahlil* ham deb beriladi. Shuni ta’kidlash joizki, token nafaqat korpus lingvistikasining, balki boshqa sohalarga ham tegishli termin bo‘lib, asosan, probel (bo‘shliq)dan probelgacha bo‘lgan belgi token hisoblanadi (Копотев, 2003. 33-37). Korpus obyekti matn, eng kichik birligi so‘z (so‘z shakl) hisoblangani uchun korpus lingvistikasida token sifatida so‘z va so‘z shakllar e’tiborga olinadi.

Korpusda tokenizatsiya bilan birga yana bir muhim jarayon bor. Bu bosqich korpusga kiritilgan ma’lumotlarni qayta ishlash uchun muhim sanaladi. *Lemmatizatsiya* deb nomlangan bu jarayonda so‘zs hakl boshlang‘ich shakli avtomatik tarzda aniqlanadi, boshlang‘ich shaklning o‘zi *lemma* deyiladi. Lemmatizatsiya flektiv tillar uchun juda muhim. Sababi lemmatizatsiya jarayonida fleksiyaga uchragan so‘zning asosi tiklanadi. Masalan, *copies → copy, bases → basis, oxen → ox; вижу → видить, иду → идти, пальчик → палец.*

Ma’lumki, deyarli barcha mamlakat xalq ta’limi maskanlarida boshlang‘ich sinflaridanoq o‘quvchi gapni o‘qiydi va undagi ot, sifat, son, fe’l, ravish, olmosh so‘z turkumlarini aniqlaydi. Korpus lingvistikasida bu so‘zning turkumlik tegi hisoblanadi. So‘z turkumlarini teglash (ingliz tilida bu **part-of-speech tagging** (**POS tagging** yoki **PoS tagging** yoxud **POST**), rus tilida **частеречная разметка** deyiladi) matnni avtomatik qayta ishlash bosqichi bo‘lib, uning vazifasi matnda qo‘llangan so‘z (shakl)larning turkumi va grammatik xususiyatlarini aniqlash hisoblanadi. Shu vazifasi bilan POS-tagging matnni avtomatik tahlil qilishning dastlabki bosqichlaridan biri sanaladi.

Korpus bazasidagi birliklarni annotatsiyalash uchun so‘z turkumlarini teglash (POS-tagging, Part of Speech tagging – so‘z turkumini anglatuvchi belgi qo‘yish) muhim ahamiyat kasb etadi. STni bunday belgilash zarurati kompyuterning omonimlarni ajratmasligi bilan bog‘lanadi.

Yaratilgan korpuslarning bunday xususiyatlari va o‘ziga xosliklari ular bilan ishlash imkoniyati hamda korpuslarning ahamiyatini oshiradi.

Korpus lingvistikasida so‘z turkumlarini teglash, grammatik kategoriyalarni teglash (Asiryan, A.K. 2017.) va so‘zlarni toifalashda noaniqliklarni bartaraf etish uchun so‘zni faqat uning lug‘atdagi shakliga asoslanib emas, balki matn (jumla)dagi ifodasi bo‘yicha uning turkumlik tegi va jumla (xatboshi, ibora)da boshqa so‘zlar bilan birikish imkoniyatini hisobga olish muhim sanaladi. Gap bo‘laklari teglarini identifikatsiyalash bir muncha qiyin jarayon. Sababi o‘zbek tilidagi jamiki so‘zlarni universal holda 12 turkum

DOI: 10.36078/1688391447

doirasida teglash imkoniyati yo‘q. So‘z uning jumla tarkibida reallashish holati va N-gramm (Abjalova, 2020. 73-77) so‘zlarning semantik valentligiga binoan polifunksional bo‘lishi mumkin. Masalan: “*Shifoxonaga bemorni keltirishdi*” va “*Shifoxonaga bemor odamni keltirishdi*” jumlalarining 1-sida *bemor* so‘zi turkumlik belgisi (kim? so‘rog‘iga javob berayotgan tushum kelishigidagi so‘z)ga ko‘ra ot turkumi, 2-jumlada esa (qanday? so‘rog‘iga javob beryapti) sifat turkumi vazifasidagi so‘z hisoblanadi. O‘zbek tili izohli lug‘atida mavjud 11 000 o‘zlashma so‘zlardan 66 ta xuddi shunday polifunksional so‘zlar aniqlandi.

So‘z turkumlarini teglash (STT) uchun lingvistik bazada so‘zlar va ularning turkumlari ko‘rsatilgan ro‘yxatning kiritilishi kifoya emas. Yuqoridagi so‘z turkumini aniqlash holatidagi kabi izchillikning yo‘qolishi yoxud bir shaklga ega polifunksional, omonim (Abjalova, 2020. 73-77) yoki ko‘p manoli so‘zlarning gapda ifodalagan turkumini topish hatto mutaxassis tilshunosni ham fikr yuritishga, izlanishga undaydi. Shuningdek, o‘zbek tilidagi ko‘pgina so‘zlar muayyan turkumga mansubligi aniqlanmagan. Har bir tabiiy tilda mavjud bunday muammolar e’tiborga olinib STTda bir necha usullarga tayaniladi.

Aksariyat hollarda so‘z turkumlarini teglashda quyidagi usul (metod, algoritm)larga asoslaniladi:

1) qoidalarga asoslangan usul;

2) stoxastik (yoxud statistik) usul.

Xulosa tarzida shuni aytish lozimki, korpus birliklarini teglash:

1) korpusdagi statistik ma’lumotlarni aniq olish;

2) korpus yordamida til o‘rganish va o‘rgatish;

3) korpusda leksik birliklarning sememalarini aniqlash;

4) kontekstda qo‘llanilgan omonim birliklarni aniqlash;

5) ko‘p ma’noli va polifunksional so‘zlar semantikasini ochish imkonini beradi. Shu bois korpus birliklarni teglash muhim ahamiyat kasb etadi.

Foydalanilgan adabiyotlar

1. Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn]: monografiya. – Toshkent, 2020. – 176 b. ISBN 978-9943-6939-0-6.

2. Abjalova, M.A., Yuldashev A. 2021. Methods for Determining Homonyms In Homonymy And Linguistic Systems. ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 11, Issue 2, February. Impact Factor: SJIF 2021 = 7.492 (https://saarj.com). ISSN: 2249-7137

 DOI: 10.36078/1688391447

3. Asiryan, A.K. 2017. Сравнение инструментов морфологической разметки. *Научный взгляд в будущее, 10.30888/2415-7538.2017-07-01-027.*

4. Копотев М. В., Мустайоки А. Принципы создания Хельсинского аннотированного корпуса русских текстов (ХАНКО) в сети интернет // Научно-техническая информация. Сер. 2: Информационные системы и процессы. № 6: Корпусная лингвистика в России. 2003. – С. 33-37.

5. Кутузов А.Б. Курс «Корпусная лингвистика». – 45 с. Лицензия: http://creativecommons.org/licenses/by-sa/3.0/

 DOI: 10.36078/1688391447

BRIEFLY ABOUT THE MORPHOLOGICAL ANALYZER

Muqaddas Tursunaliyevna ABDURAXMANOVA
Doktorant
ToshDO‘TAU
f.f.n., dotsent
MParfi2005@yandex.ru

Muxlisa Parmonovna OLIMOVA
2-kurs talabasi
O‘zMU
olimovamuxlisa556@gmail.com

Abstract. The article discusses the fact that the issue of creating linguistics software, in particular, morphological analyzers, is one of today’s urgent tasks in automatic editing and analysis in the world linguistics. Research work on the development of morfoanalyzers in world and Uzbek linguistics is discussed. The fact that existing morphological analysis algorithms will play a major rule in future work is revealed with the help of examples. Also, the reasons for the deficiencies in the system that need to be corrected are explained.

Key words: morphological analyzer; corpus; trop; algorithm; lemma; token; morphotactic opportunity.

Аннотация. В статье рассматривается, задача в создании лингвистического программного обеспечения, в частности, морфологических анализаторов, которая является одной из актуальных задач АТТ на сегодняшний день в мировой лингвистике. Обсуждается научно-исследовательская работа по развитию морфоанализаторов в узбекском языкознании. На примерах показано, что существующие алгоритмы морфологического анализа играют большую роль в дальнейшей работе. Также объясняются причины и недостатки в системе, которые необходимо исправить.

Ключевые слова: морфологический анализатор; корпус; троп; лемма; токен; морфотактическая возможность.

Linguistic support of spell checker, morphological and syntactic analysis systems has been developed in the direction of automatic analysis and editing in world linguistics. They are the basis for creating fast and economical systems that analyze and edit language material in natural language processing. These systems improve the quality of work of machine translations, electronic dictionaries, and parsers (algorithms for automatic

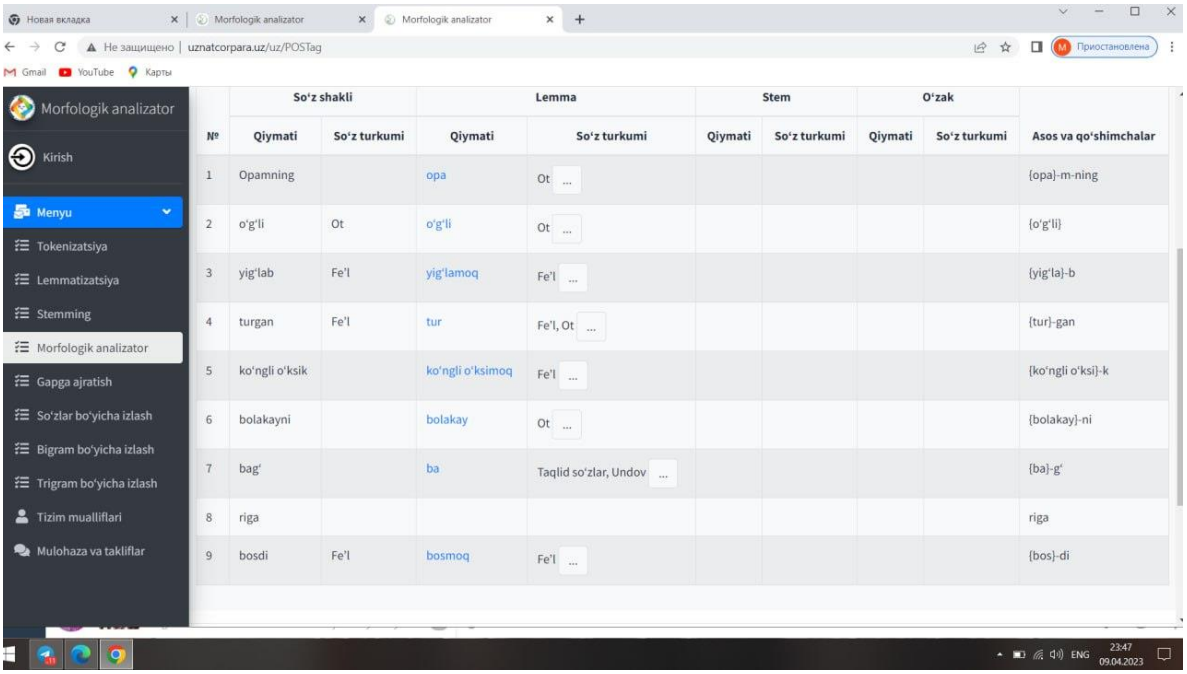
DOI: 10.36078/1688391447

analysis of certain texts). After all, in today's globalization process, it is necessary to work quickly and efficiently on large volumes of information texts in order to keep up with world computer linguistics. In this sense, automatic editing programs are being developed that they are aimed at editing texts in languages such as Russian, English, Italian, German, Azerbaijani, Arabic, Turkish, Kyrgyz.

A morphological analyzer is descriptive software that compares words and word forms with the structure in the dictionary. It determines their basis and grammatical forms. Until today, the issue of creating morphological analyzers has been studied and significant practical work has been done by several scientists in world linguistics. In particular, the "Using morphological analysis to teach vocabulary in English and French classes " program [Sullivan, 2004], created by Constance O'Sullivan and Charlotte Ebel in 2004, the scientific work of "Morphological analyzer in the development of bilingual dictionary (Kokborok - English) [Sarkar, 2015], the development of morphological analysis software for Malay and Tamil languages [Rajeev, 2011], the Arabic morphological analyzer formed by Arabic linguists Y. Jaafar and K. Bouzoubaar can be a proof of our thoughts. Also, morphological today serious work is being done on the creation of analyzers. As an example of this, we can say the algorithms that are emerging within the framework of Turkic languages. For instance, Adnan Öztürel, Tolga Kayadelen and Isın Demirsahin presented implements which is comprehensive model and an open source morpho analyzer of Turkish [4]. This morphological analysis software can be used to develop the model of Turkish language because it covers all aspects of morphology and syntax. Not only the Turkish language, but also the Turkic languages have an agglutinative character, therefore, the morphotactic condition is considered the main factor in the creation of a morpho analyzer in these languages, because what suffixes are added after the word stem is based on a certain grammatical rule. Specially, the morphotactic of the Uzbek language the possibility is as follows: *Prefix + word stem + word-forming suffix + lexical formative suffix + syntactic formative suffix*. It should be noted that pure Turkic languages do not have prefixes, but in this place, the suffixes such as *be-*, *no-*, *bo-*, *ba-*, *bar-*, *kam-* are implied and these suffixes came from Persian-Tajik languages. Linguistics is also an exact science and should not accept exceptions. In our opinion, the division of Uzbek words such as muzlatkich and o'chirg'ich into the forms muzlat+gich and o'chir+g'ich justifies our views. Based on these rules, morphological analyzers have been constructed for Uzbek language in the corpus of the Uzbek language [5] and http://uznatcorpara.uz. The software will recognize each token in the given text analyzes and forms information such as categories and different grammatical meanings. To determine the category of the search word, it based on morphological, semantic, syntactic rules. Linguistic analyzers like this in natural

 DOI: 10.36078/1688391447

language processing, improving the quality of information processing, automatic processing skills in Google translate, creating different corpora in the Uzbek language, automatic processing of corpus units, establishing a search system in the corpus are important. However, currently available morphological analyzers are not perfect. For being high efficiency of morphological analysis in some aspects of natural languages, homonymy, paronyms, polysemy, phraseologies, tropes are obstacle. Solutions to these obstacles are being sought. Technologies, for example, are being developed to eliminate errors caused by homonyms. For this, it is necessary to classify each of the similar words, i.e. comparing them with a lemma - part of a sentence and a set of morphological features. The method of collocation [Abdurahmanova, 2021] (semantic circle) can be used for distinguishing homonyms in the Uzbek language. This method was analyzed in the example of the homonymic string of the word *oshiq*. The correct interpretation of all the



№	So'z shakli		Lemma		Stem		O'zak		
	Qiymati	So'z turkumi	Qiymati	So'z turkumi	Qiymati	So'z turkumi	Qiymati	So'z turkumi	Asos va qo'shimchalar
1	Opamning		opa	Ot					{opa}-m-ning
2	o'g'li	Ot	o'g'li	Ot					{o'g'li}
3	yig'lab	Fe'l	yig'lamoq	Fe'l					{yig'la}-b
4	turgan	Fe'l	tur	Fe'l, Ot					{tur}-gan
5	ko'ngli o'ksik		ko'ngli o'ksimoq	Fe'l					{ko'ngli o'ksi}-k
6	bolakayni		bolakay	Ot					{bolakay}-ni
7	bag'		ba	Taqlid so'zlar, Undov					{ba}-g'
8	riga								riga
9	bosdi	Fe'l	bosmoq	Fe'l					{bos}-di

meanings of the word *oshiq* means that the semantic environment serves to increase productivity in the morphological analyzer. It should be said that the semantic circle can be a reliable solution to the issue of polysemy and tropes. Also, this is another factor that affects the correct operation of the morph analyzer in the process of analysis, words are not correctly divided from the morphological point of view. This is mainly due to phonetic changes which appears when a suffix is added to a word. As a result, the meaning of the words is misinterpreted, some words appear as two words so that they are divided into syllables, errors are observed in the classification into word categories:

DOI: 10.36078/1688391447

As can be seen from this interface, the sentence *Opamning o‘g‘li yig‘lab turgan ko‘ngli o‘ksik bolakayni bag‘riga bosdi* is misinterpreted according to the aspects of form value, classification, division into bases and suffixes. The errors occurred in places related to homonyms (*turgan - tur*), division into bases and suffixes (*bag‘ir, o‘g‘li*) and morphological cases (*o‘g‘li, yig‘lab, bag‘riga*).

In conclusion, we can say that many morpho analyzers have been developed in the experience of world linguistics, but the rules of one language do not apply to another language, so one of the current tasks is to create linguistic analysis methods that can correctly analyze all the rules and exceptional cases based on the internal capabilities of the Uzbek language. Frequency words of the language lexicon, a set of rules related to processing words in the language, must be inputted the morphological analyzer. In addition, the morphological analyzer should be designed as a flexible setting that can generate predictions about word analysis and such an algorithm avoids faults similar to the homonyms problem. By the way, in order to obtain accurate analysis results, it is necessary to create clear patterns of phonetic phenomena that occur within the framework of bases and additions.

References

1. Constance O'Sullivan, Charlotte Ebel. Using morphological analysis to teach vocabulary in English and French classes. Teachers as Scholars Institute Princeton University. July, 2004. Marguerite Browning, Professor.
2. Partha Sarkar, Dr. Bipul Syam Purkayastha. Morphological analyzer in the development of bilingual dictionary (Kokborok - English) - An analysis for appropriate method and approach. // International journal of engineering and technology (IJIET). Vol.4, Issue 10, April 2015 - 98-103pp.
3. Rajeev R. R., Jisha P. Jayan, Dr. Rajendran S. Morphological analyser and morphological generator for Malayalam – Tamil machine translation. // International journal of computer applications. Vol.13 – No.8, January 2011.
4. http://aclanthology.org/W19-3110.pdf
5. https://uzbekcorpus.uz
6. Abdurahmanova M. T., Rahmanova A. A. O‘zbek tili omonimlari uchun milliy teglar lug‘ati. – T.: Lesson press, 2021. – 9-11 b
7. Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer. // 14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent.

DOI: 10.36078/1688391447

8. Elov B., Hamraeva Sh., Elova D. Morfologik analizatorni yaratish usullari. O‘zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(1). 67-87.
9. Abdullayeva O. Morphological annotation system in the corpus of internet information texts in Uzbek language. // 7 th International Conference on Computer Science and Engieering, 14-16 September 2022, Turkey. 160-164.

 DOI: 10.36078/1688391447

РАЗРАБОТКА ПАРАЛЛЕЛЬНОГО КОРПУСА ПЬЕС А. П. ЧЕХОВА ДЛЯ ПРОЕКТА CHEKHOV DIGITAL

Марина Андреевна БУТУСОВА
магистрант, 1 курс
Институт филологии, журналистики и межкультурной коммуникации
Южный федеральный университет
mbutusova@sfedu.ru

Научный руководитель
Северина Елена Михайловна
доктор философских наук, профессор
emkovalenko@sfedu.ru

Аннотация. В работе рассмотрены особенности разработки русско-английского параллельного корпуса пьес А. П. Чехова для проекта Chekhov Digital. Проведён сравнительно-сопоставительный лексический и стилистический анализ текстов пьес писателя и их переводов на английский язык с помощью цифровых методов.

Ключевые слова: параллельный корпус; количественный анализ; частотный словарь; облако слов; сентимент-анализ.

Данное исследование является частью проекта по разработке параллельного корпуса пьес А. П. Чехова для семантического издания Chekhov Digital [1], которое включает в себя тексты Полного собрания сочинений и писем А. П. Чехова в 30 томах (далее ПССиП) [3], размеченные с опорой на стандарт Text Encoding Initiative. Такое цифровое представление текстов делает их машиночитаемыми и открывает новые возможности для исследовательской работы. Издание уже содержит тексты произведений 1-10 томов Полного собрания сочинений и тексты Полного собрания писем писателя, в которых в настоящее время ведётся разметка именованных сущностей. [2]

Среди следующих задач проекта Chekhov Digital – создание семантической разметки для драматического наследия писателя. Поскольку пьесы А. П. Чехова пользуются популярностью во всём мире, активно переводятся, издаются и ставятся на сцене, было принято решение создать параллельный корпус, который включает в себя как оригинальные тексты на русском языке, так и переведённые на английском. Тексты оригинальных пьес и их переводов были изучены с помощью цифровых

DOI: 10.36078/1688391447

методов. Цель исследования - изучить пьесы цифровыми методами, определить, насколько переводы адекватны оригиналу с лексической и стилистической точек зрения. Объектом исследования являются переводы пьес А. П. Чехова как материал для параллельного корпуса, а предметом – их лексическое и стилистическое соответствие оригиналу.

Из семнадцати пьес, написанных А. П. Чеховым в корпус будут включены десять: «На большой дороге» (1884 г.), «Медведь» (1888 г.), «Предложение» (1888 г.), «Свадьба» (1889 г.), «Трагик поневоле» (1889 г.), «Юбилей» (1892 г.), «Три сестры» (1900 г.), «Вишневый сад» (1903 г.) (все в переводе Джулиуса Уэста, опубликованные в 2005 г.) [8], «Дядя Ваня» (1896 г.) (в переводе Мэриан Фэлл, опубликованном в 1999 г.) [11] и «Чайка» (1896 г.) (в переводе Дэвида Виджера, опубликованном в 2006 г.) [10]. Пьесы отбирались по критерию наличия их переводов на английский язык, находящихся в открытом доступе. Отобранные тексты удовлетворяют этому критерию: пьесы на русском языке взяты с сайта Фундаментальной электронной библиотеки, которая позволяет использовать тексты с указанием источника [3]; пьесы на английском языке предоставлены ресурсом Project Gutenberg [8, 10, 11], который также использует открытую лицензию.

С помощью цифровых методов определён объём оригинальных текстов и их переводов. Выявлено, что тексты переводов содержат больше токенов, чем тексты оригиналов (например, пьеса «На большой дороге» на русском языке содержит 5404 токена, а на английском – 7968 токенов), при этом корпус был разделен на две условные группы: в первую входят ранние работы Чехова, написанные с 1884 по 1892 годы и включающие от 1678 до 5404 токенов в оригинале; во вторую группу входят поздние пьесы автора, опубликованные с 1895 по 1903 годы – от 12713 до 16536 токенов соответственно. Такое разительное различие в объёме – важная характеристика, которую требуется учесть в дальнейших исследованиях.

Были составлены два частотных словаря для текстов оригиналов и переводов. Среди самых частотных слов выявлено множество существенных пересечений: *see – видеть/понимать*, *know – знать*, *man – человек*, *love – любовь/любить*, *life – жизнь*, *hand – рука*, *want – хотеть*. Перечисленные слова встречаются как в русскоязычном частотном словаре, так и в англоязычном, при этом номера их рангов и количество употреблений отличаются. Например, слово *человек* имеет 9 ранг в русском списке и 12 в английском. Предшествующие ему слова после вычета пересечений представляют такой список: *like, one, see, get, look, take, yes*; *это, весь, Андреевна, мочь*. При этом в основном в него входят служебные, многозначные или имеющие омонимы слова. Например, *like* может переводиться двумя способами: *нравиться* и

DOI: 10.36078/1688391447

как. Следовательно, в русском языке мы увидим две отдельные позиции, а в английском – одну с суммированной графой частотности. Таким образом, мы видим некоторые различия в частотных словах в текстах оригиналов и переводов, что должно быть учтено при создании корпуса.

В ходе исследования были проанализированы лексические корреляции между пьесами. При подготовке данных использовался метод TF-IDF, который позволил нормализовать частоты и уменьшить погрешность, возникающую из-за значительной разницы в объёме текстов [6]. Корреляционная карта оригинальных текстов показала многообразие языка Чехова: самый высокий коэффициент корреляции обнаружен между текстами «Чайка» и «Три сестры» и составляет 0.5, что всё ещё считается слабой корреляцией. В текстах перевода самый высокий коэффициент корреляции равен 0.6 и найден между пьесами «Чайка» и «Три сестры», а также «Три сестры» и «Вишнёвый сад». Английские тексты показывают большее лексическое однообразие: если в оригинале самый распространённый коэффициент – 0.3, то в переводе – 0.4. Примечательно также, что в переводах передана важная черта языка Чехова: его ранние пьесы более лексически неоднородны, чем поздние, а пьеса «Предложение» (1888 г.) значительно отличается от всех прочих. Эта же тенденция сохранена и в переводах. Корреляционная карта показала, что лексическая сложность оригиналов пьес передана достаточно полно, хотя тексты переводов лексически более однородны.

Стилометрический анализ проведён с помощью библиотеки Stylo языка программирования R [5]. Этот инструмент позволяет выявить стилистические особенности текста, которые не зависят от его тематики, и разделить корпус текстов на кластеры [7]. Корпус нелемматизированных пьес на русском языке был разделен на два кластера: в первый вошли ранние пьесы, написанные до 1892 года; во второй – поздние, написанные после 1895 г. Эти две группы заметно отличаются и по объёму, поэтому, для исключения возможности влияния размера текста на результаты мы разделили поздние пьесы на акты и провели анализ повторно. Анализ результатов стилометрии показал, что поздние пьесы А. П. Чехова стилистически отличаются от ранних.

Подобный анализ был проведён для текстов перевода. Кластеризация нелемматизированных текстов продемонстрировала, что в первый кластер попали все пьесы, переведённые Джулиусом Уэстом (важно отметить, что поздние тексты в его переводе всё же выделены в отдельную подгруппу); во втором кластере находятся тексты двух других переводчиков – Мэриан Фэлл и Дэвида Виджера. Таким образом были выявлены стилистические различия языка переводчиков.

 DOI: 10.36078/1688391447

Стилометрический анализ лемматизированных текстов пьес показал различия с результатами для нелемматизированных текстов только для текстов переводов. Непредобработанные «сырые» тексты на английском языке отражают переводческий стиль, а предобработанные – стиль Чехова, демонстрируя результаты, схожие с результатами для русскоязычных текстов: ранние пьесы определяются в один кластер, а поздние – в другой. Таким образом, тексты перевода вполне соответствуют стилистике А. П. Чехова, хотя и отражают стилевые особенности письма переводчиков.

Для корпуса текстов был проведён сентимент-анализ [9], который позволяет определить эмоциональную окраску текста. Модель наивного Байеса была обучена на коротких текстах и затем использована для определения окраски выбранных пьес [4]. Оригинальные тексты по большей части оказались негативными: к позитивным были отнесены только два лемматизированных текста («Предложение» и «Свадьба») или три, когда рассматривались исключительно реплики персонажей без ремарок автора (к упомянутым ранее двум добавилась пьеса «Юбилей»). Исследование англоязычных текстов показало ещё больше негативных результатов (одинаковых вне зависимости от наличия или отсутствия лемматизации или исключения ремарок автора): к позитивным текстам отнесена только пьеса «Свадьба». Результаты сентимент-анализа показали, что тексты переводов оцениваются как чуть более негативные, чем тексты оригиналов, однако всё же близки к последним.

В ходе исследования в исходном корпусе текстов с помощью цифровых методов были выделены две группы: в первую вошли ранние малоизвестные и небольшие по объёму пьесы, а во вторую – поздние произведения А. П. Чехова, которые до сих пор ставятся на сценах по всему миру. Эти две группы отличаются не только объёмом и периодом публикации, но также лексическим разнообразием и стилистическими особенностями. Описанные результаты, полученные на материале оригиналов, оказываются верными и для переводов. Выявленная специфика переводов требует дальнейших исследований и должна быть учтена при разработке параллельного корпуса для проекта Chekhov Digital.

Использованная литература

1. Проект Chekhov Digital. URL: https://chekhov-digital.sfedu.ru/. (Дата обращения: 12.05.2023).
2. Северина Е.М., Бонч-Осмоловская А.А., Кудин А.М. Цифровые филологические практики: проект "Chekhov Digital" // Актуальные проблемы

 DOI: 10.36078/1688391447

филологии и педагогической лингвистики. 2022. №2. С. 153-165. DOI: 10.29025/2079-6021-2022-2-153-165.
3. Чехов А. П. Полное собрание сочинений и писем: В 30 т. Сочинения: В 18 т. М.: Наука, 1974—1982. URL: http://feb-web.ru/feb/chekhov/default.asp?/feb/chekhov/texts/che-te02.html (дата обращения: 10.05.2023)
4. Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 241–249). Association for Computational Linguistics.
5. Lagutina K. V. , Manakhova A. M., “Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th–21st Centuries”, Modeling and analysis of information systems, vol. 27, no. 3, pp. 330-343, 2020.
6. Lomakina L.S., Lomakin D.V., Subbotin A.N. Naïve Bayes modification for text Streams classification // Sciences of Europe. 2016. №6-2 (6). URL: https://cyberleninka.ru/article/n/na-ve-bayes-modification-for-text-streams-classification (дата обращения: 14.05.2023).
7. Mukhin Mikhail Yu., Mukhin Nikolai Yu. Idiostyle Characteristics of Lexical Compatibility in the 19th-Century Prose: Ural Stylometric Project // Журнал СФУ. Гуманитарные науки. 2020. №12. URL: https://cyberleninka.ru/article/n/idiostyle-characteristics-of-lexical-compatibility-in-the-19th-century-prose-ural-stylometric-project (дата обращения: 10.05.2023).
8. Plays by Anton Chekhov, Second Series by Anton Pavlovich Chekhov // Project Gutenberg. URL: https://www.gutenberg.org/ebooks/7986. (Дата обращения: 10.05.2023).
9. S. Smetanin, "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," in IEEE Access, vol. 8, pp. 110693-110719, 2020, doi: 10.1109/ACCESS.2020.3002215.
10. The Sea-Gull by Anton Pavlovich Chekhov // Project Gutenberg. URL: https://www.gutenberg.org/ebooks/1754. (Дата обращения: 12.05.2023).
11. Uncle Vanya: Scenes from Country Life in Four Acts by Anton Pavlovich Chekhov // Project Gutenberg. URL: https://www.gutenberg.org/ebooks/1756. (Дата обращения: 12.05.2023).

 DOI: 10.36078/1688391447

THE IMPORTANCE OF MEDIA-LINGUISTICS IN SCIENCE

Dilfuzaxon Shukhrat qizi BOLTAYEVA
Teacher
Uzbekistan State World Languages University
dilfuzaboltayeva92@gmail.com

Abstract. Media linguistics, as an integral part of linguistics, is aimed at studying the functioning of language in the media sphere. The number of scientific researches caused by the influence of modern technical mass media on language changes and aimed at the formation of specialized forms of linguistic directions has increased.

Key words: socio-informational processes (SIP); media-linguistics; media-text.

The mass media play a large role each within the lifetime of society and within the development of language. Since the half of the 20th century, the expansion of mass media has been progressing: the quantity of ancient media is increasing, technology is up, and therefore the development of the net helps to make a unified data area.

These socio-informational processes (SIP) have an effect on not solely the lifetime of society, however conjointly the functioning of the language. Mass communication has become one in every of the foremost intensive areas of speech consumption these days. the whole volume of texts distributed by the media will increase each hour, that contributes to a rise in interest during this space on the a part of researchers. This truth caused the emergence of such a science as media linguistics.

According to T. G. Dobrosklonskaya, the term "media linguistics" combines 2 elements "media" and" linguistics", which suggests that the topic of this science is" the study of language functioning within the sphere of mass communication ". this suggests that media-linguistics explores a particular sphere of auditory perception the language of mass media. In media linguistics, all ways of text process square measure used: beginning with ways of system analysis, and ending with logical, empirical, and linguistics ways. Texts for study of mass media uses the techniques of psychological feature and important linguistics, useful SIP-leaves, etc. Combining the ways of various Sciences permits North American country to create a comprehensive approach to the study of media texts.

The main class of media linguistics is media text. this idea is predicated on a mix of units of 2 series — verbal and media. Media text may be a complicated and multi-level development.

According to the tactic of media text production, it is author's or collective. this relies on what number individuals participated within the development of the media text, still as

 DOI: 10.36078/1688391447

whether or not the authorship is indicated once displaying the ultimate product. AN example of AN author's text is any material that contains a sign of authorship, like a news article. AN example of a collegial text may be a write up.

The sort of creation and therefore the form of replica of the media text ought to be thought-about along. several texts that were originally supposed to achieve the buyer verbally reach them in writing and contrariwise. for instance, AN interview during a magazine. Initially, it absolutely was oral in type, however reached the buyer in writing. Or reading a commentator's text. within the sort of creation, it's a written communication, and within the sort of replica — oral.

The channel plays a vital role in describing the media text. A channel may be a suggests that of mass communication at intervals that a media text is formed and conjointly functions. every mass communication medium contains a range of characteristic media options that considerably have an effect on the linguistic and format properties of a selected text. for instance, graphic style or illustrations during a news article.

The useful and genre affiliation of a media text is a vital element for its description. However, this element is unstable, since there's a relentless genre movement within the field of mass communication. There square measure four forms of media texts:

1. news;
2. data Analytics;
3. text-essay;
4. advertising.

The thematic dominant is another vital element for describing a media text. when analyzing this element, we will say that a stable system of topics or media topics has been fashioned at intervals the media, which incorporates news, political, cultural and alternative topics. It ought to be noted that a particular country might have its own stable media tropics, for instance, in the UK, coverage of the non-public lifetime of the house is such a stable media tropic.

A wide vary of ways square measure wont to study media texts. the foremost common ways square measure content analysis, discursive analysis, linguistics ways.

In conclusion, it ought to be noted that media linguistics may be a science that studies the functioning of language within the media sphere, that appeared comparatively recently. It originated at the junction of the 2 Sciences, which suggests that it carries their inherent features: it uses the bottom of linguistic analysis, on the one hand, and on the opposite — is incorporated into the overall system of Medialogy, that deals with the study of media. Its main class is media text, that may be a complicated and sophisticated development that carries variety of characteristic options.

DOI: 10.36078/1688391447

References

1. Dobrosklonskaya, T. G. Medialinguistics: a systematic approach to learning the language of the media: modern English media speech:study. allowance / T. G. Dobrosklonskaya. - M .: Flinta: Nauka, 2008 .-- 264 p.

2. Sour cream with. I, Media text in the cultural system: Dynamic processes in the language and style of journalism at the end of the 20th century /Smetanina S.I. - Mikhailov V.A., 2002 .-- 382 p.

3. Dobrosklonskaya TG .. Journal: Medialinguistics. Issue 3. Speech genres in mass media, 2011. - p. 17-21

 DOI: 10.36078/1688391447

СПЕЦИФИКА ПЕРЕВОДОВ ХУДОЖЕСТВЕННОЙ ПРОЗЫ А. П. ЧЕХОВА НА НЕМЕЦКИЙ ЯЗЫК: ЦИФРОВОЙ ПОДХОД

Никита Александрович ФЁДОРОВ
магистрант, 1 курс
Институт филологии, журналистики
и межкультурной коммуникации
Южный федеральный университет
Российская Федерация
nfyodorov@sfedu.ru

Елена Михайловна СЕВЕРИНА
Научный руководитель
доктор философских наук, профессор
emkovalenko@sfedu.ru

Аннотация. В работе рассматриваются вопросы разработки параллельного корпуса текстов на русском и немецком языках для семантического издания Chekhov Digital. Цифровые методы использованы для изучения особенностей перевода текстов А. П. Чехова на немецкий язык. Проведен сравнительный анализ результатов компьютерного анализа текстов произведений писателя и их переводов на немецкий язык.

Ключевые слова: частотный анализ; корреляционный анализ; стилометрия; Chekhov Digital; параллельный корпус.

В настоящее время цифровые методы исследования текстовых данных становятся ключевой частью гуманитарной науки. Одним из актуальных направлений работы в этой области является подготовка изданий литературных произведений в цифровом формате для дальнейших исследований с использованием как традиционных методов филологического и лингвистического анализа, так и методов автоматической обработки текстов. Кроме того, создание цифровых изданий литературных произведений отдельных авторов способствует систематизации их творческого наследия. Интерес представляет включение в такого рода издания не только оригинальных текстов авторов, но и переводов этих произведений, т. е. создание параллельных корпусов литературных произведений. Современные филологи все чаще работают над цифровыми (семантическими) изданиями текстов, представляющими собой «упорядоченные и универсально распознаваемые структуры данных» [5].

DOI: 10.36078/1688391447

Создавая параллельные корпусы текстов, оснащенные семантической разметкой, исследователи стремятся к автоматизации и универсализации структуры этой разметки. Как правило, соотнесение текстов в параллельных корпусах производится по предложениям, однако для оптимизации разметки в семантическом издании необходимо обратить особое внимание на значимую лексику размечаемых текстов и их структурные особенности: наиболее частотные слова, частеречные характеристики и т.п. [3, с. 289].

Проект Chekhov Digital [4] – это семантическое издание корпуса текстов произведений А. П. Чехова [6] на основе формата стандарта TEI/XML, в состав которого будет включен параллельный корпус семантически размеченных текстов писателя на русском и немецком языках. В связи с этим представляется актуальным рассмотрение характерных особенностей произведений А. П. Чехова и их переводов на немецкий язык через призму цифровых методов филологического исследования. Полученные результаты использованы при разработке параллельного корпуса для проекта Chekhov Digital.

В качестве материала исследования было выбрано 80 рассказов и повестей А. П. Чехова, а также их переводы, опубликованные на сайте «Projekt Gutenberg» [8] и распространяемые свободно по лицензии Creative Commons Attribution Non-Commercial (CC BY-NC). Источником материалов на русском языке послужило электронное научное издание произведений А. П. Чехова - ЭНИ «ЧЕХОВ» на платформе Фундаментальной электронной библиотеки (ФЭБ) [7]. Для исследования использованы следующие цифровые методы: лексический частотный анализ (построение векторных моделей), корреляционный анализ, стилометрический анализ.

Векторное представление текстов (Vector Space Model) – это математическая модель представления текстов, в которой каждому документу сопоставлен вектор, выражающий его смысл [1]. Такое представление позволяет легко сравнивать слова, искать похожие, проводить классификацию, кластеризацию и многое другое. При таком подходе принадлежность документа к классу определяется словами, тексты из одного класса содержат много схожих слов, также важна частота встречаемости слова в тексте.

Для предварительного анализа были выбраны пять текстов рассказов А. П. Чехова на немецком и русском языках: «Агафья», «Дом с мезонином», «Счастье», «Поцелуй» и «Из огня да в полымя», и созданы количественные модели оригинальных и переводных текстов с помощью мешка слов и TF-IDF, в результате был построен частотный словарь для каждого корпуса текстов. TF-IDF –

DOI: 10.36078/1688391447

статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции. Построенная на основе такой меры модель позволяет оценить схожесть текстов [9]. Анализ модели позволил выявить наиболее частотные слова в корпусах текстов, основные темы и действующих лиц в текстах.

Необходимо отметить сходство мотивов в оригинальных текстах и их переводах. Как правило, высокочастотные слова в модели оригинала встречались также и в модели его перевода, но частотность его могла достаточно сильно отличаться. Например, имена главных героев: *Савка – Ssawka, Агафья – Agafja, Лида – Lida, Женя – Schenja*; некоторые существительные: *лицо – gesicht, человек – mensch, жизнь – leben*; глаголы: *быть – sein, говорить – sagen* и др. – являются высокочастотными в оригинальных текстах, а в переводах эти элементы в форме прямого соответствия встречаются гораздо реже. В переводах часто употребляемые слова заменялись переводчиками на синонимы, к чему норма немецкого языка, как правило, склоняет носителей (большое количество составных слов-композит, стремление к аналитичности). Кроме того, по частотному словарю достаточно легко понять основную тему рассказа, мотивы, главных действующих лиц, например, частотный словарь рассказа «Счастье» содержит следующие слова: *Санька, Пантелей, солдат, старик, клад, зарывать, курган, талисман*.

Корреляционный анализ выбранных текстов А. П. Чехова показал, что лексический состав оригинальных рассказов значительно различается в рамках рассмотренных произведений: коэффициент корреляции не превышает отметку в 0.6 в парах «Поцелуй» – «Дом с мезонином» и «Поцелуй» – «Счастье». Наименьший показатель (0.4) – у пары «Из огня да в полымя» – «Агафья», т.е. эти тексты наименее схожи с точки зрения лексического состава. Корреляционный анализ текстов переводов показал, что лексический состав переводов рассказов Чехова достаточно схож в рамках рассмотренных произведений: минимальные показатели – у соотношения текстов «Agafja» и «Aus dem Regen in die Traufe», «Glueck» и «Aus dem Regen in die Traufe», «Kuss» и «Aus dem Regen in die Traufe», хотя коэффициент корреляции значительно выше (0.8). Примечательно, что во всех этих парах рассказ «Aus dem Regen in die Traufe» наиболее отличен от других с точки зрения лексики (хоть это различие и небольшое – различие в коэффициентах корреляции 0.2).

Коэффициент корреляции других текстов между собой («Kuss» – «Agafja», «Haus mit dem Mezzanin» – «Agafja», «Haus mit dem Mezzanin» – «Kuss», «Glueck» –

 DOI: 10.36078/1688391447

«Kuss», «Haus mit dem Mezzanin» – «Kuss» и т.д.) составляют 0.9, что показывает их сильную корреляцию друг с другом. Это может свидетельствовать о стандартности и универсальности языка переводчика рассказов Чехова (Alexander Eliasberg) и лексической схожести, по крайней мере, переводов этих пяти произведений.

Полученные данные свидетельствуют о том, что в отобранных рассказах А. П. Чехова лексика разнообразнее, чем в их переводах, выполненных Александром Элиасбергом.

Кластерный анализ - многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Кластерный анализ для выбранных текстов произведений писателя был проведен с помощью библиотеки Stylo языка программирования R, в которой реализован стилометрический подход - основанный на статистике метод определения авторского стиля, который позволяет «выявить различия между ранними и поздними работами одного писателя, долю участия другого человека (например, редактора) в работе над текстом и даже установить пол написавшего» [2].

Результаты стилометрического анализа всего корпуса имеющихся текстов – на русском и немецком языках, оказались чрезвычайно сложными и неинформативными в связи с большим количеством исходных данных. Поэтому тексты были объединены по годам их написания, анализ был проведен только для оригинальных текстов. В этом случае результат получился более информативным: весь корпус текстов был разделен на два больших кластера - восьмидесятые годы девятнадцатого столетия (включая ранние девяностые) и поздние девяностые, в каждом из которых были выделены еще более мелкие кластеры. Всего было автоматически выделено одиннадцать кластеров. Наиболее стилистически связанные между собой годы - это 1886 и 1887, а также 1883 и 1885 в первом кластере; 1898 и 1896 – во втором.

Проведенный количественный анализ позволил охарактеризовать лексические и грамматические особенности текстов, выявить лексические различия немецких переводов произведений А. П. Чехова от оригинальных текстов: переводчики часто обращаются к синонимам и перифразам, избегая повторяющихся слов; при этом тексты переводов в значительной мере коррелируют друг с другом, что свидетельствует о стандартности языка переводчика. Стилометрическое исследование позволило выявить схожесть хронологически разрозненных пластов творчества А. П. Чехова, что ставит перед исследователями новые вопросы, нуждающиеся в пристальном рассмотрении. Результаты проведенного исследования

DOI: 10.36078/1688391447

представляют особый интерес при подготовке материала для параллельного корпуса семантически размеченных текстов А. П. Чехова на русском и немецком языках проекта Chekhov Digital.

Использованная литература

1. Бондарчук Д. В. Векторная модель представления знаний на основе семантической близости термов //Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика, т. 6, №. 3, 2017, сс. 73-83. DOI: 10.14529/cmse170305.
2. Грачева А. А. Стилометрия: компьютерный метод атрибуции и стилистического анализа текстов // Современные СМИ в контексте информационных технологий: Сборник научных трудов 4-ой Всероссийской научно-практической конференции – 2018, Санкт-Петербург, 16 апреля 2018 года. СПб., 2019. С. 107-109.
3. Добровольский Д. О., Кретов А. А., Шаров С. А. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 263—296.
4. Проект Chekhov Digital. URL: http://chekhov-digital.sfedu.ru/ (дата обращения: 10.05.2023).
5. Северина Е.М., Ларионова М.Ч. – Новые филологические практики: семантическое издание текстов А. П. Чехова // Филология: научные исследования. 2020. № 10. DOI: 10.7256/2454-0749.2020.10.33970.
6. Чехов А. П. Полное собрание сочинений и писем: В 30 т. АН СССР. Ин-т мировой лит. им. А. М. Горького. М.: Наука. 1974-1983. URL: http://feb-web.ru/feb/chekhov/default.asp?/feb/chekhov/texts/che-te02.html (дата обращения: 10.05.2023).
7. ЭНИ «Чехов» / Фундаментальная электронная библиотека «Русская литература и фольклор». URL: http://feb-web.ru/feb/chekhov/default.asp (дата обращения: 03.05.2023).
8. Projekt Gutenberg. URL: https://www.projekt-gutenberg.org (Дата обращения: 03.05.2023).
9. TF-IDF с примерами кода: просто и понятно. URL: http://nlpx.net/archives/57 (дата обращения 03.05.2023).

DOI: 10.36078/1688391447

ПОСТРЕДАКТИРОВАНИЕ МАШИННОГО ПЕРЕВОДА СПЕЦИАЛЬНОГО ТЕКСТА: ПРОБЛЕМЫ И РЕШЕНИЯ

Лариса Анатольевна ГОРОХОВА
кандидат филологических наук, доцент
Пятигорский государственный университет
litoboika@mail.ru

Елена Анатольевна ДОЛУДЕНКО
кандидат филологических наук, доцент
Адыгейский государственный университет
ellen_313@mail.ru

Аннотация. Статья посвящена проблемам машинного перевода, в частности ошибкам нейронного машинного перевода, выявляемым и устраняемым на этапе ручного постредактирования. Корпус анализируемых ошибок был собран в процессе перевода текстов по программированию, дискретной математике, теории вероятностей и статистике с использованием технологии Smartcat. Делается вывод о роли постредактирования в обеспечении качества машинного перевода.

Ключевые слова: машинный перевод; нейронный машинный перевод; НМП; Smartcat; ошибки перевода; постредактирование.

Современные машинные переводчики используют нейросетевые алгоритмы, которые обучаются на больших объемах данных и позволяют переводить тексты более точно и гибко. Нейросеть не требует выделения фиксированных фраз и благодаря обучению умеет самостоятельно оптимально разбивать текст на составляющие, запоминая закономерности перевода. Все это привело к значительному росту качества работы машинных переводчиков. Качество текстов переводов, выполненных нейросетью, во многих случаях оказывается вполне приемлемым, однако это не значит, что они не нуждаются в постредактировании, т.е. нейросети еще не достигли того уровня совершенства, когда участие человека из переводческого процесса полностью исключается [2, 52].

Постредактирование машинного перевода – это процесс, в ходе которого профессиональный переводчик проверяет и исправляет текст, переведенный автоматически. Постредактирование машинного перевода может значительно повысить качество перевода и сделать последний более точным и читабельным. Более того, исследования показывают, что машинный перевод вкупе с постредактированием зачастую дают более качественный результат, чем человеческий перевод с нуля [3, 63]. Поэтому формирование навыков

 DOI: 10.36078/1688391447

постредактирования как части профессиональной компетенции переводчика должно стать важной составляющей обучения будущих специалистов в области перевода [1].

Постредактирование (post-editing) может быть первичным или поверхностным (Light PE), когда текст перевода соответствует по смыслу оригиналу, но содержит некоторые стилистические недочеты, и полным или вторичным (Full PE), в результате которого текст перевода становится идентичен человеческому переводу. Основной проблемой нейронного машинного перевода (НМП или NMT) является его неспособность использовать широкий контекст. Если сравнить недостатки предшественников НМП – статистического машинного перевода (SMT) и машинного перевода, основанного на правилах (RBMT), – с ошибками нейронного машинного перевода (см. Табл.1), то можно увидеть, что специфика ошибок последнего обусловлена большей частью тем, что системы НМП обучаются переводить по одному предложению, без оглядки на предыдущий контекст. Отсюда такие проблемы, как отсутствие единообразия при переводе терминологической лексики, выдуманные термины, фактологические неточности, противоречие здравому смыслу, необоснованные добавления и пропуски.

Таблица 1. Проблемы машинного перевода

SMT и SMT+RBMT	NMT
буквальный перевод непереведенные слова грамматические ошибки проблема полисемии добавления и пропуски non-translatables	проблема полисемии необоснованные добавления и пропуски неединообразие терминов выдуманные термины фактологические неточности противоречие здравому смыслу non-translatables

Автоматический (или ИИ) перевод — неотъемлемая часть инструментария переводчика, в частности технологии Smartcat. Она значительно облегчает и ускоряет работу над переводом однотипных текстов схожей тематики, т.к. позволяет заменить текст, извлеченный из исходного файла, на тот же текст на целевом языке. Вот что происходит при выборе автоматического перевода файла:
Smartcat извлекает текст из исходного файла и разбивает его на сегменты (обычно предложения) для более быстрой обработки.

Высокоточный искусственный интеллект выбирает лучшую технологию перевода (так называемый «движок») для нужд перевода, который, впрочем, всегда можно изменить в настройках в соответствии с потребностями переводчика.

В рамках реализации международного образовательного проекта в Пятигорском государственном университете, в рамках которого преподавание по

DOI: 10.36078/1688391447

направлению «Информационная безопасность» ведется на английском языке, мы столкнулись с необходимостью выполнения большого объема переводческих работ в очень сжатые сроки. Эти обстоятельства обусловили активное использование технологии Smartcat, в частности для перевода лекций по программированию, дискретной математике, теории вероятностей и статистике. Smartcat в бесплатной версии использует системы машинного перевода MT Deepl и Yandex. Несмотря на в целом удовлетворительное качество перевода, конечный продукт, безусловно, нуждался в постредактировании, что позволило нам выявить и классифицировать наиболее типичные ошибки и проблемы, с которыми не справился машинный переводчик.

К их числу относятся:

ошибки в выборе артикля (например, с опорой на микроконтекст ставится неопределенный артикль, хотя объект упоминался 2-3 сегмента назад; или же в тексте MT используется определенный артикль, несмотря на то, что речь идет о новом классе объектов):

ST	MT	PE
Первую вершину упорядоченной пары называют началом дуги, вторую – концом.	The first vertex of the ordered pair is called the start of the arc, the second is called the end.	The first vertex of an ordered pair is called the head of the arc, the second – the tail.
plt.show() #выводит график с точками	plt.show()#displays a graph with points	plt.show() #displays the plot with dots

ошибки в выборе глаголов (нарушаются клише, типичные для данной специальной тематики. MT часто переводит буквально, выбирая нетипичный глагол. Например, в программировании и дискретной математике описывает – describes (вместо defines), отдает – gives (вместо returns)):

ST	MT	PE
Имя функции, описанной в теле класса, сродни имени обычной переменной, описанной в теле класса.	The name of a function described in the class body is akin to the name of an ordinary variable described in the class body.	The name of a function defined in the class body is similar to the name of an ordinary variable defined in the class body.
Теперь для вычисления функции достаточно написать такое же	Now to calculate the function it is enough to	Now, evaluating the function is a matter of writing the same expression

DOI: 10.36078/1688391447

выражение, как для двух точек:	write the same expression as for the two points:	you would write with two points:

1) ошибки в употреблении предлогов, специфических для данной тематики (предложное управление в тексте оригинала автоматически транслируется в текст перевода, или же используется предлог, нетипичный для данного значения ЛЕ):

ST	MT	PE
Разностью двух событий A и B называется событие, которое состоится, если событие A произойдет, а событие B не произойдет.	The difference **between** two events A and B is the event that will take place if event A happens and event B does not happen.	The difference **of** two events A and B is the event that will happen if event A occurs, and event B does not occur.

2) ошибки в выборе местоимений 3 лица (одуш. / неодуш.) из-за отсутствия указания на объект в микроконтексте:

ST	MT	PE
Какова вероятность того, что он [*шар*] не белый?	What is the probability that **he** is not white?	What is the probability that **it** is not white?

3) отзеркаливание речевых недочетов, ошибок распознавания текста при сканировании и опечаток оригинала (например, паше – pashe (правильно name);

4) специфическая ошибка для Smartcat: там, где формула дается в тексте оригинала в формате графической вставки и обозначается внутри сегмента парой тегов, при переводе может нарушаться синтаксис.

Например, в приведенном ниже примере is стоит не после парных тегов, а до. В исходном тексте и тексте перевода это выглядит следующим образом:

ST	MT
Дисперсией σ^2 вариационного ряда называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической	The variance of the σ^2 variation series is called the arithmetic mean of the squares of the deviations of the variants from their arithmetic mean

5) ошибки из-за полисемии: в Smartcat MT учитывает указанную тематику и выбирает нужный вариант значения и перевода ЛЕ (например, специальный термин, а не общеупотребительное значение) – но только в том, случае, когда специальная

DOI: 10.36078/1688391447

тематика видна из микроконтекста. Иначе может быть выбрано совершенно неуместное значение и перевод:

ST	MT	PE
Какова вероятность того, что снова получится «логика»?	What are **the odds** of getting "logic" again?	What is **the probability** that the result will be the word LOGIC?
Монету подбрасывают 2500 раз. Какова вероятность того, что орел выпадет ровно 1200 раз?	What is the probability that **the eagle will fall** exactly 1200 times?	A coin is tossed 2500 times. What is the probability that **it will land heads** exactly 1200 times?
Какова вероятность, что при бросании пяти монет герб откроется более чем на двух?	What is the probability that a toss of five coins **will reveal the coat of arms** on more than two?	Five coins are tossed. What is the probability that more than two of them **will land heads**?
Мода и медиана.	**Fashion** and median.	**Mode** and median

6) буквализм в переводе терминов (особенно там, где тематика неясна из микроконтекста):

ST	MT	PE
Сумма вероятностей противоположных событий равна единице:	The union of the probabilities of **opposite** events is equal to one:	The probabilities of two **complementary** events add up to 1:

7) избыточность при переводе: MT сохраняет ЛЕ, которые являются частью клише в оригинале, но не нужны в соответствующих клише в переводе:

ST	MT	PE
Операции сложения и умножения событий обладают следующими свойствами:	The addition and multiplication **operations** of events have the following properties:	The addition and multiplication of events have the following properties:

8) транслитерация неизвестного MT слова:

ST	MT	PE
Квантили	Quantili.	Quantiles

В ряде случаев постредактор может столкнуться с целым кластером проблем в рамках одного сегмента.

DOI: 10.36078/1688391447

Так, в приведенном ниже примере можно выделить следующие недостатки машинной версии перевода:

- громоздкий синтаксис;
- «k-й степени отклонения» – имеется в виду математическая степень, т.е. power, а не degree of deviation;
- случайная величина – random variable, а не quantity (примечательно, что в других случаях тот же термин был переведен в версии МТ верно);
- «математическое ожидание» – в английском языке в теории вероятностей для обозначения данного понятия используются термины expectation, expected value и mean. Вариант mathematical expectation встречается редко и только в текстах, переведенных с русского оригинала.

ST	MT	PE
Центральным моментом k-го порядка случайной величины X называется математическое ожидание k-й степени отклонения случайной величины X от ее математического ожидания:	The central moment of the k-th order of a random quantity X is the mathematical expectation of the k-th degree of deviation of the random quantity X from its mathematical expectation:	The k-th central moment of a random variable X is the expected value of the k-power of the deviation of the random variable X from the mean.

Таким образом, мы видим, что большинство выделенных ошибок обусловлено неумением НМП опираться на контекст, выходящий за пределы одного предложения. Во всех представленных случаях вторичное постредактирование позволило выявить и исправить ошибки и неточности перевода, что еще раз говорит о превосходстве человеческого знания над искусственным интеллектом.

Использованная литература

1. Горохова Л.А., Долуденко Е.А. Машинный перевод и практика перевода: из опыта подготовки будущих лингвистов-переводчиков // Теория, практика и лингводидактика перевода. Сборник научных трудов. – Пятигорск: ПГУ, 2022. – С. 36-48.
2. Шевчук Е.В., Никифорова Ж.А. Постредактирование и типичные ошибки в автоматизированном переводе научно-публицистических текстов // Вопросы методики преподавания в вузе. 2021. №39. URL: https://cyberleninka.ru/article/n/postredaktirovanie-i-tipichnye-oshibki-v-avtomatizirovannom-perevode-nauchno-publitsisticheskih-tekstov (дата обращения: 13.05.2023).
3. Jia, Y. How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. JosTrans, (31), 2019. Pp. 60–86.

DOI: 10.36078/1688391447

ПОРТАЛ «ТЮРКСКАЯ МОРФЕМА»: ПРОБЛЕМЫ, РЕШЕНИЯ, ПЕРСПЕКТИВЫ

Айрат Рафизович ГАТИАТУЛЛИН
кандидат технических наук,
Институт прикладной семиотики
Академии наук
Республики Татарстан
ayrat.gatiatullin@gmail.com

Аннотация. В данной работе описывается портал “Тюркская морфема”, который создан, чтобы помочь в решении проблемы малоресурсности тюркских языков. Эта задача в настоящее время актуальна, в связи с тем, что по сравнению с индоевропейскими языками, практически все тюркские языки продолжают оставаться малоресурсными языками. При этом, лингвистические модели и ресурсы для других типов языков плохо приспособлены к структурно-функциональным особенностям тюркских языков. Создатели ресурсов для разных тюркских языков плохо контактируют друг с другом при создании лингвистических ресурсов разного типа. Одним из инструментов, который может помочь в решении данной проблемы должен стать лингвистический портал, который будет выполнять роль некоторого координационного центра и общей ресурсной базы в области компьютерной обработки тюркских языков. Данный портал должен служить информационно-справочной системой по тюркским языкам, библиотекой лингвистических стандартов по тюркским языкам, межпрограммным интерфейсом для связывания ранее созданных лингвистических ресурсов и программного обеспечения и еще целый ряд функций, которые описываются в данной работе.

Ключевые слова: лингвистический портал; тюркская морфема; граф знаний; малоресурсные языки.

Актуальность разработки портала «Тюркская морфема» определяется целым набором проблем, сформировавшихся в настоящее время вокруг программных разработок для работы с тюркскими языками. Рассмотрим эти проблемы:

1. Все тюркские языки (кроме турецкого) - малоресурсные языки;
2. Лингвистические модели и ресурсы для других типов языков плохо приспособлены к структурно-функциональным особенностям тюркских языков;
3. Отсутствие коллаборации в разработках для тюркских языков, как следствие отсутствие единой системы обозначений и тегов для разметки;

DOI: 10.36078/1688391447

4. Создание комплексных (интегральных) лингвистических ресурсов и моделей для тюркских языков позволит решать проблему более экономичными способами.

Рассмотрим эти аспекты актуальности.

Термин малоресурсные языки был введен еще в 2003 год Краувером [1].

Согласно его определению малоресурсные языки – это естественные языки, обладающие следующими свойствами:

1. недостаток своей системы письменности или устойчивой орфографии;

2. нехватка квалифицированных лингвистов и переводчиков для данного языка;

3. ограниченное распространение в сети Интернет;

4. нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфографических и фонетических транскрипций речи, словарей произношения и т. д.

Одной из причин нехватки, соответствующего программного обеспечения является то, что приведено во втором пункте актуальности – тюркские языки обладают большим набором структурно-функциональных особенностей, благодаря которым многие универсальные программные продукты, созданные для других языков к ним плохо применимы.

Рассмотрим некоторые из них, которые приведены в работах [2, 3]:

1. Агглютинативность,

2. Сингармонизм,

3. Отсутствие грамматического выделения единственного числа,

4. Отсутствие категории рода,

5. В тюркских языках редко встречаются исключения из правил,

6. Подавляющее большинство агглютинативных аффиксов однозначно.

7. Имена существительные обладают способностью выполнять функцию определения и др.

Еще одной причиной малоресурсности является отсутствие коллаборации в разработках для тюркских языков, и как следствие отсутствие единой системы обозначений и тегов для разметки. Если сравнить это с языками Европы, то у них создается множество лингвистических платформ и принимаются общеевропейские программы по созданию компьютерных ресурсов для языков Европы.

Например, такая платформа, как CLARIN (Common Language Resources and Technology Infrastructure) (www.clarin.eu) (Рис.1). У этой платформы есть филиалы в каждой стране Евросоюза. Так в Финляндии это www.kielipankki.fi, в Германии – clarin-d.net. В ЕС считается крайне важным вкладываться в информационные технологии для сохранения и развития европейских языков.

 DOI: 10.36078/1688391447

CLARIN является европейской исследовательской инфраструктурой, которая обеспечивает доступ к языковым ресурсам и технологиям для исследователей в области гуманитарных и социальных наук. Он поддерживает, как использование и изучение языковых данных в целом, так и повышение возможностей сравнительных исследований культурных и социальных явлений среди разных языков и дисциплин. Лингвистические платформы типа CLARIN, объединяются в проекты по созданию Европейского открытого научного облака, и интеграции сервисов, являющихся результатами междисциплинарного сотрудничества.

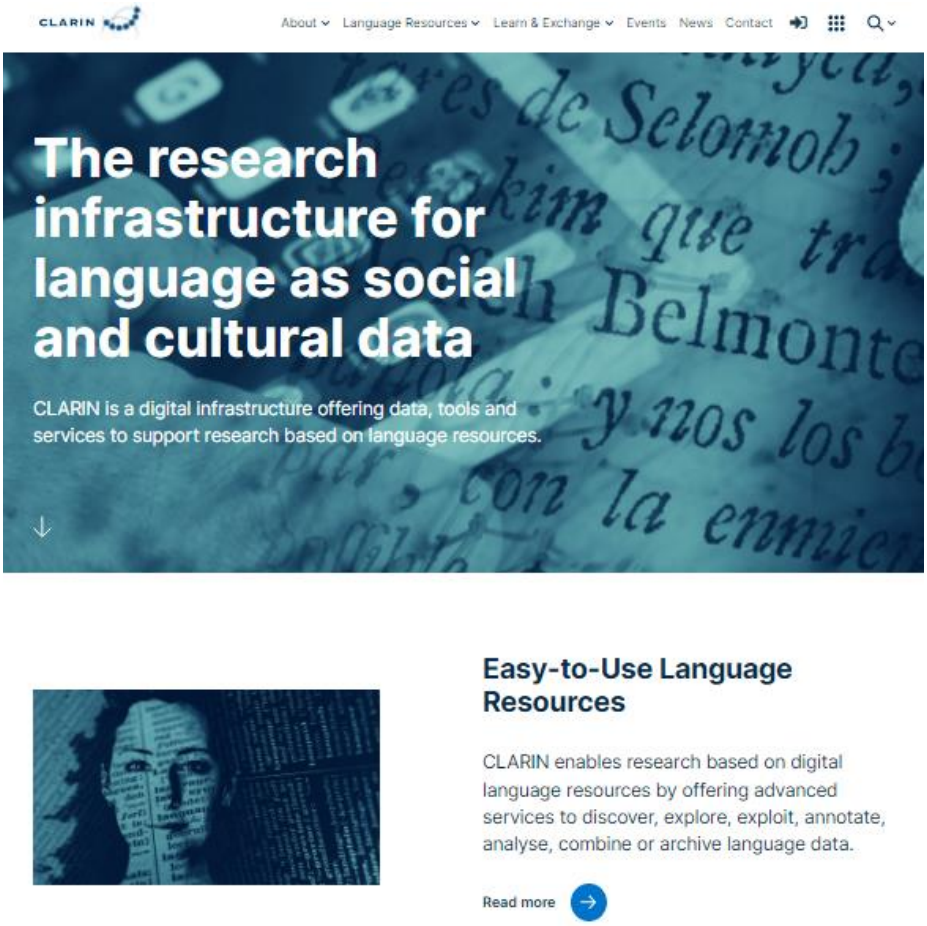


Рис.1. Интерфейс Clarin

Наличие таких платформ показывает, что подобные платформы необходимы и для тюркских языков и это было бы одним из способов решения проблемы малоресурсности. При этом лингвистические платформы для тюркских языков должны строиться с учетом структурно-функциональных особенностей тюркских языков.

DOI: 10.36078/1688391447

Также следует отметить, что степень малоресурсности тюркских языков сильно различается в зависимости от того, насколько сильно поддерживаются компьютерные разработки для этих языков государством, а также наличием квалифицированных специалистов. В связи с этим, усилия для создания языковых ресурсов и инструментов для языков с меньшим количеством ресурсов часто можно уменьшить, используя уже существующие ресурсы и инструменты для родственных, более богатых ресурсами языков.

Создание комплексных (интегральных) лингвистических ресурсов и моделей для тюркских языков позволит решать прикладные задачи более экономичным способом за счет взаимодополнения разработок между языками. Так, еще в 1988 году группой авторов В.Г.Гузев, Р.Г.Пиотровский, А.М.Щербак [4] была высказана идея, что для решения практических задач нужно создавать большой многоцелевой машинный фонд тюркских языков, который должен строиться, моделируя как общетюркскую языковую систему, так и систему каждого конкретного языка (функционирующего или мертвого, современного или древнего) со всеми ее инвентарными и структурными единицами, со всевозможными правилами знаковой репрезентации языковых единиц в речи, включая правила линейного развертывания речевых единиц.

Для осуществления такого объединения необходимо создать единую технологическую платформу, которая одновременно будет выполнять роль некоторого координационного центра и общей ресурсной базы в области компьютерной обработки тюркских языков.

Данная платформа должна реализовывать следующие функции:

1. Служить информационно-справочной системой по тюркским языкам;

2. Служить библиотекой лингвистических стандартов по тюркским языкам (термины, теги) для нового создаваемого программного обеспечения, что позволит обеспечить взаимную совместимость для разработок, создаваемых разными коллективами;

3. Служить межпрограммным интерфейсом для связывания ранее созданных лингвистических ресурсов и программного обеспечения, производящего обработку различных ресурсов на тюркских языках. Для этого в платформе должны храниться все таблицы соответствия системы обозначения посторонних разработок к системе обозначений платформы;

4. Служить библиотекой компьютерных моделей для описания структурно-функциональных особенностей тюркских языков и информационной базой для

 DOI: 10.36078/1688391447

лингвистических процессоров, которые производят компьютерную обработку тюркских языков;

5. Служить библиотекой программных модулей для создания прикладных программ, работающих с тюркскими языками;

6. Служить платформой с созданием виртуальной среды, в которой сторонние разработчики смогут реализовывать свои программные разработки по компьютерной обработке тюркских языков;

7. Служить виртуальной площадкой для общения специалистов тюркологов, которые пополняют лингвистическую базу данных и информационно-справочную систему по тюркским языкам.

В качестве реализации данных задач нами создается Интернет-портал «Тюркская морфема» [5], который представляет собой web-сайт (http://modmorph.turklang.net/ru/) (Рис.2.). Этот портал включает набор различных сервисов на базе лингвистических ресурсов с тюркскими языками и ориентирован на работу с тюркскими языками во всех аспектах: морфонологическом, морфологическом, синтаксическом, семантическом.

Данный портал постоянно пополняется новыми сервисами для работы с тюркскими языками. В перспективе внедрить в данный портал набор сервисов для работы диалектолога.

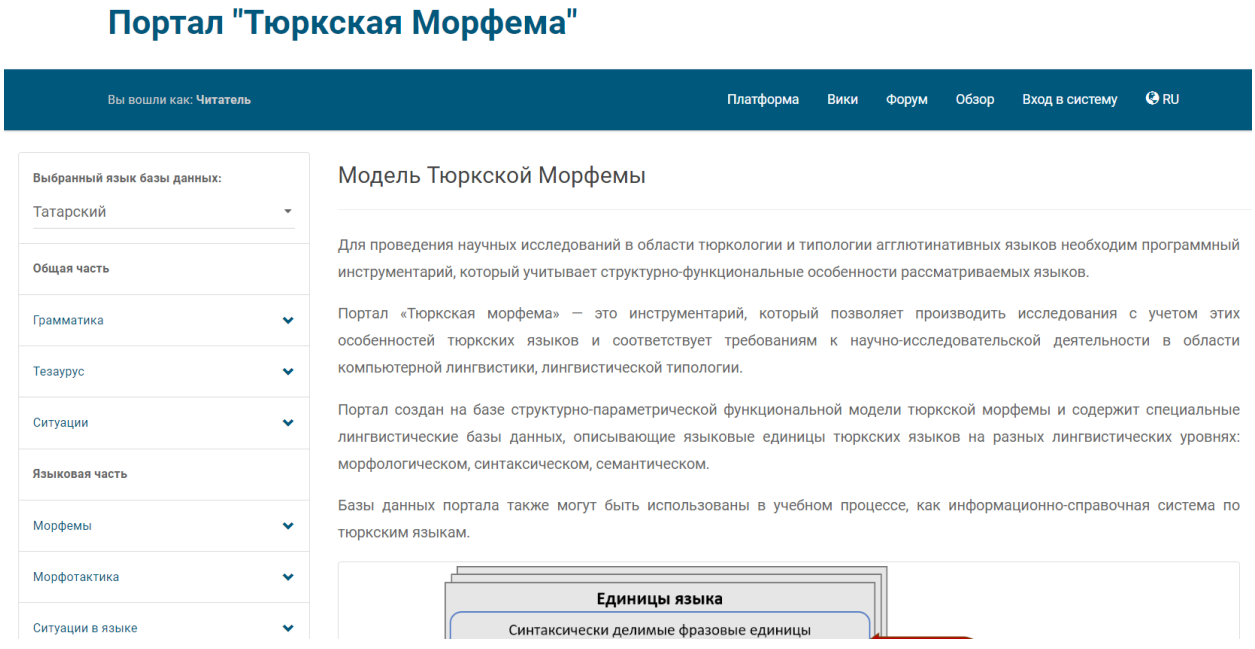


Рис.2. Фрагмент интерфейса портала “Тюркская морфема”

DOI: 10.36078/1688391447

Использованная литература

1. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Proc. International workshop on speech and computer SPECOM-2003. Moscow, Russia, 2003. Pp. 8-15.
2. Гузев В.Г. О некоторых экзотических особенностях тюркских языков («тюркские чудеса») // Актуальные проблемы мировой политики. Вып. 10 / под ред. Т.С.Немчиновой. СПб.: Изд-во С.-Петерб. ун-та, 2020. С. 231–245.
3. Suleymanov D.S. Natural possibilities of the Tatar morphology as a formal base of the NLP // In Proceedings of the First International Workshop “Computerisation of Natural Languages” (Varna, Sept. 3-7, 1999). –Sofia (Bulgaria): Information Services Plc, 1999. -P.113.
4. Гузев В.Г., Пиотровский Р.Г., Щербак А.М.О создании машинного фонда тюркских языков // Советская тюркология. 1988. №2. С.92-101.
5. Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B. (2020) About Turkic Morpheme Portal. CEUR Workshop Proceedings Institute for history, language and literature, Ufa scientific center, Russian Academy of Sciences Proceedings of TurkLang 2020, pp. 226-243.

DOI: 10.36078/1688391447

ФОРМИРОВАНИЕ РЕЧЕВОЙ БАЗЫ УЗБЕКСКОГО ЯЗЫКА

Сайёра Номазовна ИБРАГИМОВА
докторант, Научно-исследовательский институт
Развития цифровых технологий и
искусственного интеллекта
snibragimova@mail.ru
Малика Ильхамовна АБДУЛЛАЕВА
ассистент, Ташкентский университет
информационных технологий
имени Мухаммада аль-Хоразмий
malika.ilkhamovna@gmail.com

Аннотация. Научная работа посвящена формированию речевой базы узбекского языка для дальнейшего создания TTS системы для узбекского языка. Узбекский язык является одним из самых распространенных и важных тюркских языков, используемых в Узбекистане и других сопредельных регионах. Несмотря на это, существует недостаток в речевых базах и ресурсах для разработки и исследования речевых технологий на узбекском языке. Синтез речи по тексту на узбекском языке – предполагает создание фонетико-акустической базы данных. Для формирования такой базы необходимо определить принципы создания и обработки текстового и речевого корпуса для узбекского языка и особенности формирования на их основе речевой базы. Решению именно этих вопросов посвящена данная работа. Результаты исследования позволили разработать эффективный алгоритм формирования речевой базы узбекского языка, который может быть использован для различных задач, связанных с распознаванием речи, синтезом речи, автоматическим переводом и другими речевыми технологиями на узбекском языке. Этот алгоритм может быть основой для дальнейших исследований и разработок в области обработки узбекской речи и связанных технологий.

Ключевые слова: речевой сигнал; обработка сигнала; речевая база; нормализация текста; синтез; TTS система.

Синтез речи — это область искусственного интеллекта, которая занимается созданием искусственной речи с помощью компьютерных систем. Он позволяет генерировать человекоподобную речь из текстовых данных. Синтез речи находит широкое применение в различных сферах, включая технологии помощи людям с

DOI: 10.36078/1688391447

нарушениями речи, автоматизированные голосовые помощники, аудиокниги, объявления и трансляции, образовательные приложения и многое другое.

В последние годы синтез речи значительно развился благодаря прогрессу в области глубокого обучения и нейронных сетей. Современные TTS системы обладают высокой степенью натуральности и позволяют создавать речь с различными интонациями, акцентами и эмоциональными выражениями. Они способны воспроизводить не только отдельные слова и фразы, но и передавать сложные мелодические и ритмические аспекты речи, делая ее более естественной и понятной для слушателя.

Главным составляющим современных высококачественных TTS систем является речевая база с большим объемом. Формирование речевой базы для узбекского языка является сложной и многогранным процессом, который требует значительных усилий и ресурсов. Необходимо собрать обширный набор аудиозаписей на узбекском языке. Это может включать записи различных жанров речи, акцентов, диалектов и интонаций. Однако, поиск и сбор достаточного количества качественных записей может быть трудоемким и требовать сотрудничества с носителями языка и локальными сообществами. Речевая база должна быть достаточно разнообразной и покрывать различные стили и тематики.

Ключевой задачей при формировании речевой базы является достижение высокого качества и естественности синтезированной речи. Для этого необходимо обеспечить высокую четкость и чистоту аудиозаписей, минимизировать шумы и искажения. Кроме того, модели синтеза должны учитывать интонацию, ритм, акценты и другие просодические особенности узбекского языка, чтобы создать максимально естественный и понятный результат.

Речевая база TTS-систем - это база данных, состоящая из набора аудиоданных и соответствующих текстовых файлов [1]. Аудиофайлы состоят из образцов речевых элементов (звуков, слогов, слов, предложений), а текстовые файлы содержат транскрипции, соответствующие этим речевым элементам [2].

На сегодняшний день существует ряд речевых баз для мировых языков в открытом доступе. К сожалению, для узбекского языка нет такой речевой базы в открытом доступе пользования. Ниже приведены наиболее качественные и наиболее популярные из них [3-6].

Таблица 1. Список и характеристики речевых баз открытого доступа

Наименование речевой базы	**Количество дикторов**	**Язык**	**Объем речевой базы (часы)**

DOI: 10.36078/1688391447

LJ Speech	1	Английский	24
Libri-TTS	Многодикторная	Английский	585
RUSLAN	1 (мужчина)	Русский	29
NATASHA	1(женщина)	Русский	13
M-AILABS	Многодикторная	Многоязычная	1000

На практике качество синтетической речи зависит от качества речевой базы [7, 8]. Это особенно подтверждается для синтеза речи на основе конкатенативного метода и нейросетевых архитектур.

Однако, важно учитывать, что обучение нейронной сети для синтеза речи обычно требует значительного вычислительного ресурса. Обучение модели может занимать длительное время, особенно при использовании глубоких архитектур нейронных сетей. Это может потребовать использования мощных вычислительных систем или облачных платформ. После завершения обучения модели требуется оптимизация и настройка параметров для достижения наилучшего качества синтезированной речи. Этот процесс может потребовать проведения множества экспериментов и анализа результатов.

Распространенными методами формирования речевой базы для TTS систем являются:

- запись диктора, читающего заранее подготовленный текстовый, материл;
- запись диктора, произносящего спонтанную речь, нарративы и тд.

Оба метода являются затратными из-за необходимости вовлечения дополнительных специалистов и дикторов для предобработки текстовой информации и постобработки транскрипций и соответствующих аудио данных [8, 9]. Тем не менее первый метод обладает преимуществом с точки зрения возможности адаптировать разрабатываемую TTS систему под определенную область включив в речевую базу терминологии и предложения из этой области.

При создании речевой базы было уделено внимание к ряду особенностей, приведенных ниже, влияющих на его качество [10]:

1. Фонетическая структура текстовой информации, составляющий речевую базу и разнообразие лексики. Первым шагом к формированию грамотной речевой базы служит сбор и подготовка разнообразной текстовой информации. В первую очередь необходимо учитывать область, для которой разрабатывается TTS система, тем самым способствовать предрасположению системы для выбранной области. Это достигается путем включения в текст предложений, терминов и ключевых слов данного направления. Также необходима богатство фонетического и просодического охвата голосового корпуса.

DOI: 10.36078/1688391447

2. Профессиональность и грамотность диктора. Запись подготовленного текста должна осуществляться от носителя языка с хорошим, четким произношением, соответствующим установленным языковым стандартам. Речь диктора должна быть без ненужных разрывов, нелексических вокабул, фальшивых начал и наполнителей, такие как «э-э», «эм». Для грамотной записи одного часа аудио речевой базы, диктор тратит в среднем два и более часа времени. При записи аудиоданных на основе подготовленной текстовой информации от диктора также требуется учет его положения относительно микрофона. Речевой корпус для TTS систем должен состоять из мало отличающихся друг от друга аудиоданных с точки зрения интонации, громкости голоса, скорости произношения и т.д.

3. Состояние среды записи аудиоданных. Очень важным является постановка точных границ и задач для системы, которой будет служить речевая база. Для систем преобразования текста в речь требуется студийная среда, без посторонних шумов, разговоров или музыки.

4. Объем речевой базы. Общая длительность аудиоданных также меняется в зависимости от поставленной задачи. Минимальный объем речевой базы для TTS систем синтезирующей внятную и понятную речь составляет 25 часов аудиоданных с их транскрипцией. В речевую базу аудио записи попадают после жесткой многоэтапной фильтрации специалистами.

На рис.1 приведена форма создания грамотной речевой базы, состоящая из четырех главных этапов, по которым сформировалась речевая база для узбекского языка [11].

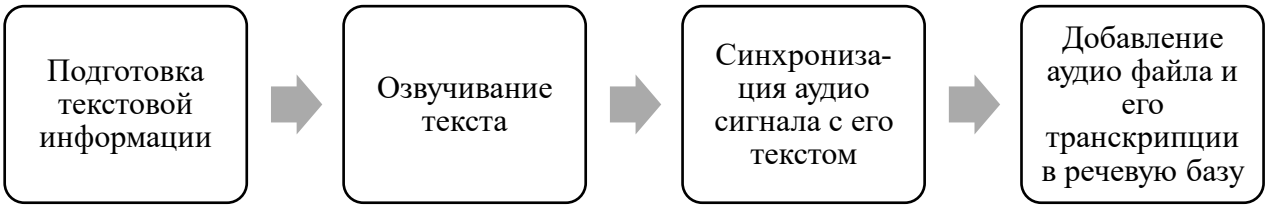


Рис.1. Этапы создания речевой базы

1-этап. Подготовка текстовой информации. Диктору предоставляется заранее подготовленный нормализованный материал в виде текста. Каждый текстовый документ проходит через ряд этапов обработки для его стандартизации. Текст, может состоять из слов, произношение которых обычно не встречается в словарях или лексиконах, таких как «БМТ», «ЎзХДП», «ТАТУ» и т.д. Такие слова называются нестандартными словами.

Нестандартные слова имеют несколько категорий:

DOI: 10.36078/1688391447

- числа, произношение которых меняется в зависимости от того, относятся ли они к валюте, времени, телефонным номерам, почтовым индексам;
- аббревиатуры, сокращения, акронимы;
- пунктуации;
- даты, время, единицы измерения и URL-ссылки.

Многие нестандартные слова также являются омографами, т.е. словами с одинаковой письменной формой, но разным произношением:

- IV, что может звучать по-разному: четыре (тўрт), четвертый (тўртинчи);
- трех- или четырехзначные числа, которые могут быть датами и обычными числами (например, 2040-йил, 2040 тонна).

Если первым компонентом обработки текстовой информации является нормализация, то вторым является просодический анализ. На данном этапе осуществляется анализ текста с точки зрения ударения и интонации и выполняется просодическая разметка нормализованного текста.

2-этап. Озвучивание текста. Этап представляет из себя обработку записанных аудио файлов для их стандартизации. Диктор производит запись аудиофайла с выразительным произношением. Формат аудиофайлов определен как .wav. Другие главные параметры аудиофайла установлены как 16 битовые моно-файлы, с частотой дискретизации 44100 Гц. В некоторых случаях при записи данные параметры могут быть не соблюдены диктором, по этой причине необходима стандартизация всех аудио файлов. Кроме выше приведенных установок, каждый аудио файл необходимо проверить с точки зрения правильности произнесенной речи и удаления речевых зон в случае если речь не соответствует тексту либо произнесен неправильно. Все аудио данные должны быть тщательно отфильтрованы во избежание избыточности аудио информации. Именно по этой причине удаляются зоны молчания.

3-этап. Синхронизация аудио сигнала с его текстом. Данный шаг является самым трудоёмким и важным для создания речевой базы. Этап выполняется экспертом, который тщательно осуществляет синхронизацию аудио и текстовых файлов.

Синхронизация аудио и его текста заключается в определении интервала произношения сегментированного текста в аудиофайле и его маркировке. Этот шаг требует особого подхода и создает необходимость создания специальных алгоритмов и состоит из 4 шагов:

1. Озвучивание текста;

DOI: 10.36078/1688391447

2. Вычисление информативных коэффициентов оригинальной и синтезированной речи;
3. Вычисление оптимального совпадения информативных коэффициентов;
4. Временное определение текущего текста в оригинальном ауди файле.

На шаге проверки файлов имеются аудио файл и его текстовая транскрипция. Проверка файлов представляет из себя механический процесс, когда эксперт, прослушивая каждый аудио файл проверяет точность совпадения произнесенных в нем звуков с транскрипцией текущего текста.

4-этап. Добавление аудио файла и его транскрипции в речевую базу. Проверенные на совпадаемость аудио файлы с их транскрипцией добавляются в речевую базу.

Эксперименты и результаты. Для формирования речевой базы узбекского языка была использована программная среда для записи Audacity проводным, конденсаторным микрофоном Hyper X QUADCAST S и была записана одним диктором женщиной. Аппаратные компоненты системы записи аудио данных:

1. Монитор №1. Монитор LG 19M38A-B - предназначенный для вывода текста диктору
2. Монитор №2. Монитор LG 19M38A-B - основной монитор, на котором осуществляется запись аудио данных
3. Наушники SONY WH 1000 XM4 для прослушивания записанных аудио данных
4. Системный блок -Dell Optiplex 3080 Micro для записи, обработки и хранения аудио данных
5. Микрофон Hyper X QUADCAST S для записи речевых сигналов
6. Клавиатура Logitech MX KEYS для ввода и изменения информации
7. Мышка Microsoft Sculpt Mouse для управления.

Основными устройствами системы записи являются монитор с оптимальным размером для отображения текста, конденсаторный студийный микрофон для высококачественной записи речи и компьютер, поддерживающий беспрерывную работоспособность программ записи звука, такие как Audacity, AdobeAudition.

В сумме сформированная речевая база узбекского языка для системы синтеза речи составила ~ 30 часов, которая была сформирована одним диктором женщиной. Всего в текстах, предоставленных для чтения использовались более 11 тысяч предложений. В предложениях использовано ~ 151500 слов на узбекском языке, 22721 из которых являются неповторяющимися словами.

Таблица 3. Параметры сформированной речевой базы

DOI: 10.36078/1688391447

Категория	**Обучение (Training)**	**Настройки параметров (Validation)**	**Общий**
Объем речевой базы (в часах)	27,5	3	30,5
Выражения	11107	584	11692
Слова	143902	7574	151476
Неповторяющиеся слова	21585	1136	22721

Статистические параметры сформированной многочасовой речевой базы узбекского языка приведены на рис. 7-8. Согласно статистике известно, что предложения с длительностью 7-8 секунд и с длиной 9 слов встречаются чаще всего [13].

Рис.7. Распределение предложений, встречающихся в речевой базе по длине.

DOI: 10.36078/1688391447

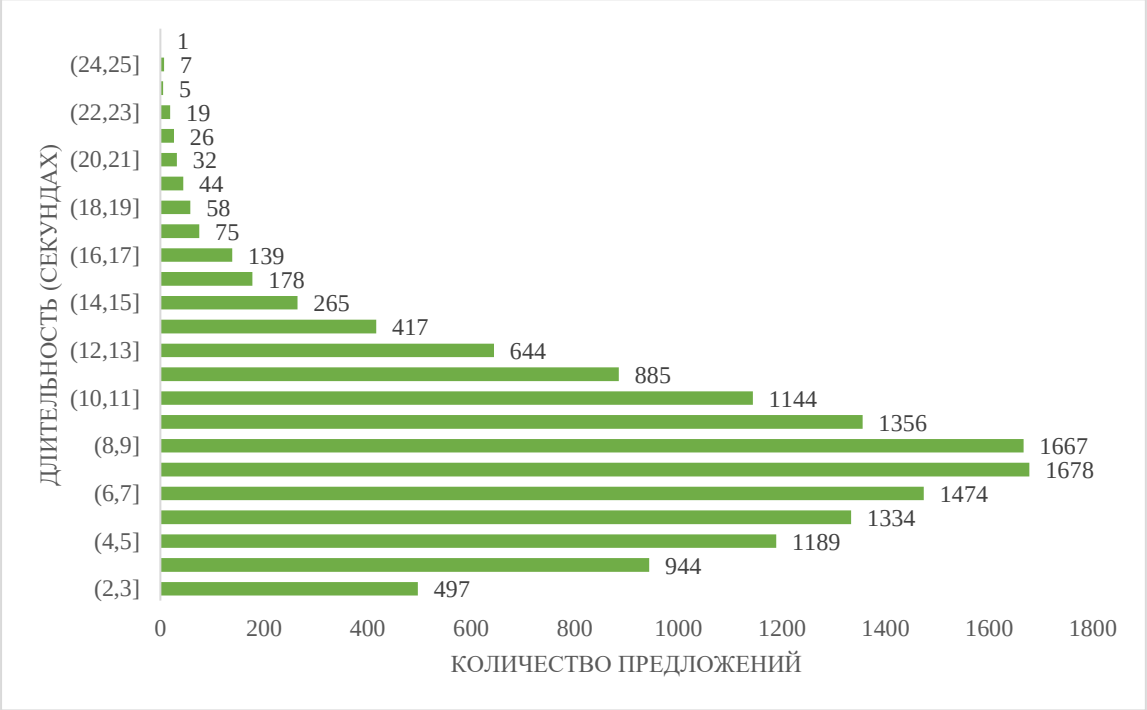


Рис.8. Распределение предложений, встречающихся в речевой базе по времени звучания.

Сформированная речевая база узбекского языка внедрился в TTS систему для узбекского языка и является неотъемлемой частью процесса синтеза речи. Синтезированная узбекская речь для оценки качества передаваемого голоса была оценена по шкале MOS от 1 до 5 и получил оценку 4,34, в то время, когда естественная речь, воспроизведенная носителем языка, получил оценку 4,45.

Использованная литература

1. Dargis, R. and Auzina, I., Towards a Modern Text-to-Speech System for Latvian, In Human Language Technologies – The Baltic Perspective, Frontiers in Artificial Intelligence and Applications vol. 307, 2018, pp.26–29.
2. Jindrich Matousek, Josef Psutka, Jirı Kruta, Design of Speech Corpus for Text-to-Speech Synthesis, Eurospeech 2001 – Scandinavia, 2001, 4p.
3. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, Librispeech: an ASR corpus based on public domain audio books, 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.
4. Muller, L., Psutka, J., Smıdl, L., Design of Speech Recognition Engine, Proceedings of TSD2000, Springer Verlag, Berlin, 2000, pp. 259–264.
5. Radova, V., UWB S01 Corpus – A Czech Read-Speech Corpus, Proceedings of ICSLP2000, vol. IV, Beijing, 2000, pp. 732–735.

DOI: 10.36078/1688391447

6. Erica Cooper, Emily Li, Julia Hirschberg, Characteristics of Text-to-Speech and Other Corpora Columbia University, USA, 1p
7. A. Acero, Acoustical and environmental robustness in automatic speech recognition. Springer Science & Business Media, 2012, vol. 201, 173 p.
8. Kishore Prahallad Automatic Building of Synthetic Voices from Audio Books CMU-LTI-10-XXX July 26, 2010, 128p.
9. A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, Using audio books for training a text-to-speech system, Pro-ceedings of the 9th International Conference on Language Resources and Evaluation, 2014, 5p.
10. Очилов М.М. Ўзбек тилидаги узлуксиз нутқни таниш технологияси, алгоритмлари ва дастурий мажмуаси. Техника фанлар бўйича фалсафа доктори (phd) диссертацияси авторефрати. Тошкент – 2022. 145Б.
11. M.I.Abdullaeva, D.B.Juraev, M.M.Ochilov, M.F.Rakhimov. Uzbek Speech Synthesis Using Deep Learning Algorithms. IHCI 2022, LNCS 13741, pp. 1–12, 2023. https://doi.org/10.1007/978-3-031-27199-1_5.
12. Хужаяров И.Ш. Интегралашган нейрон тармоқлар асосида ўзбек тили нутқини таниш алгоритмлари ва дастурий воситалари. Техника фанлар бўйича фалсафа доктори (PhD) диссертацияси авторефрати. Тошкент – 2021. Б.145.
13. Musaev M.M., Abdullayeva M.I., Ochilov M.M., Raximov M.F., Jurayev D.B. O’zbek tili nutqini sintezlovchi “Matn-nutq” dasturi, № DGU 17273. 01.07.2022.

 DOI: 10.36078/1688391447

ТАЪЛИМ ЖАРАЁНИДА ИҚТИСОДИЙ ИСЛОҲОТЛАРДАН ФОЙДАЛАНИШГА ДОИР БАЪЗИ МАЪЛУМОТЛАР

Хикоят Мусакуловна ИСРАИЛОВА
“Умумкасбий ва иқтисодий фанлар”
кафедраси катта ўқитувчиси
Ислом Каримов номидаги Тошкент давлат
университети Олмалиқ филиали,
Ш.И. ИНОМОВ
1Б-20 Мет талабаси,
Э.З.АБДУЛЛАЕВ
5Б-20 КЕМ талабаси

Аннотация. Тезис таълимда иқтисодий ислоҳотлар ва уларга тегишли чора-тадбирлардан фойдаланиш масалаларини ёритишга бағишланган.

Калит сўзлар: иқтисодий ислоҳот; аграр соҳа; мева-сабзавот; синов; тажриба; кўрсаткич; билим.

Ҳозирги қишлоқ хўжалигида ўтказилаётган изчил иқтисодий ислоҳотлар аҳолини сифатли озиқ-овқат маҳсулотларига талабни тўлароқ қондириши ва бу соҳадаги таъминотни тубдан яхшилаб жаҳон андозаларига тенглаштириш долзарб масалалардан бири ҳисобланади. Ўзбекистоннинг аграр соҳасида амалга оширилаётган иқтисодий ислоҳотларнинг ҳозирги босқичида фермер хўжаликларининг барқарор ривожланишини таъминлашда рақобат муҳитини шакллантириш орқали уларни рақобатбардошлигини ошириш, маҳсулот етиштиришда талаб ва таклиф мувозанатига эришиш, ички ва ташқи бозорда самарали фаолият юритувчи механизмларни кенг кўламда жорий қилиниши долзарб вазифалардан ҳисобланади. Мева-сабзавотчилик тармоғини мамлакат иқтисодиётидаги аҳамиятини тўғри баҳолаган ҳолда, иқтисодиётни босқичма-босқич эркинлаштириш шароитида айнан ушбу тармоқни ривожланишида туб бурилиш ясаш мақсадида Ўзбекистон Республикаси Президентининг 2006 йил 9 январдаги “Мева-сабзавотчилик ва узумчилик соҳасида иқтисодий ислоҳотларни чуқурлаштириш чора-тадбирлари тўғрисида”ги ПФ-3709-сонли фармони ҳамда 2009 йил 11 январдаги “Мева-сабзавотчилик ва узумчилик соҳасини ислоҳ қилиш бўйича ташкилий чора-тадбирлар тўғрисида”ги ПҚ-255-сонли қарори ва бошқа меъёрий ҳужжатларнинг қабул қилиниши ҳукумат томонидан бу тармоққа қанчалик эътибор билан муносабатда бўлаётганлигидан далолат беради.

 DOI: 10.36078/1688391447

Бозор иқтисодиётига ўтиш ва иқтисодий ислоҳотларни чуқурлаштириш шароитида мева-сабзавотчилик тармоғида ишлаб чиқариш самарадорлигини юксалтириш бевосита ушбу тармоқ тараққиётини тизимли тарзда комплекс амалга ошириш услубиёт ва усулларини такомиллаштириш билан бевосита боғлиқдир. Республикамиз мустақилликка эришгандан сўнгги даврда мева-сабзавотчилик тармоғида бозор муносабатларини жорий этиш борасида босқичма-босқич иқтисодий ислоҳотлар амалга оширилди. Хусусан, сўнгги йиллар давомида мева-сабзавотчилик тармоғида юз бераётган ўзгаришлар асосан қуйидагилардан иборат ҳисобланади:

- мева-сабзавотчиликни ривожлантиришни аграр иқтисодиётни модернизациялашнинг устувор йўналишлардан бири сифатида тан олиниши;

- мева-сабзавотчиликни ривожлантиришга ажратилаётган ерларнинг кенгайиб бораётганлиги;

- мева-сабзавот маҳсулотларини етиштиришда илғор технологияларни жорий этиш кўламининг ортиб бораётгани;

- мева-сабзавот маҳсулотларини сақлаш ва қайта ишлаш борасида йирик давлат дастурларининг амалга оширилаётгани ва ҳоказо.

Бугунги кунда мева-сабзавот маҳсулотлари етиштиришга республикамизда давлат тамонидан алоҳида эътибор берилиб, мева-сабзавотчилик тармоғидаги иқтисодий муносабатларга давлатнинг ўрни сезилиб турмоқда. Бозор мунсабатлари шаклланиши шароитида мева-сабзавотчилик тармоғидаги иқтисодий муносабатларнинг маркази ва унинг харакатлантирувчи кучи асосини деҳқон ва фермер хўжаликлари эгаллайди.. Бундан хулоса қиладиган бўлсак, мева-сабзавотчилик тармоғининг бозор шароитидаги барқарор ривожланиши, мева-сабзавот етиштирувчи деҳқоннинг манфаатини тўлиқ ҳисобга олиш бевосита боғлиқдир. Чунки барча шароитлар тенглиги шароитида, етиштирилган мевасабзавот миқдори, сифат кўрсаткичлари, таннархи, нарх шаклланиши омиллари (истеъмол хусусиятлари), натижада жамият учун келтирилган фойда миқдори айнан дастлабки бўғин ҳисобланган маҳсулот етиштириш жараёнида шаклланади. Хом-ашёнинг сифатли қайта ишланиши оралиқ маҳсулотлар ва шунингдек тайёр маҳсулот сифатида ҳам аксини топади. Янги технологиялар жорий этилиши ҳисобига тайёр маҳсулотларни сифатини, истеъмол хусусиятлари ўзгартириш мумкин, аммо, биогенетик жараёнлар маҳсули бўлган, етиштирилган мева-сабзавот сифатини яхшилаш мумкин эмас. Республикамизда сўнгги йилларда етиштирилаётган мева-сабзавот маҳсулотларининг сифатини оширишда, жаҳон бозори истеъмолчиларнинг талабларига мослашиш борасида бир қатор ишлар

DOI: 10.36078/1688391447

амалга оширилмоқда. Мева-сабзавотчилик тармоғидаги иқтисодий муносабатларнинг қанчалик бозор иқтисодиёти талабларига мос холда ташкил этилиши даражаси, мева-сабзавотчилик соҳаси иқтисодий ривожланиши ва маҳсулот етиштириш иқтисодий самарадорлиги даражасини белгилаб беради. Бозор шароитида мева-сабзавот маҳсулотлари ишлаб чиқариш ҳажми ва сифати ортиши энг аввало етиштирилаетган маҳсулотларга бозор талабининг мавжудлиги билан белгиланади. Шунингдек, маҳсулотнинг сифати, етиштиришга сарф қилинган ресурслар миқдори бозоридаги талаб ҳажми ўзгаришига туртки беради. Шу боисдан ҳам мева-сабзавотчилик хўжаликларида мева-сабзавот навларининг тезпишарлиги, ҳосилдорлиги, истеъмол сифати юқори бўлган ҳар бир ҳудуднинг табиий-иқлим шароитларига мос келадиган ички ва ташқи бозор талабига жавоб бера оладиган навларини илмий асосланган жойлаштириш зарурати тобора ортиб бормоқда. Қишлоқ хўжалиги соҳаси иқтисодиётини эркинлаштириш ва модернизациялаш жараёнида тармоқларни ривожлантириш мутаносиблигига эришиш, мамлакат ички бозорида қишлоқ хўжалиги маҳсулотлари бозорини мустаҳкам ривожлантириш, қишлоқ хўжаликлари маҳсулотларининг экспортини кучайтириш, пировард натижада мамлакат иқтисодиётининг барқарор ўсишини таъминлайди. Қишлоқ хўжалиги соҳасида иқтисодий ислоҳотлар давр талабларига мос равишда амалга оширилиб, унинг асосий мақсади - мамлакат ички бозори (аҳолининг истеъмол озиқ-овқат маҳсулотлари, саноат корхоналари учун эса хомашё маҳсулотлари) ва жаҳон бозори талаблари (хомашё ва тайёр маҳсулот сифатида) асосида етиштирилишига қаратилмоқда. Бу тадбирлар мамлакат раҳбарияти ва хукумати томонидан қабул қилинаётган тадбирлар орқали иқтисодий қўллаб- қувватланиб, тегишли дастурлар ва чора-тадбирлар аниқ мақсадли ва тизимли бўлишига эришилмоқда. Шу билан бирга, иқтисодиётни модернизациялаш шароитида мева- сабзавот ва узум маҳсулотларини етиштириш, истеъмол ҳажми ва турларини кенгайтиришнинг аҳамияти ортиб бормоқда. Аммо мева-сабзавот ва узум маҳсулотлари бозорида хилма-хил умумий иқтисодий манфаатлар ва алоқаларнинг аксарияти мақбул тартибга солинмаган. Шу боис, юқорида қайд этилган омиллар мева-сабзавотчилик ва узумчилик қуйи мажмуасида инновацион ва замонавий технологияларни жорий қилиш, диверсификациялашни кучайтиришни тақозо этади. Ушбу жиҳат эса республика иқтисодиёти ривожланишининг устувор йўналиши ҳисобланиб, илмийамалий аҳамиятга эгадир. Фзбекистонда етиштирилаётган мева-сабзавотларнинг рақобатбардошлик жиҳатдан устунлиги уларнинг ўзига хос таъми ва экологик тозалиги билан боғлиқ. Масалан, узумнинг маҳаллий сортлари таркибида шакар моддаси миқдори 30 %дан ошади, помидор таркибида қуруқ

 DOI: 10.36078/1688391447

моддалар улуши 5,5 %дан юқори бўлиб, бу кўрсаткичлар Европа мамлакатларидаги ушбу маҳсулотларни етиштирувчи корхоналар кўрсаткичлардан анча юқори. Ўзбекистон мева-сабзавотчилик мажмуаси рақобат афзалликларига яна қуйидагилар киради:

- ҳўл мева-сабзавотлар ва полиз маҳсулотлари, шунингдек, қайта ишланган мева-сабзавот маҳсулотлари ишлаб чиқариш ва жаҳон бозорида сотиш имкониятини берувчи табиий-иқлим шароитлари;

- мева-сабзавот маҳсулотлари сифат даражаси, хусусан экологик тозалигини таъминловчи ер ресурсларининг мавжудлиги;

- малакаси ва интизоми нисбатан юқори бўлгани ҳолда ишчи кучининг арзонлиги;

- илмий салоҳият ва амалий тадқиқотлар даражасининг юқорилиги;

- нисбатан инфратузилманинг ривожланганлиги (транспорт коммуникацияси, электрлаштириш даражаси).

Ўзбекистон Марказий Осиё ва қўшни минтақалар орасида мева-сабзавот етиштириш ҳамда улардан қайта ишланган маҳсулотларни ишлаб чиқариш, экспорт қилишда етакчи давлат ҳисобланади. Бу эса, мазкур минтақада рақобат афзаллигини вужудга келтиради. Ўзбекистон ҳудудий жиҳатдан яқин мамлакатларга мазкур турдаги маҳсулотларни экспорт қилишда сарфлаётган транспорт харажатлари ушбу ҳудудларга индустриал мамлакатларнинг экспортда транспорт харажатларига нисбатан паст. Дарҳақиқат, Ўзбекистон мева-сабзавотчилик мажмуаси маҳсулотларининг ҳудудий жиҳатдан яқин мамлакатлар бозорларида рақобатбардошлик даражасини оширишнинг улкан имкониятлари мавжуд. Фикримизча, ушбу имкониятлардан оқилона фойдаланиш мақсадга мувофиқ. Мева-сабзавотчилик соҳасидаги фаолият юритувчи агросаноат фирмаларининг мева ва сабзавотларни етиштириш, тайёрлаш, сақлаш, қайта ишлаш ва истеъмолчиларга турли кўринишларда (ҳўл мева, консерваланган, қуритилган ва ҳ.к.) етказиш билан боғлиқ бўлган ташкилий, иқтисодий ва технологик масалаларни ягона тизим доирасида мувофиқлаштириш, шу билан бирга сақлаш тизимини ривожлантириш имкониятлари юқорироқ бўлади. Ўзбекистон мева-сабзавотчилик мажмуаси ривожланишининг ўзига хос хусусиятларидан бири бу иқлим шароитининг қулайлигидир. Мазкур иқлим шароити мева ва узумнинг хилма-хил навларини турли муддатларда етиштириш ҳамда улардан ширага бой қуруқ мевалар, майиз ишлаб чиқариш имконини беради. Жумладан, кечпишар, баҳоргача сақланадиган олма, хўраки (ошхонабоб) узум навлари, шунингдек вино ишлаб чиқариш учун винобоп узумлар етиштириш имкони бор. Шунингдек, Ўзбекистон иқлимининг яна

DOI: 10.36078/1688391447

бир ўзига хослиги – вегетация даврининг узоқ муддатлилиги яъни, 180-250 кун давом этиши ва унинг эрта бошланишидир (февраль-март). Бу нафақат мева, узум ва сабзавотларнинг турли навларини етиштириш, балки қўшни мамлакатлар орсида мазкур маҳсулотлар ҳосилини анча эртароқ йиғиб-териб олиш имконини беради. Шунинг учун ҳам соҳа олимлари томонидан турли даврларда синовлардан ўтган, сайқалланган юқорида келтирилган тажрибалар таклифлар, кўрсаткичлар ҳамда билимлар мева-сабзавотчиликка ихтисослашган фермер ва деҳқон хўжаликлари учун жуда муҳим ҳисобланади.

Фойдаланилган адабиётлар

1. Ўзбекистон Республикасининг «Инвестиция фаолияти тўғрисида»ги қонуни. Ўзбекистон Республикаси қонун ҳужжатлари тўплами, 2006 й.

2. Ўзбекистон Республикаси Президентининг 2006 йил 9 январдаги ПФ-3709-сонли “Мева-сабзавотчилик ва узумчилик соҳасида иқтисодий ислоҳотларни чуқурлаштириш чора-тадбирлари тўғрисида”ги фармони.

3. Ўзбекистон Республикаси Президентининг 2009 йил 11 январдаги ПҚ-255-сонли “Мева-сабзавотчилик ва узумчилик соҳасини ислоҳ қилиш бўйича ташкилий чора-тадбирлар тўғрисида”ги қарори. “Иқтисодиёт ва инновацион технологиялар” илмий электрон журнали. № 6, ноябрь-декабрь, 2018 йил 6/2018 (№ 00038) www.iqtisodiyot.uz

4. Ўзбекистон Республикаси “Ер кодекси” ва қишлоқ хўжалигига оид қонунлари. – Т.: Адолат, 1999. 19-б. 5. Олимжонов О. ва бошқалар Фермер фаолиятининг ҳуқуқий ва молиявий асослари. – Тошкент, 2005.

 DOI: 10.36078/1688391447

ARTIFICIAL INTELLIGENCE AND THE JOB MARKET: OPPORTUNITIES AND CHALLENGES

Dilshodbek Komilovich KOMILOV
computer science teacher
Specialized Boarding School
of the Ministry of Internal Affairs
of the Republic of Uzbekistan
d.komilov01@mail.ru

Abstract. This article examines the impact of Artificial Intelligence (AI) on the job market, highlighting both opportunities and challenges. While AI presents new job roles and improved efficiency, concerns about job displacement and inequality persist. The article explores these issues and emphasizes the need for solutions to ensure a fair transition to an AI-powered future.

Keywords: artificial intelligence (ai); job market; opportunities; challenges; job displacement; inequality; efficiency; data analysis; programming; ai development.

Artificial Intelligence (AI) is rapidly transforming the world we live in, with applications ranging from self-driving cars to personalized healthcare. As AI technology continues to develop, it is having a significant impact on the job market, presenting both opportunities and challenges.

On the one hand, AI is creating new job roles in areas such as data analysis, programming, and AI development. According to a report by the World Economic Forum, AI is expected to create 2.3 million new jobs by 2025. Additionally, AI is expected to improve the efficiency of existing jobs, leading to increased productivity and economic growth. For example, AI-powered automation can streamline processes and reduce the need for manual labor in industries such as manufacturing and logistics.

However, concerns about the displacement of human workers persist. AI has the potential to automate many routine and repetitive tasks, which could lead to job losses for human workers. The fear is that AI will replace human workers, particularly in industries that rely heavily on manual labor, such as transportation, retail, and food service. A report by McKinsey & Company estimated that up to 800 million jobs could be displaced by automation by 2030, with a disproportionate impact on low-skilled workers.

Furthermore, there are concerns about the potential impact of AI on equality in the job market. Certain groups may be disproportionately affected by job displacement or may face barriers to entering the AI field due to lack of access to education or training. For

 DOI: 10.36078/1688391447

example, women and underrepresented minorities are already underrepresented in the technology industry, and there is a risk that the same patterns could emerge in the AI industry. This could lead to further inequalities in the job market.

Despite these concerns, there are also reasons to be optimistic about the impact of AI on the job market. AI has the potential to create new job roles that require a range of skills and expertise, from data analysis and programming to customer service and social skills. Additionally, AI can help to improve the efficiency of existing jobs, enabling workers to focus on tasks that require creativity and problem-solving skills.

Moreover, there is evidence to suggest that the introduction of new technology can lead to a net increase in jobs over time. A study by the National Bureau of Economic Research found that industries that adopted new technology actually experienced higher rates of job growth than industries that did not. This suggests that the impact of AI on the job market may not be as dire as some predict.

However, to ensure that the benefits of AI are realized without exacerbating inequalities in the job market, it is essential to address these issues head-on. This includes investing in education and training programs that equip workers with the skills needed to succeed in an AI-powered economy. It also means ensuring that AI is developed and deployed in an ethical and responsible manner, with consideration given to the potential impact on job displacement and inequality.

There has been extensive research on the impact of AI on the job market, with studies exploring the potential for job displacement, the creation of new job roles, and the impact of AI on inequality. A report by the World Economic Forum found that while AI has the potential to create 2.3 million new jobs by 2025, it could also lead to the displacement of 75 million jobs. The report also highlighted the need for investment in education and training programs to ensure that workers are equipped with the skills needed to succeed in an AI-powered economy.

The impact of AI on the job market is complex and multifaceted, with both positive and negative outcomes. On the positive side, AI is creating new job roles in areas such as data analysis, programming, and AI development. Additionally, AI is expected to improve the efficiency of existing jobs, leading to increased productivity and economic growth. However, concerns about the displacement of human workers persist, as AI has the potential to automate many routine and repetitive tasks.

Furthermore, there are concerns about the potential impact of AI on equality in the job market. Certain groups may be disproportionately affected by job displacement or may face barriers to entering the AI field due to lack of access to education or training. This could lead to further inequalities in the job market. However, there is evidence to suggest

 DOI: 10.36078/1688391447

that the introduction of new technology can lead to a net increase in jobs over time, suggesting that the impact of AI on the job market may not be as dire as some predict.

The methodology used to study the impact of AI on the job market varies depending on the research question being explored. Some studies use data analysis to explore trends in job growth and displacement in industries that are adopting AI technology. Others use surveys or interviews to gather data on worker attitudes towards AI and their experiences with job displacement. Additionally, some studies use simulation models to predict the impact of AI on the job market under different scenarios.

The methodology used to study the impact of AI on the job market is varied and depends on the specific research question being explored. However, there is a general consensus that investment in education and training programs is essential to ensure that workers are equipped with the skills needed to succeed in an AI-powered economy. Additionally, responsible deployment of AI technology is crucial to minimize negative consequences such as job displacement and inequality.

In conclusion, the impact of AI on the job market is a complex and multifaceted issue, with both opportunities and challenges. While there is a risk that AI could displace human workers and exacerbate existing inequalities, there is also the potential for AI to create new job roles and improve the efficiency of existing jobs. To ensure that the benefits of AI are realized without negative consequences, it is crucial to address these issues through education, training, and responsible deployment of AI technology.

References

1. World Economic Forum. (2018). The future of jobs report 2018. Geneva: World Economic Forum.
2. Autor, D. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3-30.
3. Brynjolfsson, E., & McAfee, A. (2017). The economics of artificial intelligence. In Agrawal, A., Gans, J., & Goldfarb, A. (Eds.), The economics of artificial intelligence: An agenda. Chicago: University of Chicago Press.
4. Frey, C.B., & Osborne, M.A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254-280.
5. McKinsey Global Institute. (2017). Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. San Francisco: McKinsey & Company.

DOI: 10.36078/1688391447

ПОСЛОВИЦЫ И ПОГОВОРКИ КАК ПЕРСПЕКТИВНЫЙ МАТЕРИАЛ ДЛЯ ПЕРВОГО ЗНАКОМСТВА СТУДЕНТОВ С КОРПУСОМ

Марина Самуиловна КОГАН,
кандидат технических наук, доцент
Санкт-Петербургский политехнический
университет Петра Великого

Дарья Александровна ГАВРИЛИК
магистр лингвистики, ассистент
Санкт-Петербургский политехнический
университет Петра Великого

Аннотация. В статье обсуждается ситуация по использованию подходов корпусной лингвистики (КЛ) в обучении иностранным и родному языкам (DDL). Анализ литературы показывает, что подходы DDL занимают скромное место в современной лингводидактике. По мнению авторов, знакомство с корпусными ресурсами старшеклассников в рамках проектной работы на уроках иностранного (родного) языка может привести к более активному и эффективному использованию подходов КЛ в обучении. Обосновывается почему задание для проектной работы, заключающееся в анализе употребления известных английских пословиц и поговорок в речи носителей английского языка, способно заинтересовать учащихся, соответствует их уровню владения иностранным языком и компьютерными технологиями, развивает навыки проведения научного исследования. Описаны результаты проведения педагогического эксперимента по знакомству старшеклассников с онлайн корпусами английского языка COCA и NOW.

Ключевые слова. Корпусная лингвистика; лингводидактика; корпусы COCA и NOW; Data-driven learning, пословицы и поговорки; проектная работа; английский язык.

Использование корпусов и подходов корпусной лингвистики в обучении родному и иностранному языкам, известное в зарубежной научной литературе по аббревиатуре DDL (Data-driven learning), (термин введен Т. Джонсом в 1990-х) – одно из многочисленных применений корпусной лингвистики. С начала 1990-х г, когда национальные корпусы появились в открытом доступе в сети Интернет, энтузиасты стали говорить о «корпусном повороте» в методике обучения иностранным языкам. Однако за прошедшие 30 лет ожидаемой революции в этой области прикладной лингвистики не произошло, хотя во многих других областях

DOI: 10.36078/1688391447

корпусная лингвистика кардинально изменила сложившиеся подходы, например, в лексикографии [1]. В научной литературе за последние годы утвердился термин ‘DDL-like’ подход, означающий отход от первоначального представления о том, как должен работать корпусный подход в обучении иностранным языкам: через непосредственное обращение к корпусу аутентичных текстов для формирования собственной траектории изучения ИЯ на продвинутом уровне. ‘DDL-like’ подход – это более широкий термин, охватывающий разные способы использования инструментов и подходов корпусной лингвистики для педагогических целей с разными категориями учащихся с разным уровнем иноязычной подготовки [2]. Благодаря энтузиазму П. Кроссвейта (Crosthwaite) [3] DDL-like подходы все активнее внедряются в методику обучения ИЯ в средней и даже начальной школах. В.П. Захаров и М.С. Коган сделали попытку оценить масштаб использования корпусов в изучении и преподавании языка в России на основе анализа публикаций, размещенных в электронной библиотеке e-library [4].

Однако вопрос о том, насколько активно учащиеся, познакомившиеся с корпусами и корпусными подходами в процессе обучения, продолжают ими пользоваться самостоятельно для изучения ИЯ или в использовании ИЯ на практике после окончания педагогического эксперимента или курса обучения, остается открытым. Так, М. Чарльз (Charles) в результате многолетнего эксперимента установила, что через год после окончания курса по академическому письму для аспирантов Оксфорда, которые осваивали корпусные подходы в течение 6-ти недель, только 36% продолжали их использовать (против 86% заявивших о таком намерении в опросе, проведенном сразу после завершения обучения) [5].

В контексте этого исследования актуальным становится вопрос о том, когда следует начинать знакомить обучающихся с корпусом и какие задания им предлагать, чтобы у них выработался устойчивый интерес к этому типу лингвистических ресурсов. Мы считаем, что оптимальным решением было бы приобщение к работе с корпусными ресурсами старшеклассников в рамках проектной деятельности на материале, «привязанном» к основному учебнику по иностранному языку (учебно-методическому комплексу (УМК)).

Количество разных типов заданий, которые требуют обращения к корпусу, описанных российскими исследователями и авторами из других стран, очень велико. Интересующиеся могут найти их в публикациях П.В. Сысоева, Л.К. Раицкой, А.И. Левинзон, А. Болтона и Т. Кобба, К. Лэкмана [6 – 10] и др.

В рамках настоящего эксперимента мы задались целью познакомить старшеклассников Естественно-научного лицея СПбПУ Петра Великого с

 DOI: 10.36078/1688391447

большими онлайн корпусами английского языка (the COCA[2] и NOW[3]) с разных углов зрения с целью освоения как можно более широкого диапазона функций этих корпусов в режиме непосредственного обращения к корпусам (a hands-on mode) в рамках проектной работы в курсе английского языка.

Выбор исследовательской проблемы

Для того чтобы стать основой проекта, задача должна отвечать ряду требований:

1) стимулировать учащихся освоить разные поисковые запросы к корпусу, выдвигать собственные гипотезы, проверять их, анализировать полученные результаты и т.д.

2) учитывать уровень владения иностранным языком и интересы учащихся;

3) соответствовать изучаемому материалу по обязательному учебнику (УМК);

4) позволять организовать самостоятельную работу в малых группах.

По нашему мнению, корпусное исследование пословиц и поговорок в полной мере отвечает этим требованиям. Такое задание позволит учащимся «безопасно» начать самостоятельную работу с корпусными данными, избежав многих «опасностей», описанных в литературе.

Поговорки и пословицы включены во все учебники по английскому языку. Исследователи и преподаватели ИЯ видят огромный потенциал использования пословиц и поговорок для повторения грамматических конструкций, расширения словарного запаса, развития коммуникативных навыков, просто разрядки [11], предлагая использовать ресурсы Интернета для работы с ними [12].

Интересной особенностью употребления пословиц и поговорок в речи является то, что они часто становятся объектом языковой игры, «переиначивания». И если поиск таких примеров для анализа по словарям является уделом опытных исследователей [13], то проведение подобного исследования на материале большого онлайн корпуса доступно даже школьнику. Факт наличия в корпусах идиоматических выражений отличных от их «канонической» формы, зафиксированной в словарях, подтвержден разными исследователями [14 – 16].

С педагогической точки зрения представляется бесспорным, что умение обнаружить и проанализировать неизвестные факты является очень важным навыком исследователя независимо от области его деятельности. Непредсказуемость результата делает задачу поиска вариантов употребления

2 Davies, M. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.(2008-)

3 Davies, M. Corpus of News on the Web (NOW). Available online at https://www.english-corpora.org/now (2016-)

DOI: 10.36078/1688391447

пословиц и поговорок в корпусе по-настоящему аутентичной исследовательской задачей.

Организация проектной работы с лицеистами 11 класса

Эксперимент был проведен в 11 классе Естественно-научного лицея при СПбПУ Петра Великого. В эксперименте участвовали 16 человек с высоким уровнем знаний по физике, математике, информатике и разным уровнем владения английским языком и отношением к занятиям по этому предмету: от увлеченного до поверхностного. Разный уровень владения объясняется тем, что в ЕНЛ поступают учащиеся из разных средних школ Санкт-Петербурга и других городов на конкурсной основе после окончания 9-го класса.

Базовым учебником является учебник *Forward*[4]. В каждом уроке встречаются идиоматические выражения, в некоторых уроках они изучаются в специальном разделе, как правило с забавными иллюстрациями их буквального понимания. Количество тренировочных упражнений на усвоение идиоматических выражений недостаточно, что типично для большинства учебников по английскому языку [17]. В конце каждого урока предлагается идея для проектной работы (*Project ideas* section).

Основная часть проектной работы осуществлялась лицеистами в рамках самостоятельной домашней работы в течение 3-х недель первой четверти. Аудиторно было проведено 1-е занятие в формате 30-минутной вводной лекции и последнее занятие, на котором каждая команда докладывала о результатах исследований, проведенных в корпусе. На вводном занятии учитель рассказал о целях проекта, кратко рассказал о корпусной лингвистике, представил онлайн корпусы COCA и NOW, сформулировал поисковую задачу для каждой команды. Учащиеся получили подробную письменную инструкцию по проведению корпусного исследования и представления полученных результатов в итоговом Отчете. Каждая команда получила первую часть пословицы из списка для проведения качественного и количественного корпусного анализа.

1 To err is human…
2 The grass is always greener…
3 Necessity is the mother of…
4 who laughs last …
5 Early to bed …

[4] Verbitskaya M. (ed.): Forward: English Student's Book. (Grades 10, 11). Moscow: Ventana-Graph. Person Education Limited. 2016.

DOI: 10.36078/1688391447

6 Birds of a feather.

Учащиеся должны были освоить работу в Корпусе современного американского английского языка (COCA) в режимах *List, Chart* и *KWIC* с привлечением авторитетного онлайн словаря для верификации канонической формы поговорки.

В режиме *KWIC* с выравниванием по *правому краю* учащиеся должны были найти окончания поговорок, посчитать количество стандартных/канонических и измененных форм и выписать примеры модификаций. Мы просили обратить особое внимание на повторяющиеся модификации, а также на наиболее необычные, неожиданные парадоксальные окончания пословиц/поговорок с их точки зрения. Еще одно задание касалось определения типа высказывания, содержащего пословицу или поговорку: утверждение, вопрос, восклицание или незавершенная мысль.

В режиме *KWIC* с выравниванием по *левому краю* учащиеся должны были посчитать количество употреблений пословицы или поговорки в начале предложения, в середине предложения, в виде прямой цитаты. Наконец, учащиеся должны были узнать частотность употребления канонической формы поговорки в корпусе.

После этого лицеисты должны были выполнить ряд заданий в корпусе NOW. Это задание дало им возможность:

1) познакомиться с еще одним корпусом в рамках одного проекта;

2) обратить внимание на влияние размера корпуса на результаты поиска. (О необходимости создания сверхбольших корпусов для исследования паремий и идиоматических выражений из-за низкой частотности этих единиц даже в национальных корпусах говорят многие исследователи. См, например, [18]).

3) освоить функции, непредставленные в COCA

Результаты и обсуждение

Участники эксперимента проявили неподдельный интерес на вводной лекции, задавали конкретные вопросы о разметке корпуса. В конце лекции развернулось оживленное обсуждение вопроса о том, насколько часто англичане употребляют пословицы и поговорки в речи, что позволило преподавателю еще раз подчеркнуть важность корпусных исследований, которые позволяют получить ответ на самые разные вопросы об использовании языка, включая вопрос об использовании пословиц и поговорок.

На итоговом занятии большинство учащихся отметили, что им понравилась работа в корпусе, хотя она оказалась более трудоемкой и иногда утомительной, чем

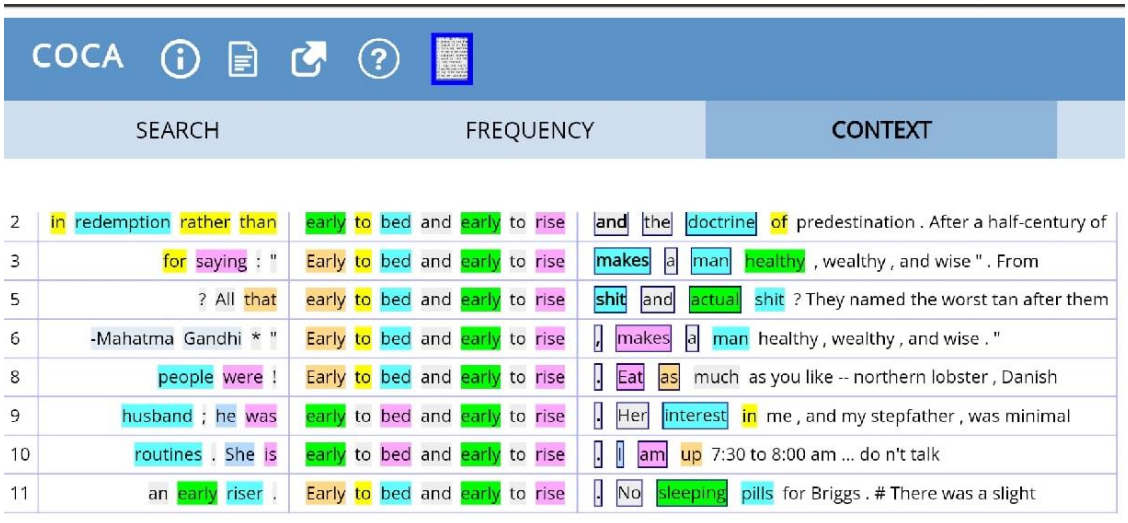
DOI: 10.36078/1688391447

они предполагали. Анализ строк конкорданса не вызвал затруднений, т.к. в 5 случаях из 6 количество употреблений, указанных паремий в COCA было меньше 100 (Таблица 1). Количество примеров для анализа в корпусе NOW было сокращено до 100 за счет использования функции выдачи случайной выборки из 100 примеров.

Таблица 1. Результаты поиска канонической формы пословиц и поговорок и их модификаций в корпусах COCA и NOW

Начало пословицы (поговорки)	Количество употреблений в корпусе всего/в канонической форме	
	COCA	NOW
To err is human/+(but/and) to forgive (is) divine	87/20	949/256
The grass is always greener /+ on the other side (of the fence)	79/32	495/179
Necessity is the mother of /+ invention	78/73	1553/1463
who laughs last/ He who laughs last, laughs best	26/4	153/59
Early to bed /+early to rise make a man healthy, wealthy and wise	71/12	522/70
Birds of a feather / + birds of a feather flock together	230/44	2345/625

Режим просмотра *KWIC* с заданной фразой в центре страницы и ближайшими словами слева и справа от нее, окрашенными в разные цвета, можно назвать дружеским по отношению к пользователю, т.к. он позволяет комфортно проводить качественные и количественные исследования паремий в канонической форме и различных модификациях (рис.1).



DOI: 10.36078/1688391447

Рис.1. Фрагмент выдачи в режиме просмотра *KWIC* с выражением *‘early to bed and early to rise’* в центре с выравниванием по правому краю

Учащиеся отметили, что хотя пословицы и поговорки часто встречаются в разных видоизмененных формах, по настоящему необычных, ярких, запоминающихся окончаний не так много. (Например: *Early to bed and early to rise makes you healthy, happy and slim; Who were early to bed and early to rise were up to 48% less at risk; Early to bed and early to rise is hardly being practiced by the younger generation; . … who laughs last... laughs the longest! The person who laughs last laughs hardest*; *… to err is human but to err twice is stupid*; *to err is human but to blame other people is politics*; *Necessity is the mother of Ontario’s hydro privatization*; *Necessity is the mother of a number of virtues …thrift, ingenuity and even imagination; Birds of feather eat together; I was sitting in a Birds of Feather community session on “Innovations in Big Data”…*) Учащиеся добросовестно проанализировали строки конкорданса и тенденции использования пословицы во времени в режиме *Chart view*. Им также понравилось задание на знакомство с расширенным контекстом наиболее интересной, по их мнению, модификацией поговорки. Корпус NOW в этом случае приводит к полному тексту статьи на внешнем сайте.

Трудности, с которыми столкнулись лицеисты, касались 2-х аспектов: невнимательное чтение инструкции в надежде на эффективность метода проб и ошибок при знакомстве с корпусом, что привело к потере времени, и неустойчивость работы сайтов на мобильных устройствах.

Выводы

Эксперимент показал, что выбранный материал для исследования и формат работы эффективны для знакомства старшеклассников с онлайн корпусами в курсе иностранного языка. Думается, что предложенный алгоритм знакомства с корпусом через проведение исследования по использованию в речи пословиц и поговорок и их модификаций можно использовать при изучении любого языка, для которого существует достаточно представительный онлайн корпус, и применять на занятиях по иностранному языку для знакомства с корпусами студентов младших курсов разных направлений подготовки.

Это задание содержит элементы настоящего исследования и компьютерной игры благодаря непредсказуемости результатов и многократному повторению запросов. Необходимость повторения запросов способствует закреплению поисковых навыков в таких сложных ресурсах как корпус COCA или NOW и в целом способствует развитию у учащихся лингвокомпьютерной компетенции. Ввиду этого мы считаем предложенный способ первого знакомства с онлайн корпусами

DOI: 10.36078/1688391447

старшеклассников и студентов младших курсов инженерных специальностей перспективным для формирования у обучающихся интереса и потребности в обращении к корпусным ресурсам в процессе изучения ИЯ и его использовании.

Однако задачу экспериментального измерения интереса студентов к корпусным ресурсам, готовность их использовать в дальнейшем еще предстоит решить в будущих исследованиях.

Использованная литература

1. Hanks P. The Corpus Revolution in Lexicography // International Journal of Lexicography. 2012, Vol. 25(4). P. 398–436.
2. Boulton A., Vyatkina N. Thirty years of data-driven learning: Taking stock and charting new directions over time // Language Learning & Technology. 2021. Vol. 25, №3. P. 66–89. URL: https://www.lltjournal.org/item/10125-73450/ (дата обращения: 10.05.2023).
3. Boulton A. Data-driven learning for younger learners: Obstacles and optimism // P. Crosthwaite (Ed.). Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners. Routledge. 2019 P. 14-20.
4. Захаров В.П., Коган М.С. Использование корпусов в изучении и преподавании языка в России: достижения, проблемы, перспективы // O‘zbek milliy va ta’limiy korpuslarini yaratishning nazariy hamda amaliy masalalari. Vol. 1, №. 01. 2021. С. 15 – 19.
5. Charles M. The gap between intentions and reality: Reasons for EAP writers’ non-use of corpora // Applied Corpus Linguistics 2 (2022) 100032. 9p. https://doi.org/10.1016/j.acorp.2022.100032
6. Сысоев П.В. Лингвистический корпус, корпусная лингвистика и методика обучения иностранным языкам // Иностранные языки в школе. 2010. №5. С. 12–21.
7. Раицкая Л.К. Дидактические возможности корпусных интернет-технологий в преподавании иностранного языка в высшей школе // Вестник Московского государственного областного университета. Серия: Педагогика. 2009. № 4. С. 123-127.
8. Левинзон А.И. Корпусное преподавание в российской школе // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 641-658.
9. Cobb T., Boulton A. Classroom applications of corpus analysis // Cambridge Handbook of English Corpus Linguistics (D. Biber, R. Reppen (Eds)). 2015. P.478 – 497.

 DOI: 10.36078/1688391447

10. Lackman K. Classroom Games from Corpora: Using Corpora to teach Vocabulary. 2010. http://www.kenlackman.com/files/CorporaGamesBook11SAMPLEPAGES.pdf (дата обращения: 10.05.2023).
11. Павлова Е.А. Приемы работы с пословицами и поговорками на уроках английского языка // Иностр. языки в школе. 2010. №5. С. 37–44.
12. Азизова Ф. С. Обучение фразеологизмов с помощью информационных технологий // Компьютер лингвистикаси: муаммо ва ечимлар (Компьютерная лингвистика: проблемы и решения, Computational linguistics: challenges and solutions) мавзусидаги халқаро онлайн илмий-амалий конференция материаллари тўплами. – Тошкент, 19 апрель, 2021 й – С.16 – 19.
13. Артемова А.Ф., Леонович Е.О. Английские антипословицы // Иностр. языки в школе. 2017. №4. С. 55–58
14. Arnaud P.J.L., Maniez F., Renner V. Non-Canonical Proverbial Occurrences and Wordplay: A Corpus Investigation and an Enquiry Into Readers' Perception of Humour and Cleverness//. Wordplay and Metalinguistic / Metadiscursive Reflection: Authors, Contexts, Techniques, and Meta-Reflection. 2015. P.135–160. URL: https://doi.org/10.1515/9783110406719-007
15. Добровольский Д.О. Корпусы текстов и двуязычная фразеография // Вестник Новосибирского государственного педагогического университета. 2015. № 5 (27). С. 23 – 37.
16. Комарова И.А., Коган М.С. Исследование английской фразеологии с помощью подходов корпусной лингвистики // Компьютерная лингвистика и вычислительные онтологии. Выпуск 3. (Труды XXII Международной объединенной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. Сборник научных трудов). СПб: Университет ИТМО, 2019. С. 40–49.
17. Cobb T: From Corpus to CALL: The use of technology in teaching and learning formulaic language. In: Understanding formulaic language: A second language acquisition perspective. London-New York: Routledge 2019. P. 192–210.
18. Benko V., Zakharov V.P. Very large Russian corpora: New opportunities and new challenges // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialogue 2016". 2016. С. 83-98. http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf (дата обращения: 10.05.2023).

 DOI: 10.36078/1688391447

КОНТЕНТ-АНАЛИЗ КАК МЕТОД ИССЛЕДОВАНИЯ РАЗВИТИЯ РОССИЙСКОЙ ИННОВАЦИОННОЙ СИСТЕМЫ

Елена Васильевна МАСЛЮКОВА,
кандидат экономических наук, доцент,
заведующий кафедрой экономической кибернетики,
Южный федеральный университет,
Российская Федерация
maslyukova@sfedu.ru

Аннотация. Методология контент-анализа для исследования развития Российской инновационной системы предполагает использование количественной статистическую обработку тестовых данных, составление словаря контент-анализа и выделение основных лексем на основе их частотности; составление матрицы контент-анализа, отражающей встречаемость каждой из лексем в каждом из фрагментов текста; проведение факторного анализа (метода главных компонент), позволяющего сгруппировать лексемы в тематические группы. В результате проведенного анализа выявлено десять микротем, связанных с отношением акторов к проводимой государственной политике в области инноваций.

Исследование выполнено за счет гранта Российского научного фонда № 21-18-00562, https://rscf.ru/project/21-18-00562/ в Южном федеральном университете

Ключевые слова: нарративная экономика качественный анализ контент-анализ факторный анализ российская инновационная система.

Развитие российской инновационной системы в значительной мере связано с реализацией государственной политики, направленной на создание условий для осуществления инновационной деятельности. Количественные показатели позволяют оценить эффективность деятельности органов государственной власти, однако ориентация исключительно на них порождают некоторые противоречия. Так, Стратегия инновационного развития до 2020 года предполагала достижение 45 ключевых показателей эффективности (КПЭ), но исходя из доклада о реализации отраслевых документов стратегического планирования России за 2019 год, в Стратегии инновационного развития до 2020 года детальной информации о выполнении мероприятий не было представлено, при это также указывается, что оценить степень реализации стратегии не представляется возможным. Как отмечают российские эксперты, большая часть количественных показателей так и не была достигнута [1].

DOI: 10.36078/1688391447

Применение качественных методов исследования российской инновационной системы позволяет выявить специфические особенности ее развития на основе анализа нарративов акторов, имеющих непосредственное отношение к данной системе. Согласно подходу Р. Шиллера, что в экономике под анализом нарративов понимается изучение распространения и динамики изменения популярных повествований, рассказов, особенно тех, которые представляют интересы человека и эмоции, и как эти изменения в течение времени связаны с экономическими колебаниями [2]. В качестве источников нарративов для анализа использовались данные 27 глубинных интервью с представителями академической сферы – учеными, занимающимися инновационной деятельностью (развернутые ответы респондентов на вопрос 4 – Какова роль государства в развитии РИС (инноваций)? Как Вы оцениваете проводимую государственную политику в области инноваций?). Респонденты представляли 6 федеральных округов: Приволжский - 8 (Нижний Новгород - 7, Казань - 1); Южный - 9 (Ростов-на-Дону - 5, Таганрог - 4), Центральный - 4 (Москва - 4), Сибирский - 2 (Новосибирск - 1, Томск - 1), Уральском - 3 (Екатеринбург - 3) и Северо-Западный - 1 (Санкт-Петербург - 1). Возраст информантов варьировался от 21 до 73 лет. Возрастной диапазон интервьюируемых связан с разнообразием работников, вовлеченных в академическую сферу: в возрасте до 35 лет (N=15), с 36 до 59 (N=7), старше 60 (N=5). Направление образования всех респондентов соответствует текущей деятельности в сфере инноваций и распределено следующим образом: естественные науки (N=11), технические науки (N=11), науки об обществе (N=5). 4 респондента обладают степенью доктора наук и 15 степенью кандидата наук, 8 респондентов занимаются инновационной деятельностью без степени.

Далее для выявления доминирующих в нарративах акторов российской инновационной системы микротем был осуществлен количественный контент-анализ выделенных нарративов.

Процедура контент-анализа включает в себя следующие этапы: разбиение текстов на фрагменты; составление словаря контент-анализа, опирающегося на принцип частотности; включение всех словоформ выбранных лексем в словарь; составление матрицы контент-анализа, отражающей встречаемость каждой из лексем в каждом из фрагментов текста; проведение факторного анализа (метода главных компонент), позволяющего сгруппировать лексемы в тематические группы; выделение главных микротем текста, интерпретация полученных результатов и содержательный анализ.

DOI: 10.36078/1688391447

В результате проведенного анализа было выделено 25 основных лексем (рис. 1).

Рисунок 1 – Облако ключевых лексем, включенных в словарь контент-анализа

На следующем этапе с целью выделения главных микротем в нарративах акторов, представляющих академическую сферу российской инновационной системы, был применен метод главных компонент. В результате было выявлено 11 факторов (Таблица 1):

Таблица 1 – Факторные нагрузки

Номер фактора	**Факторные нагрузки**	**Собственные значения**
Фактор 1	Деньги (0,52) Производить (-0,76)	2,520614
Фактор 2	Интерес (-0,61) Финансы (-0,86) Ученый (-0,79)	2,450991
Фактор 3	Исследование (0,55) Наука (0,78) Оборудование (0,63)	2,181015
Фактор 4	Технология (0,90) Развитие (0,39)	1,905533

DOI: 10.36078/1688391447

	Техника (0,90)	
Фактор 5	Бизнес (-0,72) Министерство (-0,46) Система (0,73) Россия (-0,45)	1,694748
Фактор 6	Компания (-0,79) Инновации (0,48)	1,478762
Фактор 7	Поддержка (-0,82) Проект (-0,82)	1,429791
Фактор 8	Институт (-0,67) Университет (0,75)	1,266595
Фактор 9	Бюрократия (-0,65) Проблема (0,58)	1,208539
Фактор 10	Выполнение (-0,80) Государство (-0,72)	1,179836

Анализ выделенных факторов показывает, что в рамках выделенных нарративов освещались проблемы, связанные с финансированием производства (факторы 1 и 7), ролью науки и образования для развития инноваций (факторы 2, 3 и 8), необходимостью развития технологий (фактор 4), взаимодействие с бизнесом (факторы 5 и 6), а также бюрократией и государственным заказом (факторы 9 и 10).

Таким образом, проведенный контент-анализ позволил ключевые элементы в высказываниях акторов российской инновационной системы на основе обработки лексем. Результаты количественного контент-анализа будут соотнесены с качественным нарративным анализом для более полного и системного понимания развития Российской инновационной системы.

Использованная литература

1. Вольчик В.В., Фурса Е.В., Маслюкова Е.В.. Государственное управление и развитие российской инновационной системы // Управленец. – 2021. - Т. 12. - № 5. - С. 32–49. DOI: 10.29141/2218-5003-2021-12-5-3.

2. Shiller R.J. Narrative Economics // American Economic Review. - 2017. №107(4). - P. 967–1004.

 DOI: 10.36078/1688391447

ЭЛЕКТРОННЫЕ ОБРАЗОВАТЕЛЬНЫЕ РЕСУРСЫ В ОБУЧЕНИИ РУССКОМУ ЯЗЫКУ КАК ИНОСТРАННОМУ

Мариета Муратовна МЕРЕТУКОВА
кандидат филологических наук, доцент
Адыгейский государственный университет
boss750@mail.ru

Аннотация. Тезис посвящен вопросу использования электронных образовательных ресурсов в современной практике преподавания русского языка как иностранного. Описывается эффективность образовательных Интернет-ресурсов для изучения русского языка как иностранного. Приводятся примеры открытых интернет ресурсов с описанием и возможностями действий.

Ключевые слова: русский язык как иностранный; электронные средства обучения; Интернет в обучении русскому языку как иностранному.

Активное использование электронных ресурсов становится неотъемлемой частью современной модели образовательного пространства, к отличительным чертам которой относят «большую открытость, доступность и гибкость за счет широкого применения средств самообразования на основе новых информационных и коммуникационных технологий», использование которых «способствует развитию внутренней мотивации обучающихся для получения новых знаний» [1,30].

В последнее десятилетие Интернет прочно вошел во все сферы деятельности нашей жизни, в том числе и в образовательное пространство. В настоящее время в условиях интенсификации обучения русского языка как иностранного (далее – РКИ) остро стоит вопрос об использовании электронных средств обучения, в связи с чем на первый план выходит проблема качества онлайн-ресурсов, а также форм их интеграции в учебный процесс.

Сегодня Интернет предлагает большое количество образовательных ресурсов по РКИ самого разного вида, которые можно найти на сайтах, порталах, в соцсетях, а также функционирующие в виде мобильных приложений.

В настоящей статье рассмотрим два открытых онлайн-курса для обучения русскому языку как иностранному, один из которых предлагает вуз, а второй - компания Russia Today и несколько тестовых сайтов.

«Образование на русском» – один из самых успешных учебных порталов, разработанный государственным институтом русского языка имени А.С. Пушкина. Для регистрации пользователю необходимо через социальные сети или почту

 DOI: 10.36078/1688391447

отправить заявку. После прохождения тестирования обучаемому рекомендуется программа соответствующего уровня. Описание курса элементарного уровня и его структуры дано на русском, а начиная с базового переведено на иностранные языки. Для изучения предлагаются темы, актуальные для указанного уровня.

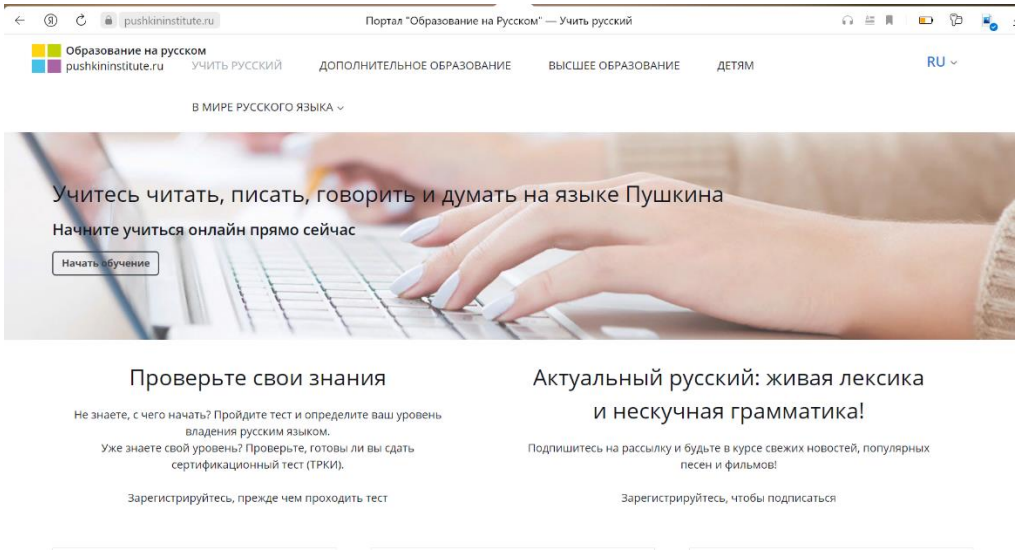


Рис. 1. Портал «Образование на русском»

В качестве плюсов можно отметить наличие множества разнообразных упражнений, всех уровней обучения, вводно-фонетического курса, удобного и яркого интерфейса. В качестве минусов – отсутствие заданий коммуникативной направленности.

Проект «Learn Russian» созданный Russia Today широко популярен во всем мире. Этому способствовало то что учебный материал как фонетический, так и грамматический, дается в интересной и увлекательной форме.

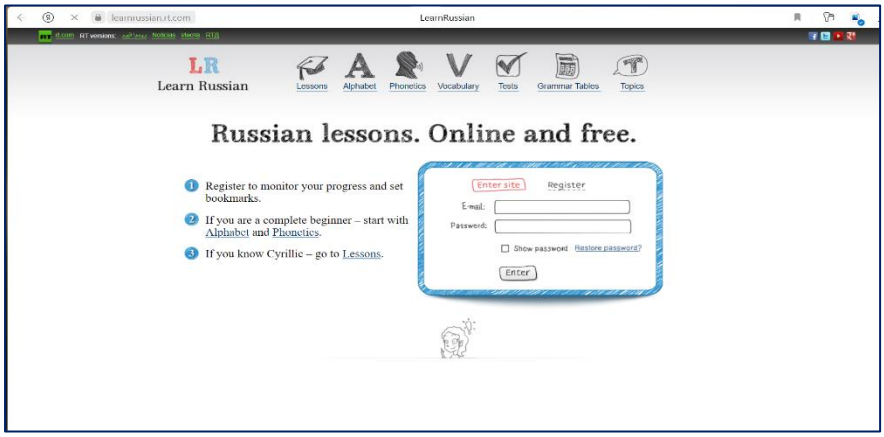


Рис. 2. Сайт «Learn Russian»

DOI: 10.36078/1688391447

Плюсами данного сайта являются разнообразные упражнения, подробное описание грамматического материала, промежуточные тесты для самоконтроля, запрограммированная возможность автоматической проверки тестов при условии, что пользователь зарегистрирован, множество заданий на аудирование и разговорные фразы. Минусы: однотипные задания по грамматике и лексике, нет разделения на уровни, нет заданий коммуникативной направленности.

Из тестовых сайтов считаем необходимым отметить оцифрованные тесты Л.Л. Бабаловой «Практикум по грамматике (РКИ, уровни А2, В1)» (http://rustest-online.ru/soderzhanie/), сборник тестов университета Лос-Анджелеса (UCLA Russian Flagship Language Skills Lounge, http://www.russian.ucla.edu/flagship/russianflagship/Welcome.html), официальные сайты с тренировочными вариантами тестов (testcons.ru, gct-msu.ru).

«Обращение к интернет-ресурсам на занятиях по РКИ реализует личностно-ориентированный подход в обучении, совершенствует навыки и умения в основных видах речевой деятельности, расширяет тезаурус обучающихся и их лингвострановедческие знания о России» [2, 71]. Однако важной проблемой является качество выложенных во всемирную сеть материалов и удобство их использования. Если от ресурсов, представленных на сайтах государственных организаций, можно ожидать соответствия нормативам в сфере ТРКИ, то ресурсы, разработанные отдельными пользователями, могут содержать не только языковые и коммуникативные девиации, но и нарушения морально-этического толка. В связи с этим стоит вопрос о профессиональной компетенции преподавателя РКИ, способного отобрать необходимый интернет-ресурс, который соответствует методическим целям и задачам обучения РКИ.

Для обучения русскому языку иностранных учащихся на этапе довузовской подготовки в Адыгейском государственном университете (далее – АГУ) активно используется специализированный компьютерный класс (лингафонный кабинет), занятия в котором ограничены рамками рабочего времени университета. Наряду с этим учащимся АГУ обеспечивается доступ к Wi-Fi университета с их личных девайсов, который предоставлен им в соответствии со статусом обучаемого. В то же время, согласно проведенному сотрудниками кафедры русского языка как иностранного АГУ опросу (рис. 1), всего лишь 1 % иностранных обучаемых довузовского этапа обучения имеют планшеты и 12% ноутбуки и, соответственно, могут изучать учебный материал вне учебной аудитории.

DOI: 10.36078/1688391447

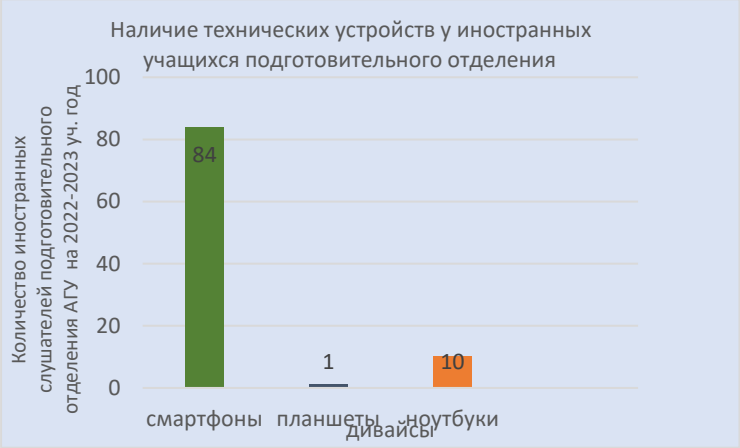


Рис. 3. Наличие девайсов у иностранных учащихся АГУ

При этом наличие смартфона не позволяет говорить о решении данной проблемы, так как смартфон не является в полной мере удобным инструментом с образовательной точки зрения.

Одним из способов решения данной проблемы видится разработка онлайн-курса по РКИ, с доступом к курсу в удобное для слушателя время и с любого девайса, в том числе и с мобильных устройств. Преподаватели кафедры РКИ АГУ в течение последнего года ведут активную работа по созданию своей собственно электронной образовательной среды для иностранных учащихся, обучающихся по дополнительным общеобразовательным программам, обеспечивающим подготовку иностранных граждан и лиц без гражданства к освоению профессиональных программ на русском языке с цель формирования у иностранных слушателей языковой и речевой компетенции в объеме, обеспечивающем возможность осуществлять учебную деятельность на русском языке и необходимом для общения в социально-бытовой, социально-культурной, учебной сферах в рамках уровня В1; заложение основ для дальнейшего совершенствования языковых знаний и умений; расширение образовательного кругозора и проникновение в русскую национальную культуру.

Использованная литература

1. Насейкина Л.Ф. Интерактивные электронные учебники в современном открытом образовании // Вестник Оренбургского государственного университета. 2010. № 5. С. 30–35.

2. Вязовская В.В., Данилевская Т.А., Трубчанинова М.Е. Интернет-ресурсы в обучении русскому языку как иностранному: ожидания vs реальность // Русистика. 2020. Т. 18. № 1. С. 69–84. http://dx.doi.org/10.22363/2618-8163-2020-18-1-69-84

DOI: 10.36078/1688391447

ИЗ ОПЫТА ИСПОЛЬЗОВАНИЯ ДАННЫХ КОРПУСОВ ТЕКСТОВ ДЛЯ ПРОВЕДЕНИЯ КОГНИТИВНЫХ ИССЛЕДОВАНИЙ

Елена Степановна МИЛЬКЕВИЧ
Южный федеральный университет
Кандидат филологических наук, доцент
Институт филологии, журналистики
и межкультурной коммуникации
Российская Федерация
esmilkevich@sfedu.ru

За последние несколько лет парадигма лингвистических исследований сильно изменилась. Этому в значительной степени способствовало создание корпусов текстов различных языков. Корпуса текстов содержат огромный эмпирический материал использования языков в различных стилях, что оказывается несомненно важным для лингвистов. Языковые данные, собранные в корпусах, дают богатый материал для изучения лексикологии, словообразования, грамматики, синтаксис, стилистики, а также для когнитивных научных познаний. Более того, специализированная разметка языковых данных значительно ускоряет процесс отбора материала исследования посредством целого ряда инструментов, таких как: отбор по ключевым словам, частеречный анализ, контекстуальный анализ, анализ повторяемых моделей употребления, анализ сочетаемости (concordance, collocations) и другие.

Теория когнитивной метафоры и когнитивной метонимии в большой степени опирается на обширную базу данных корпусов текстов, так как там представлен аутентичный материал употребления языковых структур. Ученые разрабатывают методологию проведения подобных работ, процедуру отбора валидного материала из корпусов текстов, выдвижение гипотезы и ее верификация на базе данных.

Проведение когнитивных исследований языкового материала, используя данные корпусов текстов получил свое название «корпусной анализ» [Stefanowitsch, 2006]. Он может быть двух видов: а) мотивированные корпусами текстов (corpus-driven) исследования, когда анализ корпуса позволяет выделить определенные закономерности и сформулировать их в виде исследовательской гипотезы; и б) подтверждаемые корпусами текстов (corpus-based) исследования, когда начальная исследовательская гипотеза, основанная на наблюдении, подтверждается в дальнейшем большим эмпирическим материалом корпусов текстов.

DOI: 10.36078/1688391447

Однако в случае изучения когнитивных метафоры и метонимии автоматическая обработка текста является важным только на первом этапе отбора самого материала исследования, используя инструмент отбора по ключевому слову [Милькевич, 2021]. Разрабатываются различные варианты семантической разметки данных, которые помогли бы более быстрому автоматическому отбору материала. Но на сегодняшний день все последующие этапы когнитивного анализа проходят в ручном режиме, то есть, отобранный автоматически материал корпусов текстов с помощью инструмента поиска по ключевому слову KWIC далее анализируется учеными отдельно, с целью выделения случаев истинной когнитивной метафоры или метонимии [Deignan et al, 2004]. Конечно, это представляет ряд проблем, первая из которых ограничения по количеству вторично отобранного материала (из тысячи словоупотреблений ключевого слова исследователь может оставить для дальнейшего анализа менее половины). Еще одна трудность, которую возможно решить в ходе дальнейших разработок корпусов текстов, это определение литературного или нелитературного употребления слов, так как именно последняя группа является материалом для проведения дальнейших когнитивных исследований. Также пока корпуса текстов не могут помочь определить когнитивные источники (source domain / vehicle) и когнитивные цели (target domain) для метафоры и метонимии, что является необходимым этапом для выделения моделей когнитивной метафоры и когнитивной метонимии [Evan, 2007, p.123].

Материалом данного исследования выступают примеры предложений, содержащих лексему *mouth* в современном английском языке. Источником отбора материала послужил BNC British National Corpus (Британский национальный корпус) [British …]. Это одноязычный корпус, построенный на синхронном подходе. Он содержит коллекцию, включающую более 100 миллионов примеров устной и письменной форм современного английского языка, относящихся к разным стилям. Для когнитивного анализа были отобраны первые 500 примеров словоупотреблений данной лексемы.

Как отмечают ряд исследователей, работающих с корпусами текстов, идиоматичные значения, в которых проявляется когнитивная метафора, редко попадают в первые сотни примеров корпуса и для их анализа используют словарные источники словоупотреблений. Поэтому в предлагаемом исследовании мы проводим когнитивный анализ литературного и метонимического значений лексемы *mouth*, оставляя за рамками статьи примеры ее метафорического использования.

В работе применяются как традиционно лингвистические, так и когнитивные методы исследования. Метод анализа словарных дефиниций и контекстуальный

 DOI: 10.36078/1688391447

анализ необходимы для определения основного литературного и нелитературного значений *mouth*. Метод корпусного анализа применяется при отборе материала исследования. С помощью функции поиска по ключевому слову KWIC (Key Word In Context) мы отобрали первые 500 примеров предложений с данной лексемой. Когнитивный анализ позволяет определить когнитивные механизмы, лежащие в основе номинации, выделить концепт-источник и концепт-цель, а также сформулировать основные когнитивные модели, лежащие в основе семантики лексемы *mouth*.

Изучение литературного или прямого денотативного и нелитературного метонимического значений лексемы *mouth* позволило определить когнитивные модели ее семантики. Так, в прямом значении слово mouth предстает как некая емкость (когнитивная модель MOUTH AS A CONTAINER) и как объект, имеющий свою статику и обладающий рядом физиологических характеристик (когнитивная модель MOUTH AS AN OBJECT (STATIC \ PHYSIOLOGICAL). Лексическая сочетаемость здесь ограничена определенными предлогами, глаголами и прилагательными.

Наряду с прямым значением, лексема *mouth* передает также некоторые коннотативные значения, отражающие положительные и отрицательные чувства и состояния человека. Этот факт подтверждает тезис о тесной взаимосвязи внутреннего состояния человека и его внешних физиологических проявлений.

Проведенное исследование позволило выделить три основные модели когнитивной метонимии, где концептом-источником выступает лексема *mouth*: MOUTH FOR ITS PART, MOUTH FOR A PERSON, MOUTH FOR AN ACTIVITY. Метонимические отношения между двумя концептами каждой модели связаны отношениями смежности, так как оба концепта располагаются внутри одной идеализированной когнитивной модели, отражающей рот как орган тела, человека и его внутренний мир, а также деятельность человека.

Использованная литература

1. Милькевич Е.С. (2021). Обучение методам когнитивной лингвистики в магистратуре // Известия ЮФУ. Филологические науки, № 2, с. 182-192.
2. British National Corpus (BNC). [Электронный ресурс]. URL: https://www.english-corpora.org/bnc| (дата обращения 12.01.2023) (in English).
3. Deignan A., Potter L. (2004). A corpus study of metaphors and metonyms in English and Italian // Journal of Pragmatics. Vol. 36, pp. 1231-1252.

DOI: 10.36078/1688391447

4. Evans V. (2007). A Glossary of Cognitive Linguistics. Edinburgh: Edinburgh University Press, 239 p.
5. Stefanowitsch A. (2006) Words and their metaphors: A corpus-based approach // Corpus-Based Approaches to Metaphor and Metonymy. Mouton de Gruyter, pp. 63-105.

 DOI: 10.36078/1688391447

ОНЛАЙН-ПЕРЕВОДЧИКИ В ОБУЧЕНИИ ИНОСТРАННОМУ ЯЗЫКУ (НА МАТЕРИАЛЕ УЗБЕКСКИХ ПОСЛОВИЦ И ПОГОВОРОК)

Анна Васильевна ПОЛОЯН
старший преподаватель
Южный федеральный университет,
Россия
avpoloyan@sfedu.ru

Аннотация. Обучение иностранному языку требует комплексного подхода. Как правило, инструментарий хорошего преподавателя состоит не только из учебных пособий по грамматике языка и лексических материалов. В данном тезисе предложены приемы обучения узбекскому языку с использованием машинного перевода для составления заданий по изучению пословиц и поговорок, культурных тардиций узбекского народа.

Ключевые слова: онлайн-перевод; машинный перевод; автопереводчики; обучение узбекскому языку как иностранному; электронные источники; пословицы и поговорки.

Язык изучается через культуру, и культура также познаётся через языковые особенности. Языковая картина мира любого народа не была бы полна без изучения пословиц и поговорок, хранящих в себе народную мудрость, национальный характер. И в изучении этого весьма интересного пласта культуры есть свои особенности. Несомненно, преподавателю иностранного языка должны быть известны исторические корни происхождения тех или иных устоявшихся высказываний. С течением времени многие ритуалы и традиции уходят из ритма современной жизни, оставаясь в языке лишь в поговорках. Они являются такими же хранителями культурного наследия как и песенный фольклор, классическая литература.

Современные технологии позволяют разнообразить методы обучения иностранному языку. Наряду с электронными источниками аутентичного материала преподавателю и обучающемуся доступен такой инструментарий, как электронные словари и автопереводчики.

Конечно, машинный перевод на данном этапе всё ещё совершенствуется и учеными неоднократно отмечались слабые стороны онлайн-переводчиков для узбекского языка [1-4]. В связи с переходом на латинский алфавит для улучшения машинного перевода предстоит провести ещё много работы по обновлению

DOI: 10.36078/1688391447

словарей, созданию корпусов текстов. Однако, учитывая имеющийся инструментарий, его особенности, опытный преподаватель уже сейчас может взять эту технологию на вооружение для составления заданий по изучению лексики узбекского языка и более глубокого понимания изменений, происходящих в языке.

Г. Х. Тураевой отмечается, что в 2014 г. в системе Google Translate был анонсирован машинный перевод для узбекского языка [3], но в настоящее время с узбекским языком работает около десятка различных онлайн-переводчиков. Ниже перечислены наиболее популярные из них (таб. 1):

Таблица 1. Источники машинного перевода с узбекским языком

N	Электронный источник	Ссылка
1	Google	https://translate.google.com/?hl=ru
2	Yandex	https://translate.yandex.ru/
3	Promt	https://www.translate.ru/
4	Bing (Microsoft)	https://www.bing.com/translator
5	M-Translate	https://www.m-translate.ru/
6	OpenTran	https://opentran.net/allies/scientific-translation.html
7	WebTran	https://www.webtran.ru/

Задания с использованием машинного перевода могут разрабатываться для самостоятельной работы обучающегося, но в последствии полученные результаты должны совместно обсуждаться на занятии и корректироваться преподавателем.

Первый вид заданий: преподавателем отбирается определенное количество пословиц и поговорок, содержащих лингвоспецифическую лексику и культурные реалии, для последующего перевода в различных онлайн-переводчиках и анализа полученного результата.

Таблица 2. Перевод поговорки «*Ish ishtaha ochar, dangasa ishdan qochar*»

N	Переводчик	Перевод (RU)
1	Google	*Работа пробуждает аппетит, ленивый убегает от работы.*
2	Yandex	*Аппетит к работе угасает, лень бежит от работы.*

DOI: 10.36078/1688391447

3	Promt	*Дело открывает аппетит, избегает работы.*
4	Bing (Microsoft)	*Аппетит открывается, ленивой работы избегают.*
5	M-Translate	*Работа пробуждает аппетит, ленивый убегает от работы.*
6	OpenTran	*Работа пробуждает аппетит, ленивый убегает от работы.*
7	WebTran	*Работа открывает аппетит, убегая от работы.*

Как видно из Табл. 2 переводчики по-разному переводят слово «*ochar*». Омонимия или многозначность слова – одна из главных проблем машинного перевода, но детальный анализ такого перевода является стимулом для запоминания различных значений слова, работе обучающегося со словарем, помогает закрепить изучаемую лексику при выборе наиболее удачного перевода из предложенных.

Второй вид заданий: угадать поговорку. Подходит для более продвинутого уровня, когда обучающийся уже знаком с некоторым набором пословиц и поговорок. С помощью автопереводчика или самостоятельно преподаватель «шифрует» поговорки, а обучающийся с помощью обратного перевода и своих фоновых знаний восстанавливает её исходный вид:

Собака лает, караван идет → It hurlaydi, karvon davom etadi (Google) → *It hurar, karvon o´tar* (*Ит ҳурар, карвон ўтар*)

Третий вид заданий: поисковый метод. Подбирается поговорка на узбекском языке, имеющая более глубокий смысл, чем даёт прямой перевод слов. Обучающийся на основании полученного машинным переводом набора слов подбирает аналог в русском языке, а также изучает корни происхождения этой поговорки с помощью дополнительных источников и комментариев преподавателя. Например, «*Bor tovog‘im – kel tovog‘im*» – дословно «*иди моя посуда, вернись моя посуда*» отсылает к обычаю приносить с собой в гости угощение на тарелке, которая затем возвращается также с угощением (https://uforum.uz/showthread.php?p=181735&postcount=4).

Предложенные задания не заменят собой классические упражнения, но разнообразят и дополнят их. Задания с использованием онлайн-переводчиков мотивируют обучающихся к работе с электронными источниками, пробуждают интерес к культуре и быту страны изучаемого языка.

Узбекские народные пословицы и поговорки представлены, например, на портале узбекской литературы ZIYOUZ (более 600 единиц текста, переведенных на

DOI: 10.36078/1688391447

русский язык) https://www.ziyouz.uz/ru/folklor/uzbekskie-narodnye-poslovitsy-i-pogovorki или справочник узбекских поговорок (более 1200 единиц текста, переведенных на русский язык) http://fmc.uz/maqollar.php (кириллический алфавит). Без перевода на другие языки, но в латинском алфавите и с классификацией по категориям узбекские пословицы представлены на портале SIRLAR https://sirlar.uz/ozbek-xalq-maqollar-toplami-barcha-maqollar/

Использованная литература

1. Абдурахманова, Н. Основы автоматического морфологического анализа для машинного перевода / Н. Абдурахманова // Известия Кыргызского государственного технического университета им. И. Раззакова. – 2016. – № 2(38). – С. 12-18.
2. Абдураҳмонова, Н. З. Қ. Замонавий лингвистик корпусларнинг компьютер моделлари / Н. З. Қ. Абдураҳмонова // Узбекистонда хорижий тиллар. – 2020. – No 1(30). – P. 40-48. – DOI 10.36078/1583734626.
3. Тураева, Г. Х. Проблемы машинного перевода при переводе на узбекский язык / Г. Х. Тураева // Universum: технические науки. – 2020. – № 10-1(79). – С. 47-49.
4. Хусаинов, А. Ф. Оценка влияния метода обратного перевода на качество русско-тюркских машинных переводчиков / А. Ф. Хусаинов, Л. Ш. Кубединова // Вестник Тувинского государственного университета. №3 Технические и физико-математические науки. – 2021. – № 4(86). – С. 69-77. – DOI 10.24411/2221-0458-2021-86-69-77.

DOI: 10.36078/1688391447

ПРОЕКТ CHEKHOV DIGITAL: РАЗРАБОТКА ЦИФРОВОГО ИНДЕКСА ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА

Елена Михайловна СЕВЕРИНА
доктор философских наук, профессор
Южный федеральный университет,
Российская Федерация
emkovalenko@sfedu.ru

Аннотация. Рассмотрена специфика разработки цифрового указателя (индекса) имен и названий реальных людей и объектов, упоминаемых в текстах произведений и писем А. П. Чехова и представленных в указателях академического издания. Разработка такого индекса позволяет организовать семантический поиск по текстам произведений писателя, редакционно-критическому аппарату цифрового издания Chekhov Digital.

Ключевые слова: цифровое издание; проект Chekhov Digital; семантический поиск; цифровой индекс; указатель имен и названий.

Исследование выполнено в рамках соглашения о научном сотрудничестве между Южным федеральным университетом (ЮФУ) и Национальным исследовательским университетом "Высшая школа экономики" (НИУ ВШЭ) («Зеркальные лаборатории НИУ ВШЭ»), проект № 6.13.1-02/250821-1 «Конвергенция языковых пластов русского языка в зеркале цифровых решений».

Полное собрание произведений Антона Павловича Чехова было напечатано и опубликовано в период с 1974 по 1982 год (в 30 томах, объем около 46 000 страниц). Проект Chekhov Digital – это издание нового типа, реализованное в формате стандарта TEI/XML и снабженное семантической разметкой, что открывает новые возможности для академических исследований в цифровом формате и использования литературных текстов в цифровых проектах [2]. Работа над проектом осуществляется Центром цифровых гуманитарных исследований Института филологии, журналистики и межкультурной коммуникации ЮФУ совместно с Международной лабораторией языковой конвергенции НИУ ВШЭ и лабораторией филологии ЮНЦ РАН.

Полное собрание сочинений и писем А. П. Чехова в 30 томах (далее ПССиП) [4] представляет собой академическое издание текстов произведений и писем писателя, включая ранние редакции и различные варианты текста. Подготовка

DOI: 10.36078/1688391447

данного издания была связана со скрупулезной текстологической работой с сохранившимися источниками для исправления цензурных, типографских, редакторских искажений, накопившихся в предыдущих редакциях. Были изучены архивные фонды, журналы и газеты, в которых писатель мог публиковаться, в том числе с использованием неизвестных псевдонимов и/или анонимно, а также неопубликованные записные книжки и дневники, найденные рукописи писателя. Издание состоит из двух серий: тексты произведений писателя - Сочинения (тт. I—XVIII) и эпистолярное наследие А. П. Чехова - Письма (тт. I—XII). Каждый том содержит специальные разделы, в которых представлены тексты незавершенных произведений писателя, варианты текстов и их первоначальные редакции. Тексты издания печатаются «по правилам современной орфографии и пунктуации, с сохранением индивидуальных особенностей, свойственных языку Чехова» [1, с. 7].

Цифровое семантическое издание Chekhov Digital наследует подходу, реализованному в рамках электронного научного издания творческого наследия А. П. Чехова (ЭНИ «ЧЕХОВ») Фундаментальной электронной библиотеки «Русская литература и фольклор» (ФЭБ) [5], в котором основным элементом представления текстов является не том (набор произведений и редакционно-критических материалов), а отдельное произведение как «самодостаточный фрагмент печатного издания (например, рассказы «Лошадиная фамилия» или «Унтер Пришибеев», входящие в состав 4-го тома академического Полного собрания сочинений и писем Чехова в 30 томах)» [5].

Редакционно-критический аппарат академического издания [4] играет важную роль для разработки семантического поиска. Он включает в себя историко-литературные и текстологические комментарии о произведениях писателя, описывающие историю их создания, цензурирования, источники текста, изменения, вносимые в основной текст на основе рукописей и авторизованных печатных изданий. Примечания содержат отзывы прижизненной критики и мнения современников о чеховских произведениях, информацию о творческой и сценической истории его пьес и переводах на иностранные языки [1, с. 7-8]. Семантическая машиночитаемая разметка разрабатывается также для этих материалов, что позволяет исследовать их с использованием компьютерных методов.

Академическое издание произведений и писем А. П. Чехова [4] – это не только каноничность представленных текстов писателя и тексты комментариев/примечаний специалистов-исследователей жизни и творчества писателя, но это еще и набор упоминаний в текстах реально существовавших людей,

DOI: 10.36078/1688391447

дат, событий, ситуаций и т. п. Поэтому издание содержит ряд составленных специалистами указателей, которые давали возможность извлечь эти данные из текстов в доцифровую эпоху, и сейчас представляют интерес с точки зрения организации семантического поиска по текстам произведений и писем А. П. Чехова, редакционно-критическому аппарату. К ним в первую очередь следует отнести указатели имен и названий, упомянутых в текстах писателя, комментариях/примечаниях, которые встречаются в 14/15, 16, 17 тт. серии Сочинения, сводный указатель имен и названий в 18 тт. и указатели имен и названий к каждому из 12 томов серии Письма. На основе этих указателей разрабатывается база данных имен и названий, упоминаемых в текстах писателя и примечаниях/комментариях издания, в которой содержится не только оцифрованный вариант указателей печатного издания, т.е. перевод результатов традиционных филологических практик в цифровой формат, но и дополнительная информация из других баз данных, таких как Wikidata (https://www.wikidata.org), а следовательно, данные приобретают новые специфические свойства. Разрабатываемая база данных имен и названий - цифровой указатель (индекс), наследует функциональность печатного указателя, содержащего информацию о томах и страницах упоминания соответствующего имени/названия в текстах произведений ПССиП с учетом связи имен и названий друг с другом, что позволяет реализовать семантический поиск по текстам произведений писателя и редакционно-критическому аппарату ПССиП. Подготовленные в таком формате данные становятся инструментом не только для новых филологических практик, но и позволяют получить представление о работе с большими коллекциями литературных текстов.

Цифровой указатель имен и названий писем содержит 3040 записей о названиях и 7798 записей об именах реальных людей, из них 683 записи (10 названий) – общие с записями из указателя имен и названий Л. Н. Толстого [6] (меньше всего в 1904 году (12 том писем) – 16 записей, а больше всего в 1887-1888 гг. (2 том писем) – 104 записи). Указатель имен и названий сочинений содержит 5177 записей о названиях и 8934 записи об именах реальных людей, собранных автоматически из указателей 14/15, 16, 17 и 18 томов. Оцифрованный индекс расширяет возможности доцифрового традиционного поиска по индексу за счет «нечеткого поиска» по части слова/имени, быстрого поиска через просмотр всего списка, быстрого перехода на страницу соответствующего произведения, письма или примечания/комментария. При этом появляется возможность связать не только

 DOI: 10.36078/1688391447

конкретную страницу с соответствующей записью в БД, но и найти выражение соответствующей сущности на странице.

Кроме того, на основе данных из оцифрованного индекса может быть построен частотный список встречаемости каждой сущности на страницах текстов произведений и писем писателя, в комментариях и примечаниях; исследовать появление одних и тех же имен и названий в разных контекстах. Одновременное появление различных имен и названий в одном абзаце, на одной странице или в любом другом более широком или узком контексте дает возможность выявить взаимодействие между разными фрагментами текстов писателя и примечаний/комментариев, что позволяет получить представление о социальных связях писателя, причем как в контексте взаимодействия писателя с другими людьми, например, адресаты и герои его писем, так и в контексте точки зрения А. П. Чехова на связи реальных людей и объектов друг с другом. Такого рода структуры предоставляют новые возможности для изучения идей и интересов А. П. Чехова, общих тенденций и тем его творчества и эпистолярия. Для каждого имени/названия из указателя можно выявить наиболее значимые связанные с ним имена.

Цифровой индекс также позволяет понять, каким образом создавался редакционно-критический аппарат ПССиП, влияние различных редакторов на формирование академического издания, выявить неточности и различия в подходах как отдельных томов издания, так и серий – сочинений и писем. Например, каждый том писем снабжен своим указателем имен и названий людей и объектов, упоминаемых в письмах и комментариях, в то время как для Сочинений принят другой подход – общий указатель представлен в 18 т. ПССиП, но тома, содержащие нехудожественную прозу, также содержат свои указатели, которые дополняют описания сущностей из 18 тома, в котором чаще всего указана краткая информация о человеке или объекте, с сокращениями имен/названий (насколько возможно). Например, в указателе 18 т. есть следующее описание брата писателя «*Чехов Ал. П. - Чехов Ал. П. (псевдонимы — Агафопод Единицын; А. Ед.; Ч (?); А. Седой; Алоэ)*», а в указателе 17 тома – более полное описание «*Чехов (псевдонимы — Алоэ, Агафопод Единицын, А. Седой) Александр Павлович (1855—1913), брат Чехова, писатель, журналист; автор ряда воспоминаний о Чехове*». Возможно это связано с объемностью, представленной в указателе 18 тома информации и предполагает самостоятельную работу исследователя по поиску необходимых данных. Цифровой указатель решает эту проблему, однако требует «ручной» верификации автоматически собранной информации.

DOI: 10.36078/1688391447

Цифровой индекс проекта Chekhov Digital [3] сохраняет все свойства традиционного индекса и в то же время расширяет исследовательские возможности за счет компьютерного управления информацией и семантической поисковой системы проекта.

Использованная литература

1. От редакции // Чехов А. П. Полное собрание сочинений и писем: В 30 т. Сочинения: В 18 т. М., 1974—1982. Т. 1. [Рассказы. Повести. Юморески], 1880—1882. М.: Наука, 1974. С. 5-8.
2. Северина Е.М., Бонч-Осмоловская А.А., Кудин А.М. Цифровые филологические практики: проект "Chekhov Digital". Актуальные проблемы филологии и педагогической лингвистики. 2022. №2. С. 153-165. DOI: 10.29025/2079-6021-2022-2-153-165.
3. Цифровой проект Chekhov Digital. URL: http://chekhov-digital.sfedu.ru/ (дата обращения: 08.05.2023).
4. Чехов А.П. Полное собрание сочинений и писем: В 30 т. АН СССР. Ин-т мировой лит. им. А. М. Горького. М.: Наука. 1974-1983. URL: http://feb-web.ru/feb/chekhov/default.asp?/feb/chekhov/texts/che-te02.html (дата обращения: 08.05.2023).
5. ЭНИ «Чехов» / Фундаментальная электронная библиотека «Русская литература и фольклор. URL: http://feb-web.ru/feb/chekhov/default.asp (дата обращения: 08.05.2023).
6. Orekhov B., Fischer F. The 91st Volume - How the Digitised Index for the Collected Works of Leo Tolstoy Adds A New Angle for Research, in: Digital Humanities 2018: Book of Abstracts / Libro de resúmenes. Mexico: Red de Humanidades Digitales A. C., 2018. P. 465-466.

DOI: 10.36078/1688391447

ЛИНГВИСТИЧЕСКИЙ КОРПУС В ДЕЯТЕЛЬНОСТИ ЛИНГВИСТА И ПЕРЕВОДЧИКА

Андрей Александрович СОХАНЬ
кандидат филологических наук, доцент,
Пятигорский государственный университет
sochan2001@mail.ru

Анна Александровна ПАШКОВА
магистрант Института переводоведения,
русистики и многоязычия
Пятигорский государственный университет
anna.psh2000@gmail.ru

Аннотация. Тезисы посвящены использованию лингвистических корпусов в деятельности лингвистов и переводчиков. Большое внимание уделяется применению включенных в лингвистические корпуса инструментов. В центре внимания производные прилагательные, образованные при помощи полусуффиксов типа -wert в современном немецком языке.

Деривативное словопроизводство в современном немецком языке представляет собой один из самых продуктивных способов словообразования, выделяются как суффиксальные, так и префиксальные модели. Образование новых лексических единиц при помощи аффиксов освещено в лингвистической литературе довольно подробно, однако основной акцент делается либо на словообразовательные модели как таковые, либо на словообразовательное значение, позволяющее выделять определенные группы слов со сходной семантикой. Подобных классификаций множество, но большинство из них оставляет без ответа прагматическую составляющую.

Словарный состав немецкого языка насыщен дериватами анализируемого типа, зачастую речь идет о сотнях лексических единиц, частотность употребления которых варьируется в различные эпохи под влиянием всевозможных факторов. Использование лингвистического корпуса позволяет нам не только проследить изменения семантики лексических единиц по периодам, но и установить валентные связи этих единиц, определить сферы их употребления и влияние, которое они оказывают на формирование современной языковой картины мира, что имеет определенную ценность и для лингвистических исследований, и для переводческой деятельности.

DOI: 10.36078/1688391447

В рамках нашего исследования мы использовали лингвистический корпус Der deutsche Wortschatz von 1600 bis heute (DWDS), содержащий подборку материалов на немецком языке с 17 века.

В рамках тезисов рассмотрим лексические единицы с компонентом -wert. Полусуффикс -wert рассматривается как частично синонимичный полусуффиксу -würdig. Грамматический справочник Дуден лишь вскользь упоминает, что производные с вышеупомянутыми полусуффиксами выражают некоторую модальность, а именно определенную степень долженствования. [3, 550]

Все рассматриваемые в статье адъективные производные в качестве первого компонента имеют глагол. Анализ при помощи лингвистического корпуса показывает, что в большинстве случаев используется инфинитивная форма глагола, соединяющаяся с полусуффиксом -wert при помощи интерфикса -s-. Все глаголы можно разбить на группы, которые тематически будут соответствовать также и дериватам с -wert: 1. обратить внимание, учесть, отметить - achten, beachten, bemerken, merken, berücksichtigen, erwägen, beherzigen; 2. слышать, видеть - ansehen, sehen, hören; 3. ценить, хвалить, чтить, одобрять, награждать, признавать - ehren, schätzen, rühmen, loben, begrüßen, billigen, belohnen, anerkennen; 4. следовать за чем-то, стремиться к чему-то, подражать - anstreben, erstreben, befolgen, begehren, nacheifern, nachahmen; 5. ругать, жалеть, обвинять - fluchen, bedauern, beklagen, bejammern, tadeln, erbarmen; 6. рассказывать, упоминать, называть - erzählen, erwähnen, nennen, mitteilen; 7. рекомендовать – empfehlen; 8. сочувствовать, завидовать, презирать, ненавидеть, любоваться чем-то - bemitleiden, beneiden, verachten, bewundern, verabscheuen и т.д.

Предложенное тематическое распределение крайне условно и может быть совершенно другим. Мы рассматриваем лишь наиболее частотные глаголы, участвующие в образовании производных прилагательных с компонентом -wert. Выборка примеров из DWDS показывает, что 68 из 300 лексических единиц с компонентом -wert относится к анализируемой модели, а это говорит о продуктивности данной словообразовательной модели, поскольку остальные примеры не являются прилагательными или же в качестве первого компонента используется существительное, например, preiswert.

Существительное Wert имеет довольно обширный семантический спектр:

1. Денежный эквивалент, стоимость продукта, предмета, услуги, ср.:

… meine Mutter kaufte für das soeben erworbene Geld Waren ein, denn wenige Stunden später hatte das Geld schon wieder an Wert verloren [4]

 DOI: 10.36078/1688391447

2. Совокупность положительных качеств чего-то или кого-то, их значение, важность по сравнению с чем-то, ср.: der künstlerische, unvergängliche, dokumentarische Wert eines Romans; der sittliche, ethische, moralische, ideelle, gesellschaftliche, erzieherische Wert einer Tat; der technische Wert einer Erfindung; das hat wenig, geringen, viel, einen hohen, unschätzbaren Wert (für meine Arbeit); einer Sache großen, keinen Wert beilegen, beimessen и т.д.

3. Предметы или вещи, имеющие большую материальную ценность; имущество:

Im Hochwassergebiet konnten viele materielle Werte vor der Zerstörung gerettet werden [4]

4. Выраженная в цифрах математическая или физическая величина или значение: der Wert einer physikalischen Größe; der Wert der Messung liegt um 25 Prozent höher als erwartet; Werte vergleichen, (in eine Gleichung) einsetzen; einen Wert (an einer Skala) ablesen; die Werte berechnen, erhöhen, vergrößern, verringern, verkleinern; meteorologische, mathematische, technische Werte и мн. др.

5. Почтовая марка с напечатанной стоимостью:

Die Serie wird mit einem Wert zu zehn Cent und einem zu fünfzig Cent fortgesetzt.
Auf Werten von fünf bis fünfundfünfzig Cent werden Zugvögel dargestellt [4]

6. Биржевые акции:

Er kaufte verschiedene Werte [4]

Широта семантического спектра определяет и валентные свойства существительного Wert, поэтому кроме самостоятельного употребления мы находим в DWDS примеры, в которых первоначальное значение несколько ослабевает, то есть существительное Wert переходит в разряд так называемых полусуффиксов и используется в составе производных лексических единиц с -wert в качестве второго компонента. При этом практически все дериваты оказываются объединены общим словообразовательным значением «быть достойным/стоить того, что выражено первым глагольным компонентом». Таким образом, адъективные производные могут быть распределены по тем же группам, что и приведенные выше глаголы:

1. быть достойным того, чтобы бы обратили внимание, учли, отметили:

Die Entscheidung von Herrn Dr. Voscherau ist achtenswert und hat meinen Respekt! [4]

2. быть достойным того, чтобы быть услышанным, увиденным:

Ihre Musik ist nicht unbedingt radiotauglich, eher anstrengend, aber durchaus hörenswert. [4]

DOI: 10.36078/1688391447

3. быть достойным оценки, похвалы, одобрения, награды, признания, рекомендации:

Vielleicht wird man einige von euch, ganz sicher sogar den grauen Rudi, schon jetzt als ehrenwerte ältere Person behandeln. [4]

4. быть достойным подражания, желания следовать или стремиться к этому:

Der Job, den er da anstrebt, ist nicht wirklich anstrebenswert. [4]

5. быть достойным порицания, сожаления, обвинения:

Wie diese bedauernswerte Frau selbst es von früh auf geahnt hatte, sollte Erfüllung ihrem Leben nicht beschieden sein. [4]

6. быть достойным повествования, упоминания:

Die ganze Geschichte wäre erzählenswert genug, würde sie an dieser Stelle nicht den Rahmen sprengen. [4]

7. быть достойным сочувствия, зависти, презрения, ненависти, восхищения:

Die Wärter, die während der ganzen Nacht freigekommene Raubtiere hatten erschießen müssen, boten einen bemitleidenswerten Anblick. [4]

В тезисах предлагается лишь краткий обзор словообразовательной модели „Verb+wert“, по которой в немецком языке на протяжении уже достаточно долгого времени образуются производные прилагательные. В примерах из лингвистического корпуса немецкого языка DWDS дериваты, образованные по данной модели, употребляются в качестве определения или наречия, то есть используются либо для определения одного из членов предложения, выраженного существительным, либо для определения сказуемого. Таким образом, лингвистический корпус дает возможность отслеживать не только семантику, но и синтаксическую функцию рассматриваемых лексических единиц. Кроме того, еще один инструмент корпуса – кривые, показывающие частотность употребления включенных в корпус слов – дают нам представление о популярности того или иного слова или целого ряда слов, объединенных общим словообразовательным значением.

Приведенные ниже графики наглядно демонстрируют, что в современном немецком языке употребление производных прилагательных с полусуффиксом -wert, глагольный компонент которых окрашен положительно, в последнее десятилетие значительно сократилось, например:

DOI: 10.36078/1688391447

База примеров DWDS включает не только художественную литературу, но и газеты, журналы с достаточно обширным для проведения лингвистического или предпереводческого анализа периодом поиска. Так, некоторые газеты и журналы, из которых были использованы примеры употребления анализируемых лексических единиц, датированы началом прошлого века, некоторые относятся к концу 19 века. Это позволяет говорить о том, что данная словообразовательная модель продуктивна на протяжении довольно продолжительного времени, частота

DOI: 10.36078/1688391447

употребления той или иной производной с полусуффиксом -wert зависит от предпочтений носителей языка, обусловленных различными, в том числе и экстралингвистическими факторами - политическая ситуация, доминирование в обществе тех или иных этических, моральных и др. ценностей и т.д. Широкий семантический спектр существительного Wert практически не ограничивает его валентных возможностей при употреблении в качестве полусуффикса. Он может сочетаться практически с любым глаголом в современном немецком языке при условии сохранения логических связей, что делает его своего рода универсальным словообразовательным средством.

Использованная литература

1. Большой немецко-русский словарь [Текст]: в 3 т./под. рук. О.И. Москальской. – 7-е изд. – М.: Русский язык, 2001.
2. Локтионова, В.Г. Проблема репрезентации наблюдаемых и ментальных сущностей средствами языка. - Издательство: Пятигорский государственный университет. – Пятигорск, 2018. – С. 74-82.
3. Duden. Deutsches Universalwörterbuch A-Z. 3., völlig neu bearb. u. erw. Aufl. – Mannheim; Leipzig; Wien, Zürich: Dudenverlag, 1996.
4. DWDS - Digitales Wörterbuch der deutschen Sprache [Электронный ресурс] – URL: https://www.dwds.de/wb/devisenträchtig
5. Fleischer, W., Barz I. Wortbildung der deutschen Gegenwartssprache. 4., völlig neu bearbeitete Auflage. Walter de Gruyter Tübingen, 2012.
6. Leipzig Corpora Collection – Wortschatz Deutsch [Электронный ресурс] – URL: https://corpora.uni-leipzig.de/de?corpusId=deu_newscrawl-public_2018
7. Muthmann, G. Rückläufiges deutsches Wörterbuch. Handbuch der Wortausgänge im Deutschen, mit Beachtung der Wort- und Lautstruktur. 2. Auflage. Max Niemeyer Verlag. Tübingen, 1991.

DOI: 10.36078/1688391447

TO‘G‘RI CHIZIQ VA TEKISLIKNING VEKTOR-PARAMETRLI BERILISHINING TATBIQLARIGA DOIR MASALALAR

Rokiya Begmatovna SALIBAYEVA
Angren 2-sonli KHM matematika
fani o‘qituvchisi

Annotatsiya. Tezis matematikaning keng qo‘llaniladigan tushunchalaridan biri bo‘lgan vektor tushunchasini o‘qitish masalalariga bag‘ishlangan.

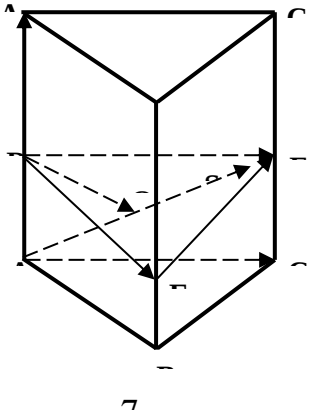
Kalit so‘zlar: vector; vektorlar ustida amallar; geometrik figuralar; to‘g‘ri chiziq, tekislik.

Ma’lumki, vektor tushunchasi zamonaviy matematikaning eng muhim tushunchalaridan biridir. Shuning uchun ham matematikadan o‘quv dasturida vektorlar ustida amallar bajarish malakalarini egallashni ta’minlash ko‘zda tutilgan bo‘lib, bunday mashqlarni bajarishda asos bo‘lib ko‘p hollarda geometrik figuralarning xossalari xizmat qiladi. Vektor metodining bir qator geometrik masalalarni yechishda, teoremalarni isbotlashda qo‘llanishini ko‘rish maqsadlidir. Juda ko‘p geometrik masalalarni yechish uchun sodda nuqtalar to‘plamlarini (to‘g‘ri chiziqlar, nurlar, kesmalar, tekisliklar, burchaklarni) vektorlar orqali ifodalash maqsadga muvofiq. Shu usulda yechiladigan stereometrik masalalarga misollar keltiramiz.

Masala 1. $ABCA_1B_1C_1$ uchburchakli prizmaning AA_1, BB_1, CC_1 qirralarida mos ravishda shunday D, E, F nuqtalar olinganki, ular

$$\overrightarrow{AD} = k \cdot \overrightarrow{AA_1}, \qquad \overrightarrow{BE} = m \cdot \overrightarrow{BB_1}, \qquad \overrightarrow{CF} = n \cdot \overrightarrow{CC_1}$$

1shartlarni qanoatlantiradi. Prizmaning A uchi va BB_1C_1C yoqining simmetriya markazi S nuqta orqali o`tkazilgan l to ‘g ‘ri chiziq DEF tekislik bilan O nuqtada kesishadi. $x = AO : AS$ nisbatni toping.



Yechish. Masalada qaraladigan vektorlarni $\overrightarrow{AB} = \overrightarrow{e_1}, \overrightarrow{AC} = \overrightarrow{e_2}, \overrightarrow{AA_1} = \overrightarrow{e_3}$ lar bilan belgilaymiz (7-rasm).

U holda $\overrightarrow{AD} = k\overrightarrow{e_3}, \overrightarrow{AE} = \overrightarrow{e_1} + m\overrightarrow{e_3}, \overrightarrow{AF} = \overrightarrow{e_2} + n\overrightarrow{e_3}, \overrightarrow{AS} = \frac{1}{2}(\overrightarrow{e_1} + \overrightarrow{e_2} + \overrightarrow{e_3})$,

$\overrightarrow{AO} = p\overrightarrow{AS}$ bo`ladi.

Bularga ko`ra esa,

$\overrightarrow{DE} = \overrightarrow{e_1} + (m-k)\overrightarrow{e_3}, \overrightarrow{DF} = \overrightarrow{e_2} + (n-k)\overrightarrow{e_3}$,

 DOI: 10.36078/1688391447

$$\overrightarrow{DO} = \overrightarrow{AO} - \overrightarrow{AD} = p\overrightarrow{AS} - \overrightarrow{AD} = \frac{P}{2}(\overrightarrow{e_1} + \overrightarrow{e_2}) + \left(\frac{P}{2} - k\right)\overrightarrow{e_3}.$$

vektor tengliklarni hosil qilamiz. Bu uchta vektorlar komplanar bo`lganligi uchun vektorlarning komplanarlik shartiga ko`ra $\overrightarrow{DO} = x \cdot \overrightarrow{DE} + y \cdot \overrightarrow{DF}$ tenglikni yozish mumkin. Bu tenglikka $\overrightarrow{DO}, \overrightarrow{DE}$ va $\overrightarrow{DF}$ vektorlar yuqorida topilgan ifodalarini qo`ysak,

$$\frac{P}{2}(\overrightarrow{e_1} + \overrightarrow{e_2}) + \left(\frac{P}{2} - k\right)\overrightarrow{e_3} = x[\overrightarrow{e_1} + (m-k)\overrightarrow{e_3}] + y[\overrightarrow{e_2} + (n-k)\overrightarrow{e_3}]$$

bo`ladi. Lekin $\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}$ vektorlar komplanar emas. Shuning uchun bu tenglikning chap va o`ng tomonlaridagi $\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}$ vektorlarga ko`ra yoyilmalar teng (vektorning berilgan bazisga ko`ra yoyilmasining yagonalik xossasi) bo`ladi:

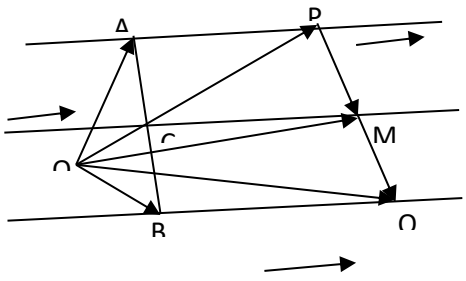
$$\frac{p}{2} = x, \qquad \frac{p}{2} = y, \frac{p}{2} - k = (m-k)x + (n-k)y.$$

Bu yerdan $\frac{p}{2} - k = \frac{p}{2}(m-k) + \frac{p}{2}(n-k)$ yoki $\frac{p}{2}(1 + 2k - m - n) = k$ ni topamiz. Agar bu yerda $1 + 2k - m - n \neq 0$ bo`lsa, O nuqta mavjud bo`lib, izlanayotgan nisbat $\frac{AO}{AS} = \frac{2k}{1+2k-m-n}$ ga teng bo`ladi.

Bu yerda *k, m, n* larga tayin sonli qiymatlar berib va S nuqtaning BB_1C_1C yoqdagi vaziyatini turlicha tanlab, turli xususiy hollarga kelinadi. DEF tekislik AS to ʻg ʻri chiziqqa parallel bo`lishi uchun $m + n = 2k + 1$ shartning bajarilishi lozimligi ravshan. Xususiy hollarning biri, $k = 0$ va $m + n = 1$ bo`lganda DEF tekislik AS to ʻg ʻri chiziq orqali o`tadi.

Masala-2. Tekislikda ikkita $\vec{p}$ $\vec{q}$ to ʻg ʻri chiziqlar

$\overrightarrow{OP} = \overrightarrow{OA} + k\vec{Q}$ va $\overrightarrow{OQ} = \overrightarrow{OB} + n$ (1)



8-rasm

(bu yerda nolmas $\vec{a}$ va $\vec{b}$ vektorlar mos ravishda p va q to ʻg ʻri chiziqlarga parallel bo`lgan ixtiyoriy vektorlar) vektorlar tenglamalar bilan berilgan. M nuqta PQ kesmani PM: MQ = ƛ nisbatta bo`ladi. k va n lar mumkin bo`lgan barcha teng qiymatlarni qabul qilgan hol uchun M nuqtalar nuqtalar to`plamini toping.

Yechish. Vektorlarni ayirish qoidasiga ko`ra (8-rasm)

$\overrightarrow{PM} = \overrightarrow{OM} - \overrightarrow{OP}$ va $\overrightarrow{MQ} = \overrightarrow{OQ} - \overrightarrow{OM}$

Masalaning shartiga ko`ra $\overrightarrow{PM} = ƛ\overrightarrow{MQ}$ yoki

DOI: 10.36078/1688391447

$\overrightarrow{OQ} - \overrightarrow{OP} = \lambda(\overrightarrow{OQ} - \overrightarrow{OM})$. Bu yerdan $\overrightarrow{OM} = \frac{\overrightarrow{OP}+\lambda\overrightarrow{OQ}}{1+\lambda}$ ni topamiz. Bu yerdagi $\overrightarrow{OP}$ va $\overrightarrow{OQ}$ larning o`rniga ularning (1) dagi ifodalarini qo`yib, soddalashtiramiz. $\overrightarrow{OM} = \frac{\overrightarrow{OA}+\lambda\overrightarrow{OB}}{1+\lambda} + \frac{(\vec{a}+\lambda\vec{b})k}{1+\lambda}$ yoki $\overrightarrow{OM} = \overrightarrow{OC} + k\vec{c}$ ga ega bo`lamiz, bu yerda $\overrightarrow{OC} = \frac{\overrightarrow{OA}+\lambda\overrightarrow{OB}}{1+\lambda}$, $\vec{c} = \frac{(\vec{a}+\lambda\vec{b})}{1+\lambda}$.

(1) Muntzam izlanayotgan M nuqtalar to‘plami C nuqta orqali o‘tib, $\vec{c}$ vektorga parallel bo`lgan to‘g‘ri chiziqlardan iborat ekanligini ko‘rsatadi.

(2) Masalaning yechimidan ko‘rinadiki, berilgan p va q to‘g‘ri chiziqlar ayqash bo‘lsa ham M nuqtalar to‘plami to‘g‘ri chiziqdan iborat bo‘ladi.

Geometrik masalalarni yechishda vektor metodni qo‘llashning eng asosiy ustunligi uning umumlashtirishlar qilishga imkon berishdir. Bunday umumlashtirishlarda qaralayotgan figuralarning necha o‘lchovli ekanligidan qat’i nazar ularni bog‘lovchi munosabatlar o‘zgarmaydi. Bunday hollarni mazkur yuqorida zikr etilgan mulohazalarda aylana va sfera, o‘rta perpendikulyar to‘g‘ri chiziq va tekisliklarning vektor tenglamalarini tuzishga doir masalalarni yechishda ko‘rish mumkin.Umuman, yassi va fazoviy nuqtalar to‘plamlarini topishda vektorlar algebrasining qo‘llanishini algebra kursida qaraladigan tenglamalar tuzishga doir masalalar bilan qiyoslash mumkin.

Foydalanilgan adabiyotlar

1. Sayfullayeva H. M. “Geometriya”. Akademik litsey va kasb-hunar kollejlari uchun T., O‘qituvchi, 2007-yil.
2. Василевских А. Б. “Обучение решению задач”. Минск, Вышейшая школа, 1979 год.

DOI: 10.36078/1688391447

THE DEVELOPMENT OF THESAURUS DICTIONARIES AND THEIR FUNCTIONS

Mohira Fayzullayevna SAPAROVA,
Faculty of Philology
Department of Roman-Germanic
Philology, an English teacher
Mamun university NEI

Abstract. Thesaurus dictionaries are an essential tool for language learners, writers, researchers, and information managers, providing a means to expand vocabulary, improve writing skills, and enhance information retrieval. However, the effectiveness of thesaurus dictionaries depends on their organization, content quality, and appropriate use in context. This thesis explores the history, development, and applications of thesaurus dictionaries, and evaluates their effectiveness as a tool for language learning, writing, and knowledge management.

Key words: Roget Thesaurus; researchers; specific; management; factors; rhetoric; communication; relationship; thematically; approach.

The development of thesaurus dictionaries has been shaped by several factors, including advances in language studies, changes in publishing technology, and the growing demand for effective tools for language learning and knowledge management. The earliest thesaurus dictionaries were simple lists of synonyms and antonyms, compiled for use in rhetoric and poetry. These lists were often organized alphabetically, without any consideration for the relationships between words.

In the 19th century, Peter Mark Roget published the first modern thesaurus dictionary, Roget's Thesaurus of English Words and Phrases. Roget's thesaurus was organized thematically, with words and phrases grouped according to their meanings and relationships. This organizational approach allowed users to explore related concepts and find alternative words and phrases more easily. In the 20th century, thesaurus dictionaries became more specialized, with many dictionaries developed for specific fields, such as medicine, law, and science [7,23-26]. These specialized dictionaries included technical terms and jargon specific to their respective fields, allowing researchers and professionals to communicate more effectively.

With the advent of digital technology, thesaurus dictionaries have become more widely available and easier to utilize. Electronic thesauruses, such as WordNet, have been developed to facilitate natural language processing and information retrieval. These

 DOI: 10.36078/1688391447

electronic thesauruses use algorithms to analyze and categorize words and concepts, allowing users to quickly find synonyms and related words. Therefore, the development of thesaurus dictionaries has been driven by a desire to improve language learning, writing, and knowledge management. [2,67-69] The ongoing development and use of thesaurus dictionaries highlight their enduring value as a tool for language studies and communication.

The history of thesaurus dictionaries can be traced back to ancient Greece, where lists of synonyms and antonyms were compiled for use in rhetoric and poetry. However, it was not until the 19th century that thesaurus dictionaries began to take their modern form. In 1805, the first modern thesaurus dictionary was published by Peter Mark Roget, a British physician and polymath. Roget's Thesaurus of English Words and Phrases was organized thematically, with words and phrases grouped according to their meanings and relationships. Roget's Thesaurus quickly became popular and has been continuously updated and revised since its initial publication. In the United States, the first thesaurus dictionary was published by Samuel Johnson in 1755. Johnson's dictionary was not organized thematically, but rather alphabetically, with synonyms and related words listed under each entry. In the 20th century, thesaurus dictionaries became an important tool for writers, editors, and researchers. [3,79-82] Many specialized thesaurus dictionaries were developed for specific fields, such as medicine, law, and science.

With the advent of digital technology, thesaurus dictionaries have become more widely available and easier to use. Electronic thesauruses, such as WordNet, have been developed to facilitate natural language processing and information retrieval. In general, the history of thesaurus dictionaries reflects the evolution of language and the changing needs of writers, researchers, and information managers. The ongoing development and use of thesaurus dictionaries demonstrate their enduring value as a tool for language learning, writing, and knowledge management. The aim of thesaurus dictionaries is to provide users with a comprehensive and organized list of synonyms, antonyms, and related words for a given term. By doing so, thesaurus dictionaries aim to expand users' vocabulary, improve their writing and communication skills, and facilitate their ability to analyze and understand language and information.

Thesaurus dictionaries also aim to provide users with a means to navigate complex fields of knowledge by mapping the relationships between different concepts and terms. This can be useful for researchers, students, and information managers who need to understand and organize large amounts of information. Thus, the aim of thesaurus dictionaries is to provide users with a tool that can help them to better express their ideas, understand language and information, and navigate complex fields of knowledge.

DOI: 10.36078/1688391447

Thesaurus dictionaries serve several functions, including:

1. Vocabulary expansion: Thesaurus dictionaries provide synonyms and related words for a given term, which can help users to expand their vocabulary and use more varied and precise language.
2. Text analysis: Thesaurus dictionaries can be used to analyze and classify texts based on their content and language, allowing for more accurate indexing and retrieval of information.
3. Writing aid: Thesaurus dictionaries can be used as a writing aid, providing alternative words and phrases to help users express their ideas more effectively and creatively.
4. Concept mapping: Thesaurus dictionaries can be used to map the relationships between different concepts and terms, allowing users to better understand and navigate complex fields of knowledge.
5. Language learning: Thesaurus dictionaries can be used as a tool for language learning, helping users to improve their vocabulary, grammar, and writing skills.

Overall, the function of thesaurus dictionaries is to provide users with a means to expand their vocabulary, improve their writing and communication skills, and enhance their ability to analyze and understand language and information.

References

1. N. Z. Abdurakhmonova, A. S. Ismailov and D. Mengliev, "Developing NLP Tool for Linguistic Analysis of Turkic Languages," 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russian Federation, 2022, pp. 1790-1793, doi: 10.1109/SIBIRCON56155.2022.10017049.
2. W.E. Collinson, "Comparative Synonymics: Some Principles and Illustrations", Transactions of the Philological Society 38:1:54–77, November 1939, doi:10.1111/j.1467-968X.1939.tb00202.x
3. Werner Hüllen, A history of Roget's thesaurus : origins, development, and design, Oxford University Press 2004, ISBN 0199254729
4. Werner Hüllen, Networks and Knowledge in Roget's Thesaurus, Oxford, January 2009, doi:10.1093/acprof:oso/9780199553235.001.0001, ISBN 0199553238
5. Gertrude E. Noyes, "The Beginnings of the Study of Synonyms in England", Publications of the Modern Language Association of America (PMLA) 66:6:951–970 (December 1951) doi:10.2307/460151 JSTOR 460151

 DOI: 10.36078/1688391447

6. Eric Stanley, "Polysemy and Synonymy and how these Concepts were Understood from the Eighteenth Century onwards in Treatises, and Applied in Dictionaries of English" in Historical Dictionaries and Historical Dictionary Research, papers from the International Conference on Historical Lexicography and Lexicology, University of Leicester, 2002, Max Niemeyer Verlag 2004, ISBN 3484391235, p. 157–18
7. W.E. Collinson, "Comparative Synonymics: Some Principles and Illustrations", Transactions of the Philological Society 38:1:54–77, November 1939, doi:10.1111/j.1467-968X.1939.tb00202.x
8. Werner Hüllen, "Roget's Thesaurus, deconstructed" in Historical Dictionaries and Historical Dictionary Research, papers from the International Conference on Historical Lexicography and Lexicology, University of Leicester, 2002, Max Niemeyer Verlag 2004, ISBN 3484391235, p. 83–94

 DOI: 10.36078/1688391447

O‘ZBEK VA INGLIZ TILLARIDAGI KONSEPTUAL METAFORALARNI ELEKTRON LUG‘ATINI TUZISH VA DASTURINI YARATISH

Muhabbat Dushambayevna RAXIMBAYEVA
Fizika-matematika fakulteti
Axborot texnologiyasi kafedrasi f-m.f.n. dots.
Urganch davlat universiteti

Xosiyat Erkinboy qizi ABDULLAYEVA
Fizika-matematika fakulteti
Axborot texnologiyasi kafedrasi magistranti
Urganch davlat universiteti

Annotatsiya. O‘zbek va ingliz tillarida internet makonida qo‘llanilishini ifodalovchi metaforalarni taqqoslash shuni ko‘rsatadiki, insonning internet makonidagi turli muloqotlarining (chatlar va ularning turlari, konferensiyalar, ICQ, tematik forumlar, Internet kundaliklari, onlayn jamoalar, ijtimoiy tarmoqlar va boshqalar) metaforik belgilarida ko‘p narsa namoyon bo‘ldi. Metaforalarni solishtirishning amaliy natijasi nafaqat metaforalarning semantikasini aks ettiruvchi, balki sinonimlar va ekvivalentlarni qidiradigan, shuningdek, metaforalarni saralash imkonini beradigan elektron ikki tilli lug‘atni yaratish edi.

Kalit so‘zlar: metafora; lingvistika; chat; forum; ICQ.

Abstract. A comparison of metaphors in Uzbek and English for the use of the Internet shows that different types of human communication in the Internet (chats and their types, conferences, ICQ, thematic forums, Internet diaries, online communities, social networks, etc.) metaphorical signs. The practical result of comparing metaphors was to create an electronic bilingual dictionary that not only reflected the semantics of metaphors, but also searched for synonyms and equivalents, as well as sorting metaphors.

Keywords: metaphor; linguistics; chat; forum; ICQ.

Bugungi globallashuv asrida jamiyatning barcha sohalari kabi tilshunoslik ham zamonaviy kompyuter texnologiyalari va internet tarmoqlari bilan integrallashib bormoqda. Hozirgi kunda til materiallarini topish, ularni o‘rganish va tadqiqot olib borish jarayonini elektronlashtirish bugungi kun tilshunosligining dolzarb vazifasidir. Bu vazifani o‘tgan asrning 60-yillari boshlarda AQSHda asos solingan korpus tilshunosligi amalga oshirmoqda. Korpus — kompyuterning ma’lumotlar bazasida saqlanuvchi

 DOI: 10.36078/1688391447

og‘zaki va yozma matnlar majmui. Korpusda yig‘ilgan materiallarning aniq yozilgan vaqti, qaysi uslubga mansubligi, qaysi manbaga tegishliligi ham batafsil yoritilgan bo‘ladi.[1]

Korpusning bu imkoniyatlari undan keng hajmda va samarali foydalanish imkonini beradi. Bugungi kunda dunyoning ko‘plab tillari o‘z korpusiga ega. Mamlakatimizda ham *O‘zbek tili milliy korpusi* va uning asosida korpusning boshqa turlarini yaratish ishlari jadal olib borilmoqda.

Korpus materialining necha tilda berilishiga ko‘ra uning bir va ko‘p tilli turlari mavjud.

Bir tilli korpuslarda mavjud til materiallari faqat bir til doirasida bo‘ladi. Ko‘p tilli korpuslar, odatda, tarjimonlar tomonidan foydalaniladi. Ko‘p tilli korpusning yana bir ko‘rinishi original matn va tarjima matndan iborat bo‘ladi. Korpusning ushbu turi qiyosiy-chog‘ishtirma tadqiqot olib borishda, tarjima nazariyasi hamda kompyuter tarjimasini o‘rganishda juda muhim manba bo‘lib xizmat qiladi. Ko‘p tilli korpusning 2 turi mavjud:

- bir-birining tarjimasi bo‘lgan matnli korpus;
- bir mavzuga oid ikki tildagi matnli korpus.

Birinchi turdagi korpus parallel matnlar korpusidir. Parallel korpus bu – tarjima qilingan matnlar juftligidir. Ikkinchi xildagi korpus — tarjima korpusi (translation corpora) deb atalib, ayni bir fikrning turli tildagi ifodasini o‘rganish uchun muhimdir. Bir va ikki tilli korpuslarda semantik teglash muhim vazifa hisoblanadi.

Hozirgi kunda nafaqat tilshunoslikda, balki tarjima nazariyasi va amaliyotida ham korpus lingvistikasi, milliy korpus terminlari keng qo‘llanilib, bu sohada ingliz, ispan, fransuz, nemis, rus va boshqa ko‘plab tillarda axborot texnologiyalariga asoslangan samarali nazariy va amaliy ishlar, tadqiqotlar, loyihalar bajarildi va bajarilmoqda. O‘zbek tilida esa bu borada qilingan ishlar kamligi tufayli bu soha masalalarini yechish juda dolzarb hisoblanadi.[2]

Oldin metafora nima ekanini bilib olsak.

Metafora bu — bir tushunchaning ma’nosini boshqasiga ko‘chirgan, ikkalasi o‘rtasida o‘xshashlikni o‘rnatadigan adabiy figuradir. Ko‘chimlarning metafora, metonimiya, sinekdoxa kabi bir necha turlari bilan farqlanadi. Biz shu ko‘chimlarni ichidan metafora haqida gaplashamiz[3].

Metafora narsa va hodisalar orasidagi o‘xshashlik asosida ulardan birini ifodasi bo‘lgan so‘zni ikkinchisini ifodalash uchun qo‘llashdir, metaforada shaxsiy o‘xshashlik: belgi-xususiyat, harakat-holat nazarda tutiladi. Shu tufayli ot, fe’l va sifat turkumlarida

 DOI: 10.36078/1688391447

metafora yo‘li bilan ko‘chish hodisasi mavjud. Masalan, tish, yengil, pasaymoq so‘zlarining ma’nolarini kuzating. Tish:

1) odamning tishi (bosh ma’no);
2) arraning tishi (yasama ma’no).

Metafora qanchalik yangi, kutilmagan bo‘lsa, shunchalik ifodali bo‘ladi. Mohir so‘z ustalari o‘z asarlarida so‘zning metaforik ma’noda qo‘llash orqali ifodali, obrazli nutqning go‘zal namunalarini yaratadilar [4].

Lug‘atning umumiy xususiyatlari. “Inson internet makonida qo‘llaniladigan metaforalarning ikki tilli lug‘ati” – Internet makonida qo‘llaniladigan metaforalar solishtirilib ikki tildagi lug‘at yaratilmoqda. O‘zbek elektron ikki tilli metafora lug‘atida internet makonidan foydalanuvchi har bir insonning ichki dunyosi metaforalarini tahlil qilish va tizimlashtirish natijalarini aks ettiruvchi elektron resurs.

Elektron lug‘at tarkibi quyidagilardan tashkil topgan: foydalanuvchiga yo‘riqnoma, lug‘atning qisqacha izohi, manbalar, yangiliklar, lug‘atning MB, ya’ni elektron lug‘at, qidiruv maydonchasi, ro‘yxatda ko‘rsatilayotgan metaforalar soni.

Bu ma’lumotlar bazasi har bir foydalanuvchiga juda keng miqyosdagi imkoniyatlarni yaratib beradi. Undan metaforalar ro‘yxatini ko‘rish hamda ko‘chirish imkoniyati mavjud. O‘zbek tilidagi metaforalar ingliz tilida ma’nosini o‘zgartirmagan holda chiqaradi. Tanlangan mavzuni dolzarbligi shundaki, yuqorida aytib o‘tganimizdek, metaforalar ko‘chirmalardan tashkil topgan. Bundan ko‘rinadiki, o‘zbek tilidagi metaforani *online* tarjima qilmoqchi bo‘lsak, biz kutgan ma’noga ega bo‘lgan gapni chiqarib beradi. Bu muammolarni bartaraf qilish maqsadida ushbu elektron lug‘at ishlab chiqilmoqda[5].

Xulosa qilib aytganda, elektron ikki tilli ushbu lug‘at foydalanuvchilarning o‘zbek tilidagi metafora lug‘ati vositalarni to‘liq aks ettirishga imkon beradi. Bazada ro‘yxatdan birma-bir kerakli metaforalarni tizim yuqorisida joylashgan qiduruv bo‘limiga kerakli metaforani yozish orqali o‘zbek tilidagi metaforalarni ingliz tilida xatosiz, ya’ni biz kutgan ma’nodagi tarjimasini chiqaradi.

Foydalanilgan adabiyotlar

1. Husniddin Ruziyev. “O‘zbek-ingliz paremalar parallel korpuslari semantik teglarini yaratishda semantik izoh berish muammolari”. — Toshkent, 2022.
2. Nazira Sobirova. “Korpus lingvistikasi va parallel korpuslar tavsifi”. — Toshkent, 2022.
3. Yormat Tojiyev, Hakimjon Shamsiddinov, Ravshanxo‘ja Rasulov. “Hozirgi o‘zbek adabiy tili”. — Toshkent, 2005.
4. M.Hamrayev, D.Muhammadiyeva. “Ona tili”. — Toshkent, 2007.
5. https://ktonanovenkogo.ru/voprosy-i-otvety/metafora-chto-ehto-takoe-primery.html

 DOI: 10.36078/1688391447

ОПЫТ ЛИНГВИСТИЧЕСКИХ И ЛИНГВО-КУЛЬТУРОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ НА БАЗЕ КОРПУСЕ АДЫГЕЙСКОГО ЯЗЫКА

Раиса Батмирзовна УНАРОКОВА
доктор филологических наук,
профессор Адыгейского государственного
университета raya_unarokova@bk.ru
Зарема Арсеновна ЦЕЕВА
кандидат исторических наук,
доцент Адыгейского государственного
университета zarema.tseeva@yandex.ru

Изучение состояния языков, презентация их богатства посредством методов корпусной лингвистики становится одним из самых продуктивных исследовательских подходов. В российской филологической науке широко используется корпусная лингвистика. Помимо Национального корпуса русского языка (НКРЯ), языковые корпусы создаются в национальных регионах России.

Адыгейский корпус создан сравнительно недавно. Данный проект разрабатывался с 2014 по 2017 гг. в НИУ «Высшая школа экономики» группой лингвистов под руководством к.ф.н. Ю.А. Ландера [http://adyghe.web-corpora.net/]. Следует отметить, что перед научным коллективом стояли новые нестандартные задачи, поскольку адыгейский языкотносится к полисинтетическим языкам [1,18]. Само создание адыгейского корпуса явилось определенным вызовом для мировой корпусной лингвистики, так как адыгейский язык типологически существенно отличается от языков, для которых изначально создавались первые корпуса. Следовательно, подход к созданию данного корпуса был иным, отличным от выработанных ранее представлений, когда архитектура корпусов ориентировалась преимущественно на крупные европейские языки. Адыгейский корпус снабжен грамматической разметкой, которая является результатом автоматического анализа словоформ. Основу его составили художественные и публицистические тексты. Некоторые особенности адыгейского электронного корпуса описаны его создателями в одноименной статье [2, 198-206].

Появление Адыгейского корпуса важно не только с точки зрения исследовательского интереса. Актуальность создания корпуса связана также с современным состоянием адыгейского языка, входящего в абхазо-адыгскую семью, распространенную на Западном Кавказе и в среде кавказской диаспоры, расселенной в десятках государств. Во-первых, адыгейский язык является младописьменным

 DOI: 10.36078/1688391447

(после Октябрьской революции была создана письменность на основе арабского алфавита, который был заменён в 1927 году латиницей, а в 1938 году — кириллицей), и далеко не все вопросы, адыгского языкознания являются на сегодня разрешенными. Во-вторых, с середины XX века сфера применения адыгейского языка, чья литературная традиция насчитывает менее 100 лет, в силу ряда причин существенно сузилась и свелась преимущественно к уровню семейно-бытового общения. По классификации ЮНЕСКО адыгейский язык относится к числу уязвимых языков [3]. Еще более сложной является языковая ситуация в среде адыгской диаспоры. Значительное сужение функционирования адыгейского языка ведет к разрушению образного мышления народа и, как следствие – изменению порождаемой им языковой картины мира. Данный процесс провоцирует обильное появление в языке заимствованной лексики, калькирование иноязычных слов и понятий, переосмысление исконно адыгских выражений.

Поэтому в современных условиях использование информационных технологий (в том числе – электронных лингвистических корпусов) для изучения и сохранения адыгейского языка является одним из важнейших факторов. Создание Адыгейского корпуса позволяет аккумулировать лексический состав языка и основные грамматические конструкции в формате портала, находящегося в открытом доступе в сети Интернет.

Московская группа лингвистов под руководством к.ф.н. Ю.А. Ландера в период с 2017 по 2021 гг. проводила корпусные лингвистические исследования по адыгейскому языку [4,99-104].

В 2022 году был инициирован совместный проект НИУ ВШЭ и АГУ «Корпусные исследования адыгейского языка». Цельюданного проекта является изучение динамики развития и степени сохранности адыгейского языка, а также объективация его состава, структуры и образной системы.

Важным достижением в ходе реализации проекта стало значительное расширение Адыгейского корпуса (с 8,25 млн. словоупотреблений до 11,1 млн. словоупотреблений), благодаря чему в нём теперь намного лучше представлено разнообразие жанров и авторов. В то время как ранее основную часть корпуса составляли современные газетные тексты, в результате работы проекта в него введено большое число художественных произведений (добавлено около 1,5 млн. словоупотреблений) и научных текстов (более 420 тыс. словоупотреблений). Кроме того, создан уникальный фольклорный подкорпус на базе фольклорно-этнографического архива Центра адыговедения Адыгейского государственного университета (более 1600 текстов, более 600 тыс. словоупотреблений),

DOI: 10.36078/1688391447

позволяющий фильтровать тексты по фольклорным жанрам, исполнителю и его анкетным данным и т.д.

Расширенный фольклорными, научными и художественными текстами, лингвистический корпус в этой ситуации демонстрирует объективное состояние языка и помогает наметить пути его сохранения и перспективы развития.

В ходе реализации проекта были проведены лингвистические, лингвокультурологические и фольклористические исследования, осуществленные посредством использования многоуровневого, многофункционального адыгейского корпуса. Подготовлен к публикации ряд научных статей различной направленности.

Так, с применением корпусных методов проанализированы особенности репрезентации в фольклорных и публицистических текстах концепта «мужество» в адыгейском языке. При помощи корпуса удалось отобрать большой массив языковых данных и с их помощью изучить базовые лексические единицы рассматриваемого концепта как с формально-семантической, так и с количественно-статистической сторон. Это способствовало более детальному анализу лексико-семантических единиц и выявлению динамики функционирования концепта «мужество» в различных типах текстов. Установлено, что объективированное в фольклоре значение концепта «мужество» связано исключительно с проявлением этого качества в боевых условиях. В то же время исследование показало, что семантическое поле концепта «мужество» в современной адыгской языковой картине мира значительно расширилось, и ныне его содержание пересекается со значением таких понятий как «человечность», «твердость», «терпение». Контекст употребления данных понятий указывает на то, что в системе ценностей современного адыгского социума этическая категория «мужество» приобретает новые смыслы, отражающие не только воинскую, но и гражданскую доблесть.

Материалы фольклорного подкорпуса лингвистического корпуса адыгейского языка позволили ввести в научный оборот фольклористики новые термины и понятия: *тепчъэхэр* (ритуальные приговоры, ритуальные обращения к участникам обряда); *кlаджэхэр* (импровизационные возгласы и призывы). Через них манифестируются высокие идеалы традиционной культуры, в особенности – этикетные нормы взаимообхождения в обществе. Исследование показало, что высокая сохранность эстетики обрядового пространства (особенно свадебного) в условиях современности является не просто данью традиции, а еще одним каналом/механизмом трансляции культуры и связи поколений.

Исследование этнокультурных и когнитивных особенностей гендерного стереотипа фемининности в зависимости от культурных паттернов позволило

 DOI: 10.36078/1688391447

раскрыть социально-культурные и аксиологические особенности семиотизации представлений о женщине в адыгской лингвокультуре. При этом отбирался именно материал фольклорного подкорпуса, поскольку литературная форма языка нивелирует генетически первичные свойства самобытности, а паремиологический, фразеологический и фольклорный материал в большей степени сохраняет их [Бижева, Улаков2011].

При помощи Адыгейского корпуса осуществлен поиск ландшафтной лексики и архаичных топонимов, зафиксированных в фольклорных текстах, а также анализ фольклорного топонимикона. Выявлены словообразовательные особенности в географических названиях. Открылись новые возможности пополнения лексикографической базы словарей ландшафтной терминологии и исторической топонимии. Фольклорныйподкорпус позволяет оперативно пополнять лексикографическую базу ономастических изысканий. В то же время, апробирование метода на базе адыгейского лингвистического корпуса выявило сложность и специфику работы поисковика, обусловленную полисинтетизмом языка. В силу этого, показать частотность словоупотреблений нередко оказывается технически затруднительным и требует доработки в «ручном режиме». В то же время метод позволяет успешно выявлять коммуникативный и культурный контекст функционирования искомых единиц, установить значение малоупотребительных и исчезнувших из современного языка ландшафтных терминов и топонимов.

Кроме того, проведены лингвистические исследования, касающиеся принципов выделения местоименных серий на материале адыгейских демонстративов, а также анализ специфики формирования образа идеального героя в адыгском фольклоре. Применение корпусного инструментария позволило уточнить типы универсальных клишированных определений, которые используются в качестве основного художественно-стилевого средства для создания образа прославленного воина.

Таким образом, реализация проекта «Корпусные исследования адыгейского языка» показала, что введение в корпус фольклорных и диалектных коллекций дает возможность решать разнообразные задачи в области языкознания, фольклористики и лингвокультурологии. Проведение разноплановых исследований позволило протестировать потенциал Адыгейского корпуса для научных исследований и проектной работы (в том числе при преподавании языка, фольклора, этнической культуры). Работа с электронным корпусом адыгейского языка продемонстрировала результативность механизмов поиска необходимой информации. Комплексная автоматизация исследований и прикладных разработок при помощи корпусных

 DOI: 10.36078/1688391447

механизмов позволяют оптимизировать адыговедческую научную деятельность и выводить ее на новый исследовательский уровень.

В то же время, выявлена настоятельная необходимость дальнейшего пополнения Адыгейского корпуса новыми текстами различных стилей и жанров в целях расширения исследовательского поля.

Использованная литература

1. Ландер Ю.А, ЛетучийА. Б., СумбатоваН. Р., ТестелецЯ. Г. Основные сведения об адыгейском языке // Я. Г. Тестелец (ред.). Аспекты полисинтетизма: очерки по грамматике адыгейского языка. – М.: Изд-во Российского гос. гуман. ун-та, 2009. – С. 17—120.
2. Ландер Ю. А., Архангельский Т. А., Багирокова И. Г. Некоторые особенности адыгейского электронного корпуса // В кн.: Кавказская филология: история и перспективы. К 90-летию МухадинаАбубекировичаКумахова: сборник научных статей. – Нальчик: ИГИ КБНЦ РАН, 2019. – С. 198-206.
3. Atlas of the World's Languages in Danger [Электронныйресурс]. – 2010 – URL: https://web.archive.org/web/20220215051145/http://www.unesco.org/languages-atlas/index.php
4. Ландер Ю. А., Архангельский Т. А. Адыгейский корпус и орфографическое слово // Актуальные проблемы изучения кавказских языков. Материалы международной научно-практической конференции. – Махачкала: Издательство ДГУ, 2017. – С. 99-104.
5. Бижева З.Х., Улаков М.З. Лингвокультурологический аспект модернизации этносоциального пространства Северного Кавказа // Известия Кабардино-Балкарского научного центра РАН. – 36 (44). – Нальчик, 2011. – С. 259-263

DOI: 10.36078/1688391447

FILOLOGLARGA BINAR MUNOSABATLARNI O‘QITISH TO‘G‘RISIDA BA’ZI MULOHAZALAR

Nadira Raxmanovna UMAROVA
Zamonaviy axborot texnologiyalari kafedrasi katta o‘qituvchisi
O‘zbekiston davlat jahon tillari universiteti
nodiraumarova1960@mail.ru

Annotatsiya. Tezis filologlarga matematikani o‘qitish masalalariga bag‘ishlangan bo‘lib, unda matematikaning muhim tushunchalaridan biri bo‘lgan binar munosabatlarni o‘qitishda tilda mavjud bo‘lgan ayrim ma’lumotlardan foydalanish masalalari muhokama qilindi.

Kalit so‘zlar: binar munosabat; refleksivlik; simmetriklik; tranzitivlik; ekvivalentlik; tolerantlik; tartib munosabati.

Munosabat tushunchasi matematikaning filologiyada keng qo‘llanishga ega bo‘lgan tushunchalaridan biridir. Munosabat tushunchasini kiritish uchun matematikaning asosiy tushunchalaridan biri bo‘lgan to‘plamlarning dekart (to‘g‘ri) ko‘paytmasi tushunchasidan foydalanamiz.

A va B to‘plamlarning to‘g‘ri ko‘paytmasi – elementlari birinchi elementi A to‘plamga tegishli, ikkinchi elementi B to‘plamga tegishli bo‘lgan juftliklardan iborat bo‘lgan to‘plamdir.

A={a, o, i, e} va B= {b, d, k } to‘plamlar berilgan bo‘lsa, ularning to‘g‘ri ko‘paytmasi AxB={(a,b),(a,d),(a,k),(o,b),(o,d),(o,k),(i,b),(i,d),(i,k),(e,b), (e,d),(e,k)} to‘plamdan iborat bo‘ladi. Shuni aytib o‘tish kerakki, elementlarning joylashish tartibini o‘zgartirish mumkin bo‘lsa ham, tashkil etuvchi elementlarning o‘rnini almashtirish mumkin emas, ya’ni AxB ≠ BxA. Yuqoridagi misolimizda ko‘radigan bo‘lsak, AxB to‘plamning (a,b) elementi BxA to‘plamning (b,a) elementi bilan bir xil emas ((a,b) ≠ (b,a)).

AxA to‘g‘ri ko‘paytmaning ixtiyoriy qism to‘plami A to‘plamda binar munosabat bo‘ladi va R⊂AxA ($(a,b) \in R$) ko‘rinishida belgilanadi. $(a,b) \in R$ yozuv o‘rniga aRb yozuvdan ham foydalanish mumkin.

Gres P.V. “Matematika dlya gumanitariyev” o‘quv qo‘llanmasida (M.-Logos, 2007) binar munosabatlarga “Obyektlar juftliklari orasidagi munosabatlar binar (ikki o‘rinli) munosabatlar deyiladi” [1,52], deb ta’rif bergan.

Binar munosabatlarga sonlar orasidagi tenglik (=), katta yoki kichik bo‘lish (>,<) munosabati, sonlarning EKUK va EKUBga ega bo‘lish munosabati, o‘zaro tub bo‘lishi munosabati, to‘g‘ri chiziqlar orasida parallel, perpendikulyar bo‘lish, geometrik figuralar

 DOI: 10.36078/1688391447

orasida teng bo‘lish, o‘xshash bo‘lish, gomeomorf bo‘lish, insonlar orasida sinfdosh bo‘lish, tengdosh bo‘lish, kasbdosh bo‘lish, tanish bo‘lish, do‘st bo‘lish, dushman bo‘lish, davlatlar orasida hamkor bo‘lish, chegaradosh bo‘lish, bayroqlarida bir xil ranglar bo‘lishi kabi munosabatlarni misol qilib keltirish mumkin.

Filologiyada ham harflar, tovushlar, so‘zlar, gaplar orasida uchraydigan binar munosabatlarga doir misollarni ko‘p keltirish mumkin. Masalan, alifboda ikkita harfning ketma-ket kelishi, ikkita harfning bo‘g‘in tashkil qilishi, ikkita so‘zning sinonom, omonim, paronim, o‘zakdosh bo‘lishi, ikkita so‘zning gap yoki so‘z birikmasi tashkil qilishi va h.

$(a,b) \in R$ yozuv o‘rniga aRb yozuvdan foydalanish R binar munosabatning xossalarini ifodalashda qulayliklar tug‘diradi:

1°. $\forall a \in A\,(aRa)$ refleksivlik, har bir element o‘z –o‘zi bilan shu munosabatda. Masalan, tenglik, tengdoshlik, o‘z-o‘ziga xizmat qilish, $\leq$, $\geq$.

2°. $\forall a \in A \neg(aRa)$ antirefleksivlik, masalan, $<,>$, yoshi katta bo‘lish, bo‘yi baland bo‘lish.

3°. $\forall a,b \in A\,(aRb \Rightarrow bRa)$ simmetriklik (tenglik, tengdoshlik, o‘xshashlik, umumiy bo‘luvchiga ega bo‘lish, sochlarining ranglari bir xil bo‘lishi, sinfdosh, kursdosh bo‘lish)

4°. $\forall a,b \in A\,(aRb \,\&\, bRa \Rightarrow (a=b))$ antisimmetriklik (qat’iymas tengsizlik $\leq$, $\geq$., qism to‘plam bo‘lish)

5°. $\forall a,b,c \in A\,(aRb \,\&\, bRc \Rightarrow aRc)$ tranzitivlik (tenglik, tengdoshlik, o‘xshashlik, umumiy bo‘luvchiga ega bo‘lish, sochlarining ranglari bir xil bo‘lishi, sinfdosh, kursdosh bo‘lish, bir mahallada yashash)

Refleksivlik, simmetriklik va tranzitivlik xossalariga ega bo‘lgan munosabatlar ekvivalentlik munosabati deyiladi. Tenglik, tengdoshlik, o‘xshashlik, parallellik, bitta mahallada yashash, soch ranglarining bir xil bo‘lishi, sinfdosh bo‘lish, kursdosh bo‘lish, kasbdosh bo‘lish munosabatlari ekvivalentlik munosabatlariga misol bo‘la oladi.

Ekvivalentlik munosabati to‘plamni ekvivalent elementlar sinflariga ajratadi. Aksinchasi ham o‘rinli, agar to‘plam sinflarga ajratilgan bo‘lsa unda ekvivalentlik munosabati berilgan bo‘ladi. (Sinflarga ajratish — berilgan to‘plamni o‘zaro kesishmaydigan, yig‘indisi shu to‘plamga teng bo‘lgan qism to‘plamlarga ajratish).

Masalan, haqiqiy sonlar to‘plamini ikki sinfga ajratish mumkin: algebraik va transsendent sonlar to‘plami.

Bitta mahallada yashash munosabati ekvivalent binar munosabat bo‘ladi va mahalla odamlari to‘plamini mahallada qo‘shni bo‘lgan o‘zaro kesishmaydigan odamlar qism to‘plamlariga ajratadi (bir uyda yashamaydigan qo‘shnilar).

 DOI: 10.36078/1688391447

“Bitta dahada yashash” munosabati shahar odamlarini o‘zaro kesishmaydigan dahalarga ajratadi.

Filologiyada mavjud bo‘lgan ayrim binar munosabatlarning qanday xossalarga ega bo‘lishini ko‘rib chiqaylik.

Rus tilida mavjud bo‘lgan “otlar bir jinsga tegishli” munosabati rus tilida mavjud bo‘lgan otlar to‘plamini uchta ekvivalentlik sinflariga ajratadi.

Gapning ikkinchi darajali bo‘laklari — aniqlovchi, to‘ldiruvchi, hol — ekvivalentlik sinflaridir (grammatk ma’nosiga ko‘ra bo‘linadi).

Harakatni bildiruvchi so‘zlar turkumi — fe’llar ham “bir xil shaklga ega” (aniq va noaniq (infinitiv) shakli), “bir xil zamonga ega” (o‘tgan, hozirgi va kelasi zamon) munosabatlariga nisbatan ekvivalentlik munosabatlari bo‘ladi.

Ekvivalentlik sinflariga misol sifatida yana quyidagi tushunchalarni olish mumkin:

- gapning bosh bo‘laklari — ega va kesim;
- kesimlarning sodda kesim, qo‘shma kesimlarga ajratilishi;
- gaplarning darak, so‘roq va undov gaplarga ajratilishi;
- gaplarning sodda va murakkab gaplarga ajratilishi;
- sodda gaplarning yig‘iq va yoyiq bo‘lishi;
- sodda gaplarning shaxsi ma’lum va shaxsi noma’lum gaplarga ajratilishi;
- bog‘lovchilar yordamida va intonatsiya yordamida tuzilgan murakkab gaplar;
- tovushlarning unli va undosh tovushlarga ajratilishi;
- eskirgan (arxaizm) va yangi kiritilgan so‘zlar.

Yana bir munosabat haqida to‘xtab o‘taylik. Refleksivlik va simmetriklik xossalariga ega bo‘lgan munosabatlar tolerantlik munosabati deyiladi. Insonlar orasida tanish bo‘lish, qarindosh bo‘lish, so‘zlar orasida ko‘pi bilan bitta harf bilan farqlanish.

Uchta harfli so‘zlar orasida “ikkita bir xil harfga ega bo‘lish” munosabati o‘zbek tilidagi “borni yo‘q qilish” yoki “yo‘qni bor qilish” lingvistik masalasini yechishda namoyon bo‘ladi: $bor \to yor \to yoq \to yo‘q$ va $yo‘q \to to‘q \to toq \to boq \to bor$.

Refleksiv, antisimmetrik va tranzitiv bo‘lgan munosabat tartib munosabati deyiladi. (Katta yoki teng ($\le$), kichik yoki teng ($\ge$), yoshi katta yoki teng bo‘lishi).

Yana bir misol. Tildagi so‘zlar uchun “A va B so‘zlarning birinchi m ta harflari bir xil bo‘lib, alifboda A so‘zining $(m+1)$-inchi harfi B so‘zining $(m+1)$ -inchi harfidan oldin turadi “munosabati, “Birinchi harfi bir xil bo‘lish” (ekvivalentlik) munosabatlaridan til lug‘atlarini tuzishda foydalaniladi.

Filologiyadagi ba’zi bir munosabatlarning qaysi xossaga ega ega ekanligini ko‘rib chiqaylik.

DOI: 10.36078/1688391447

So‘zlar orasida “sinonim so‘zlar bo‘lish” munosabati ekvivalent binar munosabatdir, chunki refleksivlik, simmetriklik va tranzitivlik xossalariga ega. Sinonim (ma’no jihatidan yaqin yoki bir xil bo‘lgan, bir xil tushunchani bildiruvchi so‘zlar).

1. Sinonim bo‘lish munosabati refleksiv, chunki har bir so‘z o‘zida ifodalangan ma’noni ifodalaydi.

2. Sinonim bo‘lish munosabati simmetrik. Agar a so‘zi b so‘ziga sinonim bo‘lsa, u holda b so‘zi ham a so‘ziga sinonim bo‘ladi.

3. Agar a so‘zi b so‘ziga sinonim bo‘lsa va b so‘zi c so‘ziga sinonim bo‘lsa, u holda a so‘zi c so‘ziga sinonim bo‘ladi. Haqiqatan ham shunday.

Antonim bo‘lish munosabati qanday xossalarga ega bo‘lishini ko‘raylik.

1. Refleksivlik xossasi bajarilmaydi, chunki hech qanday so‘z o‘z-o‘ziga antonim emas.

2. Agar a so‘zi b so‘ziga antonim bo‘lsa, b so‘zi ham a so‘ziga antonim bo‘ladi. Bu to‘g‘ri (katta-kichik).

3. Agar a so‘zi b so‘ziga antonim bo‘lsa va b so‘zi c so‘ziga antonim bo‘lsa, u holda a so‘zi c so‘ziga antonim bo‘ladimi? Yo‘q, albatta.

Ko‘rinib turibdiki antonim bo‘lishlik faqat simmetriklik xossasiga ega.

So‘zlar orasida birinchi harfi bir xil bo‘lish munosabati ham ekvivalentlik munosabati bo‘ladi.

Xuddi shunday so‘zlar orasida omonim so‘zlar bo‘lish, paronim so‘zlar bo‘lish, o‘zakdosh bo‘lish, so‘zlarning harflari soni teng bo‘lishi, qofiyadosh bo‘lishi va yana boshqa bir qator munosabatlarni ko‘rib chiqish mumkin.

Filologiya ta’lim yo‘nalishi talabalariga binar munosabatlar mavzusini, matematikani o‘qitishda filologiyada mavjud bo‘lgan tushunchalardan foydalanish talabalarning fanni yaxshi o‘zlashtirishlariga asos bo‘ladi, deb hisoblaymiz.

Foydalanilgan adabiyotlar

1. Грес П.В. Математика для гуманитариев: Учебное пособие. – М.: Логос, 2007.
2. А.А.Азизов, И.М.Дмитриева, М.А.Рогожина “Русский язык.11” Ташкент : Укитувчи,1997
3. Умарова Н.Р. Эквивалентные бинарные отношения в филологии. Гармонично развитое поколение- условие стабильного развития Республики Узбекистан Сборник научно- методических статей № 3 Ташкент-2017
4. Н.Умарова, Ғ.Ахмадалиев. Бинар муносабатлар ва хоссалари Узбекский научно-исследовательский институт педогогических наук имени Т.Н.Кары-Ниязи. Гармонично развитое поколение - условие стабильного развития

DOI: 10.36078/1688391447

республики Узбекистан. Сборник научно-методических статей. Ташкент. 2015 г. апрел. 388- ст.
5. Н.Умарова, Ғ.Ахмадалиев. Применение теории отношений в филологии. Узбекский научно-исследовательский институт педогогических наук имени Т.Н.Кары-Ниязи .Гармонично развитое поколение - условие стабильного развития республики Узбекистан. Сборник научно-методических статей. Ташкент. 2015 г. март. 314- ст.

 DOI: 10.36078/1688391447

STEAM YONDASHUV ASOSIDA INTEGRATSIYALASHGAN DARSLARDA GEOGEBRA DASTURIDAN FOYDALANISH

Nigora Alisherovna UMAROVA
Ijtimoiy gumanitar fanlar kafedrasi katta o‘qituvchisi
Angren universiteti
muhammadnigora@mail.ru

Annotatsiya. Tezis integratsiyalashgan dars, uni o‘tishda STEAM yondashuv va integratsiyalashgan darslarni o‘tishda GeoGebra dasturidan foydalanish masalalariga bag‘ishlangan.

Kalit so‘zlar: integratsiya; fanlararo aloqa; STEAM yondashuv; geogebra dasturi.

Bu tezis “STEAM yondashuv asosida integratsiyalashgan darslarda GEOGEBRA dasturidan foydalanish” mavzusidagi ilmiy ishimiz doirasidagi ilk izlanishimizdir. Mavzu nomidagi asosiy tushunchalar haqida ma’lumotlarni yig‘ish bilan shug‘ullanib ko‘raylik.

Mavzuda ishtirok etgan birinchi tushuncha — integratsiya tushunchasidir. Uzluksiz ta’lim jarayonini yuqori darajaga ko‘tarishda integratsiya, ya’ni fanlararo aloqadorlikdan foydalanish muhim ahamiyat kasb etadi. Integratsiya termini lotincha integratio — tiklash, to‘ldirish, integer — butun so‘zidan olingan bo‘lib, fanlarning yaqinlashishi va o‘zaro aloqa jarayoni bo‘lib, differensiatsiya bilan birga kechadi [1]. Integratsiya atamasi qo‘shilish, birlashish ma’nolarini bildiradi. Integratsiyalash turli manbalarda mavjud bo‘lgan materiallarni ma’lum bir maqsad asosida birlashtirib taqdim etishdir.

Ma’lumotlarning haddan tashqari ko‘pligi, turli fanlarda takrorlanishi o‘quvchilar tomonidan yaxshi o‘zlashtirilmasligiga sabab bo‘lib, oqibatda o‘quvchining bilim olishga bo‘lgan qiziqishi kamayadi.

Bugungi kunda o‘quvchilar olishi kerak bo‘lgan bilim, ko‘nikma va malakalar hajmi ortib, uni egallashga kerak bo‘lgan vaqt miqdori kamayib ketyapti. Shu ajratilgan vaqt davrida qo‘yilgan vazifani amalga oshirish uchun turli usullardan foydalanish zarur bo‘ladi, bu usullar ichida eng maqbuli fanlararo integratsiyadan foydalanishdir.

Integratsiya tushunchasiga olimlar tomonidan turlicha ta’riflar berilgan. Masalan, N.S.Svetlovskaya integratsiyani “ilgari bir -biridan farq qiladigan bir nechta turli birliklar (o‘quv fanlari, faoliyat turlari va boshqalar) da aniqlangan bir xil turdagi elementlar va qismlar asosida yangi yaxlitlikni yaratish, so‘ngra ushbu elementlar va qismlarni ilgari mavjud bo‘lmagan maxsus sifatli butunlikka moslashtirish” deb talqin qiladi, uning

 DOI: 10.36078/1688391447

fikricha, integratsiyaning muhim sharti materialni bir qator fanlar va metodikada yagona maqsad va funksiyasiga tabiiy bo‘ysunishi asosida qurishdir [3].

“Integratsiya” tushunchasini L.N.Baxareva “fanlarning yaqinlashishi va aloqa jarayoni” deb talqin qiladi va “... fanlararo aloqalarni ta’limning yangi sifat darajasini o‘zida mujassam etgan yangi butun “bilim monolitini (yaxlitligini)” yaratishga hissa qo‘shuvchi yuqori shakli ...” deb ifodalaydi [4, 48].

U integratsiyani ta’limning fanlar sistemasidan iborat ekanligini inkor qilmaydi, ta’limni mukammallashtirish, kamchiliklarni bartaraf etish va fanlararo aloqadorlik, o‘zaro bog‘liqliklarni chuqurlashtirish yo‘li deb hisoblaydi. Muammoga bunday yondashuv integratsiya va differensiatsiya orasidagi munosabatlarni tushunishga asoslangan.

Integratsiya masalalari bilan shug‘ullanuvchi yana bir tadqiqotchilar I. D. Zverev va V. N. Maksimovalar integratsiyani uzluksiz bog‘langan, birlashgan, yaxlitlikni yaratish jarayoni va natijasi deb hisoblaydilar. Ta’limda u fanlararo ta’lim muammolarini ochib berishda turli o‘quv predmetlarining bir sintezlangan kurs (mavzu, bo‘lim, dastur) elementlarini birlashtirish, turli fanlarning ilmiy tushunchalari va usullarini umumiy ilmiy tushunchalar va bilish usullari bilan uyg‘unlashtirib, fan asoslarini birlashtirish va umumlashtirish orqali amalga oshiriladi [5, 47].

Yana bir olim V. S. Kukushkinning fikricha, “integratsiya bu bir yoki bir nechta turli o‘quv fanlari bo‘yicha bir-biridan farq qiluvchi bilimlarning yaxlitlik xususiyatiga ega bo‘lgan tizimga birlashishi jarayonidir”. Umumlashtirilishi mumkin bo‘lgan bilimlarni yaxlitlash, yagona butunga birlashtirish, o‘quvchilarga hozirgi davrda juda muhim bo‘lgan asosiysini ajrata olishga o‘rgatishni ko‘rsatib, tahlil qilish va umumlashtirishni o‘rganishga yordam berish [6].

Y.M.Kolyaginning talqinida ta’lim tizimiga nisbatan “integratsiya” tushunchasi ikki xil ma’noga ega. Integratsiya ta’lim maqsadi — “o‘quvchining atrof muhit to‘g‘risidagi yaxlit tasavvurini shakllantirish” va o‘quv vositasi — “fan bilimlarini yaqinlashtirishning umumiy platformasini toppish” deb qaralishi mumkin. Ta’lim maqsadi sifatida u o‘quvchilarga dunyoni elementlari o‘zaro bog‘liq bo‘lgan yaxlit bir butunlik sifatida tasavvur qilishga o‘rgatadigan bilimlarni beradi. Va integratsiya ta’lim vositasi sifatida bilimlarni rivojlantirish, kengaytirish va yangilashga qaratilgan. Shu bilan birga, integratsiya ta’limni an’anaviy o‘quv fanlarini o‘qitish bilan almashtirmay, olingan bilimlarni yagona tizimga birlashtirishi kerak [7].

Ta’limda integratsiya turli yo‘nalish va darajalarda amalga oshiriladi:

1. Intra-predmet — individual o‘quv fanlari doirasida (tushunchalar, bilim, ko‘nikmalarni birlashtirish);

 DOI: 10.36078/1688391447

2. Ikki yoki undan ortiq fanlar orasida — fanlararo (faktlar, tushunchalar, tamoyillarni tahlil qilish);
3. O‘ziga xos fanlararo aloqadorlik — trans-predmet (ma’lum bir fanning boshqa fanlar bilan bog‘lanishi.

Fanlararo integratsiya bir o‘quv fanini o‘rganishda boshqa bir fanning qonun, nazariya, usullaridan foydalanish demakdir. Shu tartibda amalga oshirilayotgan tizimlashtirish o‘quvchilar ongida fanlarga boshqa ko‘z bilan qarash tushunchasini shakllantiradi. Matematikaning ayrim tushunchalari juda ko‘p fanlarda ishtirok etadi. Masalan, son, proporsiya, to‘plam, munosabat, graf, tenglama, funksiya, hosila tushunchalari. Matematika va fizika, matematika va biologiya, matematika va kimyo, matematika va geografiya, matematika va chizmachilik, matematika va jismoniy tarbiya, matematika va ona tili fanlari orasidagi fanlararo aloqadan foydalanish ta’limni, matematikani o‘qitishni yuksak darajaga ko‘taradi. Matematik qonunlarning fizika va kimyo fanlarida qo‘llanilishiga juda ko‘p misollar keltirish mumkin. Matematikaning ko‘p tushunchalarini geografik tushunchalar bilan uyg‘unlashtirish mumkin. Masalan, globus, qutb, kenglik, uzunlik, ekvator, shaharlar o‘rni, karta, azimut, koordinata chiziqlari, tekisligi, nuqtaning koordinatasi, vertikal burchak, aylana, shar tushunchasi va h. Fanlararo integratsiyaga doir ma’lumotlarni matematika va gumanitar fanlar orasida, ayniqsa, o‘zbek tili fani bilan juda ko‘p ko‘rinishda berish mumkin. Matematik mantiq bo‘limining mulohaza (rost yoki yolg‘on qiymat qabul qiluvchi darak gaplar), ular ustida mantiqiy amallar mantiqiy qo‘shish, mantiqiy ko‘paytirish, (yoki, va bog‘lovchilari), mantiqiy xulosa chiqarish amali (“Agar… bo‘lsa, u holda….bo‘ladi” bog‘lovchisi), murakkab formula (qo‘shma gap) til va matematika fani orasidagi tushunchalarning aloqasini ko‘rsatishga misol bo‘ladi.

Ilmiy ish mavzusidagi ikkinchi bir tushuncha — STEAM yondashuv tushunchasi ustida ham to‘xtalib o‘taylik. STEAM: S — science, T — technology, E — engineering,
A — art va M — math, ya’ni tabiiy fanlar, texnologiya, muhandislik, san’at, matematika. Nomlanishidan ko‘rinib turibdiki, STEAM texnologiyasi ta’limni hayot bilan bog‘lovchi texnologiya bo‘lib, uning asosiy g‘oyasi nafaqat nazariy bilimlar, amaliy bilimlarni ham muhimligini ko‘rsatishdir.

Ta’limdagi bu yondashuv shiori “Mind and hand – “Aql va qo‘l” bo‘lgan Amerikaning Massachusets Texnologiyalar universitetida ishlab chiqilgan bo‘lib, unda STEAM kurslari yaratilib, dasturi o‘quvchilarning ilmiy-texnika yo‘nalishlarida kompetensiyalarini rivojlantirishga qaratilgan bo‘lib, o‘quvchilarda muammolarni keng qamrovli tushunish, ijodiy fikrlash, muhandislik yondashuvi, tanqidiy fikrlash, ilmiy metoforalarni tushunish

 DOI: 10.36078/1688391447

va qo‘llash, dizayn asoslarini tushunish kabi ko‘nikmalarni rivojlantirish vazifasini amalga oshiradi.

STEAM – ta’lim texnologiyasi:

- loyihalash metodiga tayanib, uning asosini bilish va badiiy izlanish tashkil etadi;
- bolaning rivojlanishini bevosita tashqi olam bilan bog‘laydi;
- dunyoni tizimli o‘rganish imkonini beradi;
- atrofda ro‘y berayotgan jarayonlar haqida mantiqiy fikrlashga o‘rgatadi;
- voqeliklar orasidagi aloqani anglab olishga o‘rgatadi;
- atrof-muhitdan o‘zi uchun noma’lum bo‘lgan, noodatiy va qiziqarli yangiliklarni ochish imkonini beradi;
- ta’lim mavzu doirasida integratsiyalab olib boriladi;
- fanlararo aloqa va loyihalash metodlari uyg‘unlashtirilgan holda amalga oshirilib, tabiiy fanlar texnologiya, muhandislik ijodiyot va matematika bilan integratsiyalashadi;
- amaliy mashg‘ulotlar orqali ilmiy-texnik bilimlar real hayotda foydalanishi namoyish qilinadi;
- darslarda o‘quvchilar modellarni ishlab chiqadi; ularni sinovdan o‘tkazadi, natijasi kutilganday bo‘lmasa, uning sabablarini izlaydi, o‘ylaydi va kamchilikni bartaraf etish chorasini ko‘radi, har bir sinovdan keyin modelni yanada takomillashtiradi, muammolarni o‘z kuchlari bilan yengib maqsadlariga erishadilar. Yutuqdan keyin o‘z kuchlariga ishonch hissi paydo bo‘ladi;
- hayotda uchraydigan qiyinchiliklarni yengish uchun zarur bo‘lgan tanqidiy tafakkur va muammolarni hal etish ko‘nikmalarini rivojlantiradi;
- guruhda ishlash mobaynida o‘quvchilarda o‘z fikrini bayon qilish, muhokama va munozara qilish imkoni paydo bo‘ladi [8].

Mavzudagi uchinchi katta tushuncha GeoGebra dasturi, uning yaratuvchisi Markus Xoenvarter. “GeoGebra bu — dinamik geometriya dasturi, ya’ni u yordamida biz konstruksiyalarni nuqtalardan, segmentlardan, chiziqlardan va hokazolardan yaratishimiz mumkin, ularni turli xil geometrik formulalarni tiklashga undash va shu bilan kerakli tushunchalarni yaxshiroq o‘rganish yoki yaxshiroq o‘rgatishda ishlatish mumkin” [10].

GeoGebra (geometriya va algebra portmantosi) — boshlang‘ich maktabdan universitet darajasiga qadar matematika fanini o‘rganish uchun mo‘ljallangan interfaol geometriya, algebra, statistika va hisob-kitob ilovasi. GeoGebra ish stollari

 DOI: 10.36078/1688391447

(Windows, macOS va Linux), planshetlar (Android, iPad va Windows) va veb uchun ilovalar bilan bir nechta platformalarda mavjud [9].

GeoGebra dasturiy ta’minot to‘plamida tuzilmalar nuqtalar, vektorlar, segmentlar, chiziqlar, ko‘pburchaklar, kesimlar, tengsizliklar, yashirin ko‘phadlar va funksiyalar yordamida amalga oshiriladi. Elementlarni sichqoncha va sensorli boshqaruv elementlari yoki kiritish paneli orqali kiritish va o‘zgartirish mumkin. GeoGebra raqamlar, vektorlar va nuqtalar uchun o‘zgaruvchilarni saqlashi, funksiyalarning hosilalari va integrallarini hisoblashi mumkin va Root yoki Extremum kabi buyruqlarning to‘liq to‘plamiga ega. O‘qituvchilar va talabalar GeoGebradan geometrik taxminlarni shakllantirish va isbotlashda yordam sifatida foydalanishlari mumkin [9].

Matematikaning tilda ishlatiladigan ayrim tushunchalarini Geogebra dasturi yordamida quruvchi instrumentlaridan ayrimlarini ko‘rsatib o‘taylik.

Kesma (gap bo‘laklarini belgilashda ishlatiladi):

1. «Прямая» instrumentida oq uchburchakga bosamiz.
2. Ro‘yxatdan «Отрезок» ni tanlaymiz.
3. Kesmani uchi – 2 ta nuqtani qo‘yamiz.

Nur (yo‘nalishga ega bo‘lgan chiziq, matematikada graf qirrasi, binar munosabatlarni ifodalasa, tilda gap bo‘laklari orasidagi bog‘lanishni ifaddalaydi):

1. «Прямая» instrumentida oq uchburchakga bosamiz.
2. Ro‘yxatdan «Луч» ni tanlaymiz.
3. Maydonda 2 ta nuqtani tanlaymiz – birinchisi – nur boshi, ikkinchisi – nur o‘tadigan nuqta [11].

Foydalanilgan adabiyotlar

1.Х.Н.Ҳакимов. Ўқувчиларни баркамол шахс қилиб тарбиялашда интеграцион ѐндашув // Педагогик маҳорат.- Бухоро, 2017. -№4, 238-240

2. Х.Н.Ҳакимов. Ўқувчи шахсини жисмоний ва маънавий-ахлоқий шакллантиришда интеграциялашган дарсларнинг хусусиятлари// Осиё мамлакатлари тамаддуни ва ипак йўли. Халқаро илмий анжуман материаллари. Самарқанд, 2019.270-272 бетлар

3. Светловская И.С. Об интеграции как методическом явлении и ее возможностях в начальном обучении // Начальная школа. 1990. №5. С. 57-60.

4. Бахарева Л.Н. Интеграция учебных занятий в начальной школе на краеведческой основе // Начальная школа. 1991. №8. С. 48-51.

5. Зверев И.Д., Максимова В.Н. Межпредметные связи в современной Педагогика. -1981.- 195с.

 DOI: 10.36078/1688391447

6. Кукушин В.С. Современные педагогические технологии // В. С. Кукушин. Начальная школа. Изд. 2-е. Ростов н/Д, 2004. – 384 с.

7. Колягин Ю. М., Алексеенко О.Л. Интеграция школьного обучения // Начальная школа, 2001. № 9. 28 – 31 с.

8. G.Ergasheva, D.Boltayeva. STEAM yondashuv orqali o‘quvchilarda tayanch kompetensiyalarni shakllantirish. Fan, ta’lim va amaliyot integratsiyasi. “Bilig – ilmiy faoliyat” FTAI Respublika ilmiy-amaliy konferensiya nashri | http://bilig.academiascience.org ISSN: 2181-1776.

9. GeoGebra Materials: http://www.geogebra.org/materials

10. https://blog.desdelinux.net/uz/geogebra-matematikasi

11.Planimetriyani geogebra dasturi asosida o’rganish. N.V. Juraeva,CHDPI “Matematika” kafedrasi dotsenti dosent F.M. Xoldorova

“Aniq va tabiiy fanlarni o’qitish metodikasi” (matematika) Academic Research in Educational Sciences. Zamonaviy ta'limda matematika, fizika va raqamli texnologiyalarning dolzarb muammolari va yutuqlari, TVCHDPI.|CSPI CONFERENCE 3 | 2021.

 DOI: 10.36078/1688391447

СТИЛОМЕТРИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ ПЕРЕВОДА НА РУССКИЙ ЯЗЫК ПРОИЗВЕДЕНИЙ ДЖ. К. РОУЛИНГ

Вероника Дмитриевна ВЕДЕРНИКОВА
магистрант, 1 курс
Институт филологии, журналистики и межкультурной коммуникации
Южный федеральный университет
vedernikova@sfedu.ru

Елена Михайловна СЕВЕРИНА
Научный руководитель
доктор философских наук, профессор
emkovalenko@sfedu.ru

Аннотация. Работа посвящена изучению специфики выявления идиостиля переводчиков цифровыми методами. Проведен стилометрический анализ текстов переводов серии книг «Гарри Поттер» Дж. К. Роулинг, позволивший выявить специфику стиля переводчика.

Ключевые слова: стилометрия; идиостиль переводчика; частотный анализ; корреляционный анализ; кластерный анализ.

Вопрос о создании качественного перевода актуализирован спорами о роли идиостиля переводчика в процессе перевода. Известный переводчик с французского языка, доктор филологических наук С. Н. Зенкин отметил, что «хороший перевод имеет тенденцию исчезать, стушевываться перед оригиналом; хороший перевод – это незаметный перевод…» [3]. Безусловно, основная задача перевода - передача смыслов автора, однако, переводчику дозволяется в рамках контекста и литературного жанра выбирать языковые формы для их выражения. В связи с этим возникает необходимость в определении и изучении тех языковых форм, которые определяют специфический стиль переводчика и могут влиять на передачу авторских смыслов.

Для исследования были взяты тексты перевода на русский язык серии книг «Гарри Поттер» Дж. К. Роулинг, опубликованные издательствами «Махаон» и «Росмэн». Следует отметить, что переводы издательства «Махаон» осуществлены одним переводчиком -М. Спивак [1], в то время как переводы издательства «Росмэн» выполнены группой переводчиков под руководством главного редактора М. Д. Литвиновой [2].

DOI: 10.36078/1688391447

Материал исследования представлен в формате plain text (txt), в кодировке UTF-8, и составил 904 822 токена (перевод издательства «Махаон») и 928 837 токенов (перевод издательства «Росмэн»), что позволяет рассматривать переводы как относительно равнозначные с точки зрения лексического объема. Для определенных задач, в частности для составления частотного словаря, тексты были предобработаны: токенизированы и лемматизированны с удалением пунктуации, стоп-слов и приведением к нижнему регистру.

Для определения идиостиля переводчика были выбраны такие стилометрические методы, как частотный, корреляционный и кластерный анализ. Стилометрия позволяет провести статистический анализ отклонений между литературными стилями разных авторов и жанров. Стилометрические методы во всем их разнообразии имеют две общие черты: текстовые элементы должны быть каким-то образом преобразованы в числа, а числа уже исследованы статистическими методами [7]. И одним из наиболее простых и демонстративных стилометрических методов выступает частотный анализ.

Частотный анализ является отличным «первым знакомством» с исследуемым текстом, поскольку наглядно представляет самые частотные формы. В ходе нашего исследования частотный анализ проводился на предобработанных текстах с последующим составлением на базе частотных словарей терм-документной матрицы вхождений элементов в определенный текст.

Результаты частотного анализа показали, что при наличии единой языковой формы в оригинальных текстах имена героев и названия атрибутов фэнтэзийного пространства различаются в переводных текстах: так в переводе издательства «Махаон» переводчик передает «*Dumbledore*», «*Voldemort*» и «*Hagrid*» как «*Думбльдор*», «*Вольдеморт*» и «*Огрид*» в то время, как переводчики издательства «Росмэн» указывают «*Дамблдор*», «*Волан-де-Морт*» и «*Хагрид*». Кроме того, различается перевод слова «*wizard*»: в переводе издательства «Махаон» – это «*колдун*», а в переводе «Росмэн» – «*волшебник*». Данные лексические формы, являясь синонимами, передают разные конотативные значения. Согласно Русскому семантическому словарю, «волшебник» может быть как «добрым» (положительная коннотация), так и «злым» (отрицательная коннотация) в то время, как «колдун» всегда «злой» (отрицательная коннотация) [5]. Исследование коллокаций в НКРЯ показало, что слово «*волшебник*» употребляется с коллокатами «*злой*» и «*добрый*» почти в одинаковых пропорциях, а слово «*колдун*» в основном используется в коллокации «*злой колдун*» [4].

 DOI: 10.36078/1688391447

Корреляционный анализ позволил проследить степень лексической связанности каждого текста друг с другом. Для исследования использовались терм-документные матрицы, составленные на основе частотных словарей, анализ которых показал достаточно высокий коэффициент корреляции практически для всех текстов в переводе двух издательств. Однако есть небольшие различия между оригинальными и переводными текстами. Например, в переводе издательства «Росмэн» коэффициент корреляции между 4 и 5 текстами серии снижается с 1.0 до 0.9. Кроме того, в переводе издательства «Махаон» коэффициент корреляции между 5 и 7 текстами серии увеличивается с 0.9 до 1.0. Такие отличия в коэффициенте корреляции могут непрямо свидетельствовать как о различии в языковых нормах, так и о проявлении переводческого идиостиля.

Кластерный анализ позволил провести разделение текстов на сравнительно однородные группы. Кластеризация проводилась на исходных (непредобработанных) оригинальных и переводных текстах с помощью пакета Stylo для языка R [7]. Для изучения авторского и переводческого идиостилей дополнительно были взяты следующие тексты: для кластерного анализа оригинальных текстов – тексты серии книг «Cormoran Strike» R. Galbraith (J. K. Rowling) («The Cuckoo’s Calling», «The Silkworm», «Career of Evil», «Lethal White»); для кластерного анализа переводных текстов – авторские и переводные тексты М. Спивак («Твари, подобные Богу», «Черная магия с полным ее разоблачением» и перевод К. Эдварса «Дочь хранителя тайны»).

Кластерный анализ оригинальных текстов показал, что оригинальные тексты подразделяются на два кластера: первый кластер состоит из всех текстов «Harry Potter» и первых двух текстов серии «Cormoran Strike» («The Cuckoo’s Calling», «The Silkworm»), второй кластер – из оставшихся двух книг серии «Cormoran Strike» («Career of Evil», «Lethal White»). Первый кластер дополнительно разделяется на две группы -текстов серии «Harry Potter» и первые два текста серии «Cormoran Strike». Внутри кластера текстов серии «Harry Potter» также есть разделение на кластеры: в первый кластер вошли тексты серии с 1 по 4, во второй – с 5 по 7.

Такая кластеризация позволяет высказать предположение о схожести стилей написания серии «Harry Potter» и ранних работ R. Galbraith (J. K. Rowling). Выделение более поздних работ серии «Cormoran Strike» в отдельный кластер позволяет утверждать о смене стиля в написании криминальных романов. Кроме того, выделение в отдельные кластеры первого и седьмого текстов серии «Harry Potter» может также свидетельствовать о стилевых изменениях: как известно,

 DOI: 10.36078/1688391447

первый том серии планировался к публикации как самодостаточный роман, но успех после его публикации сподвиг автора к созданию продолжения [6].

Кластерный анализ переводных текстов показал, что переводные тексты также подразделяются на два кластера: в первый кластер вошли все тексты серии «Гарри Поттер», во второй – авторские и переводные тексты М. Спивак. Такая кластеризация позволяет говорить о том, что авторский сигнал оказывается сильнее, чем сигнал переводчика. Первый кластер также подразделяется на два кластера: на переводы издательства «Росмэн» и переводы издательства «Махаон». Выделение первого текста в отдельный кластер и в переводе издательства «Махаон», и в переводе издательства «Росмэн» подтверждает сохранение переводчиками авторского сигнала в переводах, однако разделение остальных текстов по кластерам, отличных от кластеров оригинальных текстов может свидетельствовать о проявлении идиостиля переводчика.

Таким образом, специфика переводческого стиля, выявленная с помощью цифровых методов, заключается в выявлении набора отклонений от литературного стиля автора. Такие методы как частотный, корреляционный и кластерный анализ позволяет в сравнении с оригинальными и схожими по направленности текстами установить тенденцию изменений идиостиля не только автора, но и переводчика. Полученные результаты также должны быть проанализированы в контексте истории создания текста и в сопоставлении с результатами, полученными другими стилометрическими методами.

Использованная литература

1. Издательства реального мира. Махаон. URL: https://harrypotter.fandom.com/ru/wiki/%D0%9C%D0%B0%D1%85%D0%B0%D0%BE%D0%BD (дата обращения: 12.05.23).
2. Издательства реального мира. Росмэн. URL: https://harrypotter.fandom.com/ru/wiki/%D0%A0%D0%9E%D0%A1%D0%9C%D0%AD%D0%9D (дата обращения: 12.05.23).
3. Калашникова Е. И. Волевич, С. Зенкин: «Хороший перевод – это незаметный перевод» // Русский журнал. Круг чтения. 2001. URL: http://old.russ.ru/krug/20010514-pr.html (дата обращения: 11.05.23).
4. Национальный корпус русского языка. URL: https://ruscorpora.ru/ (дата обращения: 10.05.23).

DOI: 10.36078/1688391447

5. Русский семантический словарь: Толковый словарь, систематизированный по классам слов и значений / Под общей редакцией Н. Ю. Шведовой. М.: Изд-во «Азбуковник», 1998.
6. Саликова Н. А. История создания и жанровые особенности гепталогии Дж. К. Роулинг «Гарри Поттер» // Веснік МДПУ імя І. П. Шамякіна. 2013. №2 (39). С. 117-121.
7. Eder M. Stylometry with R: A Package for Computational Text Analysis / M. Eder, J. Rybicki, M. Kestemont // The R Journal Vol. 8. 2016. pp. 107-121.

 DOI: 10.36078/1688391447

КОМПЬЮТЕРНЫЕ ПРОГРАММЫ ОБРАБОТКИ КОРПУСОВ ТЕКСТОВ

Замира Гидовна ХУАЖЕВА
кандидат филологических наук, доцент
Адыгейский государственный университет
(г. Майкоп, республика Адыгея)

Аннотация. Тезис посвящен компьютерным методам обработки текстов, которые основываются на синтаксисе, морфологии и грамматическом анализе. Работа программ опирается на статистическую основу — корпус текстов, которые предварительно аннотированы разработчиками и использованы для «обучения» программы, а также алгоритмическое индексирование той или иной словарной базы, обычно — словаря, каждый элемент словника, которого снабжен морфологическим модификатором (модификаторами).

Ключевые слова: компьютерный метод; методы обработки текстов; компьютерные программы; корпус текстов.

В основе программ компьютерных методов обработки текстов лежат статистическая база, в которую включены словари и морфологические модификаторы. Основная задача такого анализа состоит в том, чтобы вычленить основные смысловые доминанты и тематическую структуру произведения. По сравнению с другими программами, которые предлагаются в тезисе, они позволяют получить более точный результат. В то же время, они позволяют проверить гипотезу на более широком спектре данных с большей степенью уверенности в том, что она верна. Именно с этой точки зрения и будут рассматриваться современные программы обработки русского текста.

В первую группу входят компьютерные программы, предназначенные для анализа текстов на русском языке. Если речь идет о грамматическом срезе, то он может быть полезен для формирования целостной картины языка.

Russian Morphological Dictionary (А.Зализняк) работает с входным ASCII-текстом, в словаре около 120 тысяч слов. Реализован на SWI-Prolog для Windows. Программное обеспечение позволяет быстро и точно определить грамматический признак слова, опираясь на словарь. Если речь идет о текстах социофонной принадлежности, то это может послужить доказательством того, что вы имеете отношение к социальной сети [1,87]. Не говоря уже о том, что в этом словаре отсутствуют такие понятия, как наречие «детский» и многие другие. В то же время,

 DOI: 10.36078/1688391447

возникает проблема с определением грамматических особенностей новых слов в литературном языке.

Быстрый и удобный парсер русского языка, разработанный А. Зализняк на основе этого словаря. Программу можно загрузить как на Windows, так и на операционную систему. Приложение может работать как в консольном, так и в консольном режимах. В данном случае речь идет о морфологическом анализе литературного текста. Появляются гипотезы, основанные на частотности прилагательных [2, 45]. Это связано с сложностью установки программы, ввода необходимых параметров.

Выделение имен в текстах с неструктурированным русским языком (Nameless English Regulation). Типы сущности: люди, организации и т.д. Правила выделения сущностей основываются на правиле. Некоторые сущности могут быть связаны с внутренними словарями, если они есть. Идеально подходит для работы с системами, которые разрабатываются на основе Web-технологий. Mono работает на платформах Macintosh и Bluetooth.

В эту группу входят программные продукты для автоматического анализа текстов. Они позволяют выявить наиболее часто встречающиеся лексические единицы и составить общее представление о семантических процессах, которые протекают в изучаемом речевом продукте.

Научный инновационный продукт «ТекстАналитик 2.0» был разработан в Научно-производственном центре «МегаСистемы». Для построения семантико-семантической сети используются термины, выделенные в тексте. Позволяет искать скрытые смысловые связи между фрагментами текста. Анализируя текст, можно построить иерархическое дерево тем и подтем. Кроме того, есть возможность отредактировать текст.

Основные проблемы, с которыми сталкивается бизнес-сообщество в последние годы:

1. Анализ содержания текста осуществляется автоматически, с использованием гиперссылок;
2. Анализ содержания текста осуществляется автоматически, с использованием гиперссылок;
3. Скрытые смысловые связи слов текста с текстом;
4. В результате автоматического реферирования текста формируется его смысловой портрет;
5. В процессе кластеризации информации проводится анализ ее распределения по темам, тематическим группам;

 DOI: 10.36078/1688391447

6. Автоматическое преобразование гипертекста в текст;

7. Есть возможность ранжировать информацию о семантичности текста в зависимости от степени ее значимости;

8. Автоматическое формирование полноценной базы знаний, содержащей гипертекстовую структуру.

Автоматизированная система поиска и анализа информации в сети Интернет – GalaktikaZOOM, позволяет анализировать и обрабатывать огромный объем текста, а также извлекать из него необходимые сведения.

В процессе обработки запроса система формирует список документов, в которых содержится информация о том объекте, который ищут пользователи. В процессе работы с информационными портретами пользователь получает общее представление о объекте и информацию, необходимую для получения качественного результата.

Автоматическая обработка текстов (АОТ) является одним из инструментов, используемых в системе автоматической обработки текста. Пакет включает в себя лингвистические процессоры, которые обмениваются информацией между собой. Вход процессора, в свою очередь, является входом другого процессора. Среди представленных на рынке продуктов, можно выделить следующие:

- модуль графемического анализа текста;

- немецкие и англоязычные языки могут быть использованы в морфологическом анализе;

- автоматическое уничтожение омонимов с помощью автоматического оружия;

- семантический анализ текста состоит из четырех модулей, каждый из которых имеет свою специфику;

- система поиска лингвистических данных (Конкордс).

Нужно отметить, что существуют программные продукты, предназначенные для сбора данных в стиле и авторской манере текстов.

Слоган "Свежий взгляд" основан на анализе русского текста с помощью DOS. Программа вычисляет те места в тексте, где сходные по звучанию слова расположены близко друг к другу [2, 89]. Это приводит к парониму: «Минимальный возможный объем информации, который можно получить» и т.д. Из сайтов, посвященных поиску и анализу текстов в сети Интернет среди современных технологий, одним из самых популярных является компьютерная графика.

 DOI: 10.36078/1688391447

Автоматическое реферативное исследование не является целью для филологического анализа, однако с помощью этой программы можно получить информацию о том, какие темы волнуют профессиональное сообщество.

Худломер работает над автоматической классификацией текстов, в том числе и русскоязычных. Автором было собрано, проанализировано и систематизировано более четырех корпусов текстов. В список включены художественные и научно-популярные произведения, а также протоколы обмена сообщениями. В результате была получена эмпирическая кривая распределения длины слов в различных текстах. Эта кривая является эталоном, который используется при классификации товаров и услуг.

Литература может быть разговорной, научной или научно-популярной. В то же время, с помощью этой программы можно наблюдать за функциональной принадлежностью текста независимо от обсуждаемой темы. Позволяет анализировать тексты на английском и русском языках, а также проводить лингвистическую экспертизу текстов. Если речь идет о лингвистическом анализе, то он включает в себя не только лексику, но и семантическую составляющую.

Семантическая рубрика - это модуль, который позволяет отсортировать текст по семантическому признаку [3, 98]. Программное обеспечение включает в себя корпоративную и персональную поисковую систему. Поиск вопросов и ответов по сети Интернет реализован на основе метапоисковой системы «Аскннет». Поисковая система AQUA позволяет анализировать текстовую информацию и делать логические выводы. Программные и программные продукты, представленные на данном сайте, не являются коммерческим продуктом. Лингвистический анализ включает в себя графические, грамматические и морфологические аспекты.

Таким образом выбор программного продукта зависит только от методов предобработки текста.

Использованная литература

1. Баранов, А.Н. Введение в прикладную лингвистику: Учеб. пособие / А.Н. Баранов. – М.: Эдиториал УРСС, 2014. 260 с.
2. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2016. 248 с.
3. Захаров, В.П. Корпусная лингвистика: Учеб.-метод. пособие / В.П. Захаров. – СПб.: СПбГУ, 2018. 128 с.

DOI: 10.36078/1688391447

ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Евгения Викторовна ЮГАЙ
PhD, и.о.доцент на кафедре
«Английского языка и литературы»
Навоийского государственного
педагогического института

Аннотация. Данная статья рассматривает теоретико-методологические основы компьютерной лингвистики, её понимание и методы, используемые в теоретических и практических исследованиях компьютерной лингвистики.

Ключевые слова: компьютерная лингвистика; диалоговые агент; машинный перевод (MT); ответы на вопросы (QA).

Сегодняшняя тенденция развития компьютерных технологий и интернета даёт нам огромные возможности развития и исследования таких направлений в лингвистике, как морфология, синтаксис, лексикология, семантика уже на более высоком уровне, как компьютерная лингвистика. Компьютерная лингвистика — это научная и инженерная дисциплина, связанная с пониманием письменного и устного языка с точки зрения вычислений и созданием артефактов, которые с пользой обрабатывают и производят язык, либо в целом, либо в диалоговой среде. В той мере, в какой язык является зеркалом разума, компьютерное понимание языка также обеспечивает понимание мышления и интеллекта. А поскольку язык является нашим самым естественным и самым универсальным средством общения, лингвистически компетентные компьютеры значительно облегчили бы наше взаимодействие с машинами и программным обеспечением всех видов и предоставили бы нам под рукой, действительно отвечающие нашим потребностям, обширные текстовые и другие ресурсы интернета.

Учитывая то, что это довольно-таки новое направление научной деятельности в нашей Республике, то само исследование компьютерной лингвистики будет являться актуальным для местных исследователей данного направления. Само же развитие компьютерной лингвистики дало о себе знать еще в середине 50-х годах XX века. Когда машинный перевод (также известный как механический перевод) не смог сразу дать точные переводы, автоматическая обработка человеческих языков была признана гораздо более сложной, чем первоначально предполагалось. Вычислительная лингвистика родилась как название новой области исследований,

 DOI: 10.36078/1688391447

посвященной разработке алгоритмов и программного обеспечения для интеллектуальной обработки языковых данных. «Термин «компьютерная лингвистика» был впервые введен Дэвидом Хейсом, одним из основателей Ассоциации компьютерной лингвистики (ACL) и Международного комитета по компьютерной лингвистике (ICCL)». То есть то, что началось как попытка перевода между языками, превратилось в целую дисциплину, посвященную пониманию того, как представлять и обрабатывать естественные языки с помощью компьютеров. Так в этом направлении для перевода с одного языка на другой необходимо понимать грамматику обоих языков, включая как морфологию (грамматику словоформ), так и синтаксис (грамматику структуры предложения). Чтобы понять синтаксис, нужно было также понять семантику и лексику (или «словарь») и даже кое-что из прагматики использования языка. На сегодняшний день исследования в области компьютерной лингвистики проводятся на кафедрах компьютерной лингвистики, в лабораториях компьютерной лингвистики и на кафедрах лингвистики. Некоторые исследования в области компьютерной лингвистики направлены на создание работающих систем обработки речи или текста, в то время как другие нацелены на создание системы, обеспечивающей взаимодействие человека и машины. «Программы, предназначенные для общения человека с машиной, называются диалоговыми агентами».

Практические цели этой области широки и разнообразны. Вот некоторые из наиболее известных: эффективный поиск текста по заданной теме; эффективный машинный перевод (MT); ответы на вопросы (QA), начиная от простых фактических вопросов и заканчивая вопросами, требующими вывода и описательных или дискурсивных ответов (возможно, с обоснованиями); обобщение текста; анализ текстов или разговорной речи на предмет темы, настроения или других психологических характеристик; диалоговые агенты для выполнения конкретных задач (закупки, устранение технических неполадок, планирование поездок, соблюдение графика, медицинские консультации и т. д.); и, в конечном итоге, создание вычислительных систем с человеческими способностями к диалогу, овладению языком и получению знаний из текста.

Методы, используемые в теоретических и практических исследованиях компьютерной лингвистики, часто опираются на теории и открытия теоретической лингвистики, философской логики, когнитивистики (особенно психолингвистики) и, конечно же, информатики. Однако ранние работы с середины 1950-х до примерно 1970-х годов, как правило, были скорее нейтральными в отношении теории, а основной задачей была разработка практических методов для таких приложений,

 DOI: 10.36078/1688391447

как машинный перевод и простой контроль качества. В машинном переводе (МП) центральными вопросами были лексическая структура и содержание, характеристика «подъязыков» для конкретных областей (например, сводки погоды) и преобразование из одного языка в другой (например, с использованием довольно специальных грамматик преобразования графов или грамматик переноса). В вопрос-ответ (QA) забота заключалась в характеристике шаблонов вопросов, встречающихся в конкретной области, и связи этих шаблонов вопросов с формами, в которых могут храниться ответы, например, в реляционной базе данных.

Так как компьютерной лингвистикой могут заниматься специалисты в самых разных областях и в самых разных отделах, так и исследовательские области могут охватывать широкий спектр тем. Поэтому, наше исследование можно разделить на четыре основные области дискурса: лингвистика развития, структурная лингвистика, лингвистическое производство и лингвистическое понимание. Но более подробно мы раскроем их в нашей следующей статье.

Использованная литература

1. Shimchuk , A. O., & Yugay, E. V. (2022). Внедрение интерактивных игр на уроках английского языка для младших школьников. international conference dedicated to the role and importance of innovative education in the 21st century, 1(4), 105–110.
2. https://plato.stanford.edu/entries/computational-linguistics/
3. "Deceased members". ICCL members. Archived from the original on 17 May 2017. Retrieved 15 November 2017.
4. Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Pearson Prentice Hall.

DOI: 10.36078/1688391447

АВТОМАТИЧЕСКИЙ СБОР СЛОВ-СИНОНИМОВ ИЗ ОНЛАЙН-ТЕЗАУРУСА

Асем ШОРМАКОВА
PhD, старший преподаватель
Казахского Национального
Университета имени аль-Фараби

Аннотация. Тезис посвящен автоматическому созданию списка синонимов. Основная идея заключается в создании автоматического каталога (списка) синонимов, которые создаются из определенного набора слов. Описаны инструменты и ссылки для создания автоматизированного каталога. Приведены примеры сбора синонимов из тезауруса для автоматического создания каталога.

Ключевые слова: синонимы; тезаурус; английский; казахский язык.

Основная задача направлена на автоматическое создание каталога синонимов определенных слов. Для создания такого списка *thesaurus.com* использует синонимы неправильно переведенного английского слова. Этот процесс полностью автоматизирован. Найденные синонимы перечислены и затем переведены на казахский язык с помощью МП . Таким образом, каталог на казахском языке создается автоматически [1-4]. Этот процесс показан на рис.1.

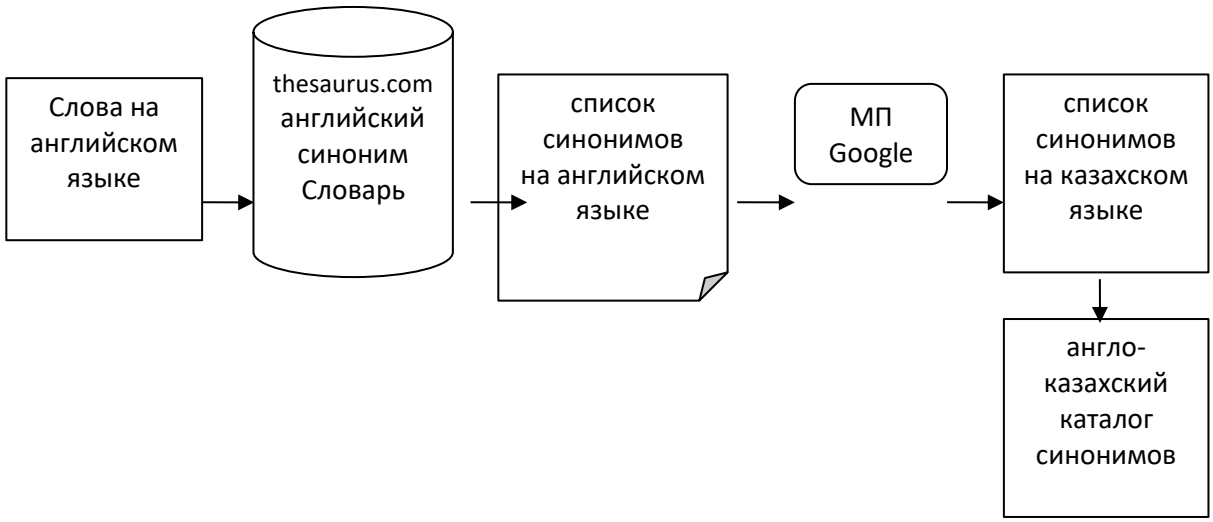


Рисунок 1 – Общая схема каталога синонимов

DOI: 10.36078/1688391447

Библиотека *PyDictionary* использовалась для формирования каталога неправильно идентифицированных слов. *PyDictionary*-это модуль словаря Python 2/3 для получения значений, переводов, синонимов и антонимов слов . С онлайн источника *Synonym.com WordNet* используется для получения синонимов и антонимов с сайта, переводчик *Google* для переводов и для получения значений слов. *WordNet* ® - большая лексическая база данных на английском языке. Кроме того, при создании каталога использовался онлайн-словарь. Словарь *thesaurus.com* взят с сайта *dictionary.com* из большой и бесплатной онлайн-тезауруса в мире[5]. *Dictionary.com* — это ведущий в мире онлайн-источник определений, происхождения слов и многого другого. Объем источника *thesaurus.com* состовляет более чем 3 миллионами синонимов и антонимов. Библиотека *Beautiful Soup* использовалась для использования синонимов с *thesaurus.com. Beautiful Soup* — это библиотека Python, используемая для извлечения данных из файлов HTML и XML, или парсер для анализа файлов HTML/XML, написанных на языке программирования Python [6].

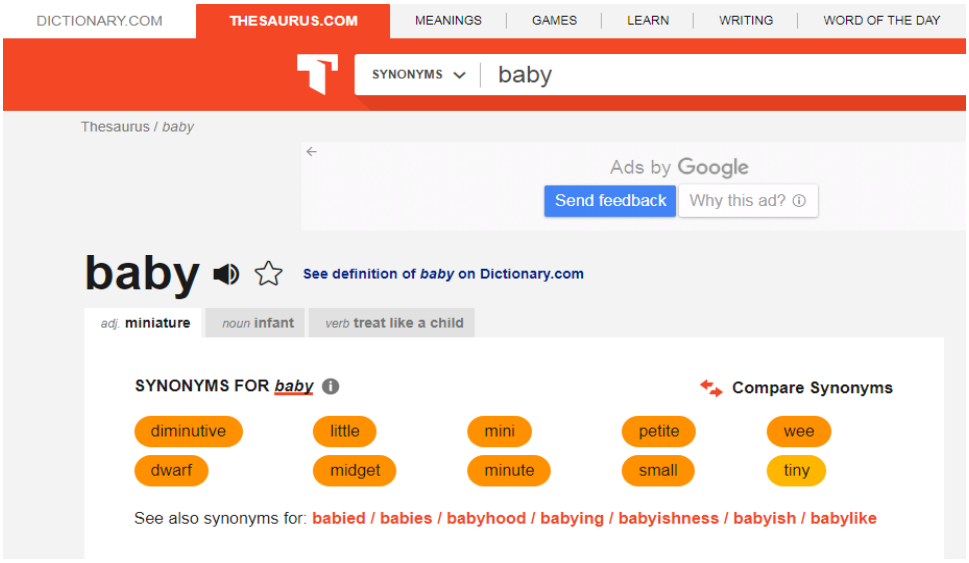


Рисунок 2 – Синонимы неправильно переведенного слова «*baby*» на Thesaurus.com

Слова, показанные на рис. 2, являются всеми возможными эквивалентами неправильных слов, например, английскими версиями неправильных слов из каталога. Мы можем вызвать thesaurus.com из нашей программы.

К пимеру у слово *baby* имеются следующие синонимы:

Baby: *diminutive, dwarf, little, midget, mini, minute, petite, small, wee, tiny* итд.

DOI: 10.36078/1688391447

Найденные синонимы на английском языке переводятся на казахский язык с помощью Google переводчика [7] и заносятся в каталог. Каждое новое неправильное слово в казахском языке и его эквиваленты записываются в новой строке справочника. Краткий фрагмент записи каталога выглядит так:

1. *Аз, жеткіліксіз, шамалы, шектеулі*
2. *...*
40. *Сыпайы, жақсы, әділет, мейірімді, сүйкімді, қолайлы, тамаша, ерекше*
41. *Сәби, ергежейлі, кішкентай, кіші, шағын* ... итд.

Словарь, используемый в каталоге, можно увидеть в следующей таблице 1.

Таблица1 – Сведения о каталоге

английский тезаурус	Количество всех синонимов в казахском языке
Более 3,5 млн.	Более 7000

По сути, это каталог синонимов казахского языка объемом 1000 строк. Каталог индексируется по первому слову. В каждой строке каталога приведены синонимы неправильных слов, то есть в каждой строке каталога встречается до пятнадцати-семнадцати синонимов при 1000 строк неправильно переведенных слов. Каждое найденное новое неверное слово и его синонимы сохраняются в новой строке каталога. В итоге написано об инструментах, используемых при автоматическом создании каталога англо-казахских синонимов. Были описаны библиотеки *PyDictionary, Beautiful Soup*, написанные и используемые на Python. Необходимые слова были автоматически взяты из онлайн-словаря синонимов *thesaurus.com*, переведены на казахский язык с помощью Google переводчика и сохранены в каталоге в новом виде. В краце описаны и объяснены на примерах модель и алгоритм автоматического формирования каталога синонимов из определенных слов.

Использованная литература

1. Шормакова А.Н., Тукеев У.А. «Технология машинного перевода с обучением английского языка на казахский язык». *Материалы международной конференции студентов и молодых ученых «Мир науки»*, 23-26 апреля 2012г. – Алматы: Қазақ университеті, – с. 154.

 DOI: 10.36078/1688391447

2. Forcada M. L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J. A., Sánchez-Martínez F., Ramírez-Sánchez G., Tyers F.M. "Apertium: a free/open-source platform for rule-based machine translation", *Machine Translation*, (Special Issue on Free/Open-Source Machine Translation) 25:2, 127-144

3. Шормакова А.Н., Айткулова А. "Добавление новой англо-казахской языковой пары в платформу машинного перевода Апертиум". *51-я Международная научная студенческая конференция «Студент и научно-технический прогресс»*, Новосибирск, 12-18 апреля 2013, Секция “Информационные технологии”. - с. 241.

4. Sundetova A., Forcada M.L., Shormakova A., Aitkulova A. "Structural transfer rules for English-Kazakh machine translation in the free/open-source platform Apertium", in *Proceedings of the I International Conference on Computer Processing of Turkic Languages (TurkLang-2013)* (Astana, 3-4 oct. 2013) , p. 322-331.

5. Онлайн словарь [Электронный ресурс]: https://www.thesaurus.com/ Запрос: 10.05.23.

6. Библиотека питон [Электронный ресурс]: https://www.crummy.com/ software/BeautifulSoup/ Сұраныс күні: 10.05.23.

7. Машинный перевод Google [Электронный ресурс] https:// https://translate.google.com// Запрос: 10.05.23.

 DOI: 10.36078/1688391447

THE DEVELOPMENT OF UZBEK STEMMER FOR UZBEK LANGUAGE

Nilufar Abdurakhmonova
National university of Uzbekistan, Tashkent, City, Uzbekistan
Ismailov Alisher Shakirovich
Tashkent Finance Institute, Andijan City, Uzbekistan
Raima Shirinova
[3]National university of Uzbekistan, Tashkent, City, Uzbekistan

Abstract. There has been a massive growth in the volume of data produced around the world in last few decades. The first data growing had begun about fifty years ago when the researchers experienced the first scientific publications coming from many different fields, later it became socialization of the Internet. The researchers needed some mechanisms to find information from these big data collections. Currently, these mechanisms are a big part of the information retrieval process. These mechanisms are a full, extensive, and specialized area of research in information technology. The main goal of information retrieval is to systematically analyze data and to extract some related data or documents that user needed or required information. Information need means that mechanism answers not only direct questions (e.g., what is stemming?), but also searches for documents that related to an entered term or group of terms that can be specific (e.g., stemming) or it is not be specific that user is not sure what she/he is looking for (e.g., algorithm to find words' roots). These straight questions, terms, and words are also known as *queries* and they represent the user's input to the system. The output could be a document title or a group of document titles and they can be ranked according to a matching percentage between the documents and the query during the process of calculation.

Keywords: stemming, information retrieval, natural language processing

Introduction

One of the important parts of the information retrieval pipeline is a stemming [1][6]. Lovins [7] defines a stemming algorithm as “a procedure to reduce all words with the same stem to a common form, usually by stripping each word of its derivational and inflectional suffixes”. The main objective of the stemming process is to remove all the possible affixes and thus reduce the word to its stem [2][8]. Using Stemming, many contemporary search engines related words with prefixes and suffixes to their word stem, to make the search broader that means that it can ensure that the greatest number of relevant matches is included in search results. Stemming has also applications in machine translation, document summarisation [9, 10], and text classification [3][11]. There are two approaches

 DOI: 10.36078/1688391447

in stemming algorithm. The first approach is stemming which means normally context-free and the main objective is to identify affixes and remove them. The second approach is lemmatization [4][12]. In lemmatization, the developer has to have a good knowledge of the language and its grammar. Lemmatization also requires a dictionary look up, therefore, lemmatization is more complex than basic stemming. Hence, in lemmatization more results that are accurate expected [5][13]. For example, a word ‘better' has a lemma ‘good'. This kind of words cannot fix in basic stemming unless the algorithm has a look-up table.

There has been a lot of resource development on stemming and stemming algorithms. However, most of the developments are conducted for popular languages with a wide range of speakers such as English, Spanish, and French. Normally, those available stemmers or stemming algorithms cannot be applied to languages that is completely different from European languages [14]. Hence, the Uzbek language lacks the resources on stemmer, information retrieval, and linguistic applications.

The review of literature review shows that stemming algorithm has not been proposed yet for extracting Uzbek root words from the Uzbek corpus, which is applicable to the above-mentioned functions. Therefore, the objective of the current research is to develop a stemming algorithm for the Uzbek language. The Uzbek stemming algorithm uses both inflectional and derivational morphemes. The output is in the form of meaningful root words without affixes. Furthermore, the accuracy and strength of the proposed algorithm will be evaluated.

Use only styles embedded in the document. For paragraph, use Normal. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text.

1.2. Problem Statement

There are more than 200 languages in the world. Every natural language has its unique characteristics and rules. For the developer, the main problem is that it is very difficult to apply same stemming algorithm on every natural language [16]. Each language has its prefixes and suffixes and as well as individual exceptions, which means it needs handling differently from one to another language. Which means you have to develop a new stemming algorithm for most of the languages.

When we look at deeper on the process of developing a stemming algorithm, it is clear that each language has a different kind of difficulties for developing a stemmer. Since we propose to develop a stemmer of Uzbek language, we have to look at what kind of problems we have to face during the development of Uzbek stemmer. First, the Uzbek language is

 DOI: 10.36078/1688391447

very different from English and western languages in terms of morphological form, grammatical rules, and composition of words. The Uzbek language is an agglutinative language with rich morphological structure. The Uzbek words created by adding suffixes and prefixes to root word. And some words are composed of a combination of two or three affixes append to root word, which makes it difficult to identify and remove the specific affixes. For example,

arra – arrala *(handsaw "noun") - (saw "verb"),*

kuch– kuchli *(strength) - (strong),*

hosil – serhosil *(crop, harvest) - (highly productive).*

Kel - Kelolmaganlardanmisiz ? (come) – (are you one of those who could not come?)

The first difficulty for developing stemmer of Uzbek is that the Uzbek language has a very rich affix which requires to identify and define them in the stemming algorithm.

The second problem is a combination of suffixes that append to a single word which makes it difficult to identify and remove them in order to stem the word.

1.3. Research Objectives

The aim of this research is to study the possibility of applying automatic word conflation to Uzbek language and to design a stemming algorithm for the Uzbek language. In order to achieve this aim, the research covers the following three stages:

- To study the existing stemming algorithms;
- To develop an Uzbek stemming algorithm;
- To test and evaluate the designed Uzbek stemmer.

The objective of this research is to achieve automatic word stemming for Uzbek language and to examine the performance and effectiveness of the Uzbek stemming algorithm.

1.4. Motivation of the research

The Uzbek data on the internet is increasing every year. To get information is always include information retrieval. There are may be a lot of research is done on stemming and information retrieval in the Uzbek language. However, there lack of information on the internet about stemming algorithm [17]. Even if there is information about stemming, it is almost impossible to find the implemented algorithm from the internet to test it. The motivation of this research is to develop a stemming algorithm for Uzbek language and test it. The Uzbek stemmer available on the internet through this link: http://stemming.uz

DOI: 10.36078/1688391447

II. CONSTRUCTION OF THE UZBEK STEMMING ALGORITHM

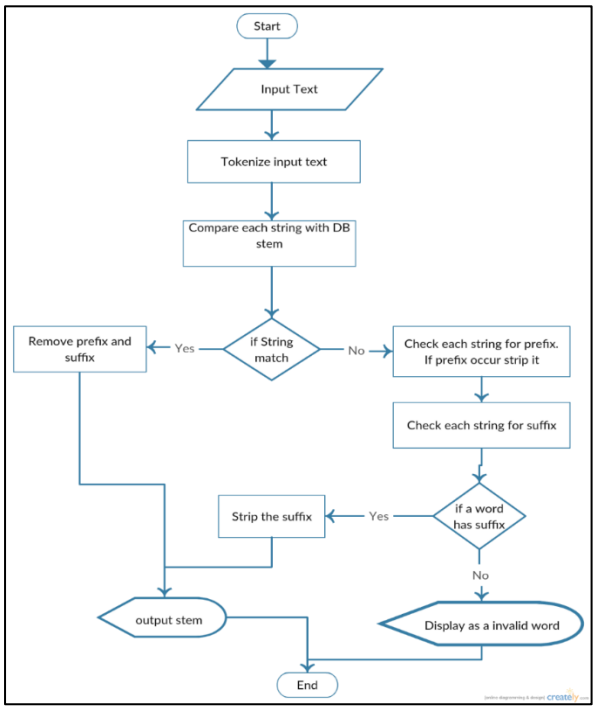
The base model is an original version of the model that can be enhanced with new features [44]. This section describes the base model of the Uzbek stemming algorithm. The Uzbek stemmer is an affix removal stemmer. The Uzbek stemming algorithm has two procedures:

1. Dictionary check-up
2. Affix removing

In dictionary check-up procedure we have a set of root words in the database (stem lexicon). The algorithm will compare the input word with stem lexicon using string similarity method. In this step, the input words only compare right to left character comparison. If an input word match with database root word (all characters must match according to word's length) then the algorithm will display it as an output stem.

The second procedure of the Uzbek stemmer is affix removing. The affix removing procedure has two steps, prefix removing, and suffix removing. The affix removing procedure will be used only if the input word does not match with stem lexicon. In the first step of the affix removing procedure, the algorithm will check the input word according to defined prefixes, if the input word has prefixes then the algorithm will remove them. Next step is to scan the prefix removed a word for suffixes if a word has suffixes then stemmer will remove the suffixes and display the affixes stripped word as an output stem.

2.4.3. Flowchart of the Uzbek stemming algorithm



The flow of the Uzbek stemming algorithm has shown in Figure 4.4.

DOI: 10.36078/1688391447

Figure 4.4: Uzbek stemming algorithm flowchart

The process starts with input text, which is entered Uzbek language words. The algorithm will tokenize the input text, which includes the process of removing all the symbols, punctuations, and separate each word by space. Once, all the input words tokenized then the algorithm will compare each input word with stem lexicon. If the input word matches with stem lexicon then shows the matched word as an output stem. If the input word does not match with stem lexicon then, the input word will go to next procedure, which is removing prefix and suffix by checking each input word with defined prefix and suffix in the algorithm. If the input word does not match with stem lexicon and the input word does not have any prefix or suffix that defined in the algorithm then the algorithm will reject this word and display it as an *invalid word*.

2.5. Pseudo code

Pseudo code is a detailed, however readable description of what a computer program or algorithm must do, expressed in a formally styled natural language more rather than in a programming language [32]. The following section will display the pseudo code of the Uzbek stemming algorithm.

Table 4.2: Pseudo code for tokenization

Input -> Tokenize(input);
If TokenizeValidate(input) then return input;

Table 4.3: Pseudo code for compare input text with stem lexicon

for each Word in Words do
if Input Match with DB stem then stem-> extractStem(input); else return null;

2.6. Implementation of Uzbek stemmer

The Uzbek stemmer is implemented by using Java object oriented programming language and SQLite database management system for supporting database. It has a user-friendly interface. The Uzbek stemmer is available in two languages English and Uzbek languages. Figure 4.5 and figure 4.6 shows the graphical user interface of the Uzbek stemmer. Uzbek

 DOI: 10.36078/1688391447

stemmer available online that everyone can download and test it on their personal computer. There is only one requirement is PC should have Java 8 kit.

3.1. Experiment of the Uzbek stemmer

Several of experiments is conducted to test the overall performance of the Uzbek stemmer. The implemented Uzbek stemmer was tested on different data sets to be evaluated. The data sets were created for the testing of the Uzbek stemmer. This data sets cover areas such as history, human body, historical figure and art. The researcher tested the algorithm on a sample of three data sets contains combined 4086 words collected from all categories (all three text files content is shown in Appendix A1, A2, and A3).

Sample 1 holds 2261 words and stemmer generated 1996 stems. Table 5.1 shows some output of sample two generated by the Uzbek stemmer.

Table 5.1: Testing sample 1

Input text	Output	Expected Output
Input	Output	Expected Output
Asrning	Asr	Asr
Ismi	Ism	Ism
Amir	Amir	Amir
ko`ragoniy	ko`ra	ko`ragoniy
Kuchli	Kuch	Kuch
tarag`ay	Tara	tarag`ay
Tarixiy	Tarix	Tarix
Burqul	Bur	Burqul

Sample 2 holds 602 words and stemmer generated 574 stems. Table 5.2 shows some output of sample 2 that generated by the Uzbek stemmer.

3.2.1. Correctness

In stemming algorithm there are two main errors, over stemming and under stemming. Over stemming – is occurs when two different inflected words are stemmed to the same root form, that they should not have been stemmed to same root. It is also called as a false positive. Under stemming – is occurs when two different inflected words should be stemmed to the same root form, but they are not stemmed to same root. It is also known as a false negative. Stemming algorithms try to reduce each type of error, however reducing one type of error may lead to increasing the other type of error.

To evaluate the performance of the Uzbek stemmer, manual counting approach was applied. The Uzbek stemmer was tested three sample text files that are prepared by taking

 DOI: 10.36078/1688391447

randomly from the sample text document that contains Uzbek text. Table 5.4 shows the evaluation by error rate and accuracy of the Uzbek stemmer.

Table 5.4: Evaluation of error of the Uzbek stemmer

Text	Sample 1	Sample 2	Sample 3
Total words	2261	602	1223
Stemmed words	1996	574	1133
Over-stemming	116 (5.81%)	34 (5.92%)	63 (5.5%)
Under-stemming	18 (0.90%)	6 (1.04%)	7 (0.61%)
Accuracy	93%	93%	94%
Average Accuracy	93.30%		93.30%

Sample text 1 contains 2,261 words. In addition, there are over-stemming and under stemming observed. Uzbek stemmer generated 1996 stems from the sample text 1. Out of these words 0.90% (18) words were under stemmed, and 5.81% (116) words were over-stemmed. Total percentage of the stemming error occurred was 6.71 % (134 words). As a result, the accuracy of the Uzbek stemmer becomes 93.29 %.

Sample text 2 holds 602 words. The stemmer generated 574 stems from the sample text 2. Out of these words 1.04% (6) words were under stemmed, and 5.92% (34) words were over-stemmed. Total percentage of the stemming error occurred was 6.96 % (40 words). The accuracy of the Uzbek stemmer for sample text 2 is 93%.

Sample text 3 contains 1223 words. Uzbek stemmer generated 1133 stems from the sample text 3. Out of these words 0.61% (7) words were under stemmed, and 5.5% (63) words were over-stemmed. Total percentage of the stemming error occurred was 6.1% (70 words). As a result, the accuracy of the Uzbek stemmer becomes 94%.

3.2.2. Retrieval effectiveness

Retrieval effectiveness can be measured with recall and precision as well as stemmers' speed, size, and so on [52][54].

The researcher evaluates the Uzbek stemmer's retrieval effectiveness by applying recall and precision measurement. For the calculation of precision and recall measurement following items considered:

 DOI: 10.36078/1688391447

True positive (TP) – output which matches with database root word.

False positive (FP) – output which has specific prefix/suffix that has removed during the stemming process.

False negative (FN) – output which is not match with database root word and not match specific prefix/suffix. It is also called invalid words.

The precision calculated as following:

$$Precison = TP/(TP + FP)$$

Where the precision equal to a total number of true positive output divided by a collection of a total number of true positive output and false positive output.

The recall calculated as following:

$$Precison = TP/(TP + FN)$$

Where the recall is equal to a total number of true positive output divided by combination of total number true positive output and false negative output.

Figure 5.1 shows the precision and recall measurement rate for the Uzbek stemmer. The sample text 1 produced 96.79% precision and 87.93% recall, sample text 2 generated 98.43% precision and 95.27% recall, and sample text 3 has 97.35% precision and 92.45% recall.

3.2.3. Compression Performance

Another way to evaluate the stemmer is word compression ratio. For calculation the compression rate(C), the formula is shown as follow [31]:

$$C = 100 * (W - S)/W$$

Where

W is number of unique words before Stemming

S is number of unique stems after Stemming

Table 5.5 shows the compression ratio for tested three sample texts. The sample text 1 has 48.07%, sample text 2 generated 29.73% and sample text 3 has 37.53% compression ratio after the calculation. The average compression ratio is 38.44% for the stemmed words. It reduces the sample text by 38.44%. The experiment proved that large documents give high compression ratio, the reason is that in large documents, the frequency of the same concept words high.

 DOI: 10.36078/1688391447

Table 5.5: Evaluation compression ratio of the Uzbek stemmer

Text	Sample 1	Sample 2	Sample 3
Total words	2261	602	1223
Stemmed words	1996	574	1133
Unique stemmed words	1174	423	764
Word compression ratio	48.07%	29.73%	37.53%
Average compression ratio	38.44%	38.44%	38.44%

3.3. Comparison of the Uzbek stemmer with existing stemmer.

Table 5.6 shows the comparison of the Uzbek stemmer index compression factor with Lovins, Porter and Paice/Husk stemmer. Table 5.6 is experiment done by Lennon, Frakes, Paice and Harman. From the table 5.6 we can conclude that the Uzbek stemmer ICF is higher than Porter stemmer ICF, it is close to Lovins stemmer ICF and the Uzbek stemmer ICF is lower than Paice/Husk stemmer ICF.

Table 5.6: The comparison of Uzbek stemmer ICF with existing stemmers

Stemmer	Lennon et al.	Frakes	Paice	Harman
Lovins	from 30.9% to 45.8%	29%	44.60%	38.23%
Porter	from 26.2% to 38.8%	17%	38.90%	28.74%
Paice/Husk	-	33%	51.30%	-
Uzbek stemmer	from 29.73% to			

3.4. Strength and weakness

The experiment and evaluation showed that the Uzbek stemmer has high accuracy rate with average of 93.30 %. The Uzbek stemmer generated average 97.52% precision and 91.88% recall, which considered the high result. The compression ratio of the Uzbek

DOI: 10.36078/1688391447

stemmer is from 29.73% to 48.07% after testing three samples. We can summarize the strength of the Uzbek stemmer is it produces high accuracy output and it generates high precision and recall, it has a low error rate, and it has high index compression factor rate. The weakness of the Uzbek stemmer is that, it is time-consuming and it may reject valid Uzbek words if the word is outside the lexicon and the word should not have defined affixes in the algorithm.

4.1. Future work recommendations

In this research, the Uzbek stemmer used two approaches to developing the stemmer. The database check-up approach where all the root words stored in the database and stemming done by comparing input word with database root word. The second approach is to affix removal approach where Uzbek prefixes and suffixes defined in the algorithm and algorithm check each word if the word contains defined prefixes or suffixes then the algorithm will remove them.

For the future recommendation, researchers may try to achieve stemming by lemmatization method where the algorithm has to understand the part of speech (POS). The lemmatization method provides more detailed information about lemma (stem), and it can differentiate the noun and verb to analysing purpose.

The second recommendation is to propose and develop a platform for Turkic family language. All the Turkic family languages are agglutinative language and morphology of the all the Turkic family languages are similar to each other. Hence, the researchers may purpose single stemming algorithm framework that includes all the Turkic family languages.

REFERENCE

[1] Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. Information Research, 19

[2] Christopher D. Manning, P. R. (2008). Introduction to Information Retrieval. Cambridge University Press.

[3] Chowdhury, G. (2010). Introduction to Modern Information Retrieval, Third Edition. Facet Publishing ©2010 .

[4] Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press

[5] Bruce Croft, J. L. (2013). Language Modeling for Information Retrieval. Springer Science & Business Media.

 DOI: 10.36078/1688391447

[22] Adamson, G. W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information storage and retrieval*, *10*(7), 253-260.
[23] Bijal, D., & Sanket, S. (2014). Overview of Stemming Algorithms for Indian and Non-Indian Languages. arXiv preprint arXiv:1404.2878.
[24] Lovins, J. B. (1968). *Development of a stemming algorithm* (p. 65). Cambridge: MIT Information Processing Group, Electronic Systems Laboratory.
[25] Jivani, A. G. (2011). A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl, 2(6), 1930-1938.
[26] Karaa, W. B. A. (2013). A new stemmer to improve information retrieval. International Journal of Network Security & Its Applications, 5(4), 143.
[27] Porter, M. F. (1980). An algorithm for suffix stripping. Program: Electronic Library and Information Systems, 40(3), 211-218.
[28] Paice, C. D. (1994). An evaluation method for stemming algorithms. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 42-50). New York, NY: Springer-Verlag.
[29] Lennon, M., Pierce, D. S., Tarry, B. D. & Willett, P. (1988). Document retrieval systems. In P. Willett (Ed.), Document retrieval systems, (pp. 99-105). London: Taylor Graham Publishing.
[30] D. Suleymanov, A. Gatiatullin, N. Prokopyev, Abdurakhmonova N. Turkic Morpheme Web Portal as a Platform for Turkology Research International / Conference on Information Science and Communications Technologies ICISCT 2020 (Indexing Scopus) Applications, Trends and Opportunities 4th, 5th and 6th of November 2020, Tashkent Uzbekistan https://ieeexplore.ieee.org/document/9351500
[31] N. Abdurakhmonova, I. Alisher and R. Sayfulleyeva MorphUz: Morphological Analyzer for the Uzbek Language Bosma 2022 7th International Conference on Computer Science and Engineering (UBMK), 2022, doi: 10.1109/UBMK55850.2022.9919579. pp. 61-66.
[32] N. Abdurakhmonova, I. Alisher and G. Toirova Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing 2022 7th International Conference on Computer Science and Engineering (UBMK), 2022, doi: 10.1109/UBMK55850.2022.9919521 pp. 73-75
[33] N. Z. Abdurakhmonova, A. S. Ismailov and D. Mengliev, "Developing NLP Tool for Linguistic Analysis of Turkic Languages," 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russian Federation, 2022, pp. 1790-1793, doi: 10.1109/SIBIRCON56155.2022.10017049

 DOI: 10.36078/1688391447

ВОЗМОЖНОСТИ КОРПУСНОЙ ЛИНГВОДИДАКТИКИ В ОБУЧЕНИИ ИНОСТРАННОМУ ЯЗЫКУ В КОНТЕКСТЕ ОТКРЫТЫХ ОБРАЗОВАТЕЛЬНЫХ РЕСУРСОВ

Дмитриев Александр Владиславович
кандидат филологических наук, доцент
Санкт-Петербургский политехнический университет Петра Великого
avd84@list.ru
Коган Марина Самуиловна
кандидат технических наук, доцент
Санкт-Петербургский политехнический университет Петра Великого
m_kogan@inbox.ru

Аннотация. Одним из препятствий для широкого использования корпусной лингводидактики для изучения ИЯ является тот факт, что многие корпуса, описанные в научных статьях, создаются исключительно для исследовательских целей и недоступны научной и педагогической общественности. Проблемой также является недостаточная компетенция преподавателей иностранного языка в области корпусной лингвистики. Находящиеся в открытом доступе корпуса являются ценным ресурсом для лингвистического анализа, но разработка дидактических материалов на их основе для учебных целей при изучении родного и иностранного языков является весьма трудной и трудоемкой задачей для большинства преподавателей иностранных языков в школе и вузе. Решением проблемы может стать разработка набора заданий/упражнений, привязанных к конкретному доступному корпусу.

Ключевые слова: ООР, онлайн корпуса, корпусная лингводидактика, дидактические материалы, повышение квалификации

Развитие интернета и информационного общества привело к появлению важного тренда на открытое образование, науку, информацию. Эта тенденция поддержана документами ООН и UNESCO, решениями двух всемирных международных конгрессов по развитию открытых образовательных ресурсов (ООР), важнейшими характеристиками которых являются бесплатный доступ, право на использование, адаптацию и распространение третьими лицами без ограничений или с незначительными ограничениями [1, 2, 3].

Теоретически одной из ключевых особенностей концепции ООР в свете преподавания иностранного языка (ИЯ) является возможность исследователя обратиться к имеющемуся потенциалу корпусной лингводидактики. Однако на практике между исследованиями в области корпусной лингвистики и методикой

 DOI: 10.36078/1688391447

преподавания ИЯ имеется существенный разрыв, на что в свое время обратили внимание Болтон (Boulton) и Кобб (Cobb), а позже и А. Чамбес (Chambers) [4, 5].

Причинами такого разрыва, на наш взгляд, являются, во-первых, недоступность многих корпусов, описанных в научных статьях, во-вторых, отсутствие общедоступных заданий, созданных на базе находящихся в свободном доступе онлайн корпусов, для непосредственного использования в учебном процессе, и в-третьих, недостаточная подготовка преподавателей ИЯ по корпусной лингводидактике [5].

Последняя проблема, кажется, гораздо более серьезная. С одной стороны на современном этапе в соответствии с требованиями ФГОС все студенты филологических специальностей в России получают базовые знания о корпусной лингвистике, некоторые имеют публикации по теме использования подходов корпусной лингвистики в обучении ИЯ [6], однако открытым остается вопрос, достаточно ли обучающимся односеместрового курса по использованию корпусов и подходов корпусной лингвистики в обучении ИЯ и насколько студенты готовы регулярно использовать эти знания в практической деятельности после окончания обучения [7, p. 234]. С этой проблема сопряжена также другая: недостаточная проработанность методики внедрения корпусных технологий в образовательный процесс в высших учебных заведениях, как у лингвистов, так и у студентов, изучающих английский язык для специальных целей, о чем неоднократно нами упоминалось ранее [8, 9, 10].

Для решения этой проблемы существует большое количество самых разнообразных ООР, которые способны повысить квалификацию по корпусной лингвистике – как у студентов и выпускников, так и у самих преподавателей. К ним относятся коллекции YouTube, МООС и записи вебинаров, организованные ведущими центрами по корпусной лингвистике. Среди них можно назвать следующие:

1. Лекции таких специалистов в этой области, как Э. Ле Фолл – по освоению Sketch Engine и созданию собственного специального корпуса в корпусе COCA, Л. Энтони5 – по использованию основных функций программы AntConc версии 3.4.0 (которую он сам и разработал), В.А. Плунгяна, Д.О. Добровольского в Постнауке, Д. Бузаджи, Б.В. Орехова, В.П. Захарова и некоторых других известных исследователей, а также видеозаписи заседаний конференции «Корпусная лингвистика»6, которая проходила в дистанционном режиме на кафедре Математической лингвистики СПбГУ в 2021 г. В рамках этой конференции впервые была организована секция «Корпусы в учебных целях» и Круглый стол «Использование корпусных подходов в преподавании».

[5] Ссылка на серию видео по освоению программы AntConc Версии 3.4.0. URL: https://www.youtube.com/playlist?list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj.

[6] Ссылка на записи заседаний конференции Corpora 2021. URL: https://www.youtube.com/playlist?list=PLvN-Hgi1JTKTVM29etGSGf9nrAHAiUT8.

DOI: 10.36078/1688391447

2. MOOC *Corpus Linguistics: Method, Analysis, Interpretation*[7], в рамках которого, в частности, рассматриваются крупные онлайн корпуса и новый корпусный менеджер LancsBox, разработанный В. Брезиной; курс *Введение в корпусную лингвистику*[8], дающий представление о практическом применении достижений современной корпусной лингвистики и позволяющий освоить базовые методы корпусного преподавания в повседневной работе; SPOC[9] *Improving writing through corpora: Data-driven learning*[10], способствующий развитию навыков академического письма с прямым обращением к корпусу [11]; целью учебного курса *Corpora in Language Teaching*[11], разработанного на основе общедоступных корпусов и корпусных инструментов, находящихся в открытом доступе, является знакомство участников с базовыми концепциями корпусной лингвистики, применение корпусных подходов в обучении иностранным и родному языку и развитие навыка создания учебных материалов на основе корпусного подхода.

3. Серия вебинаров LAEL Webinars, организованных в 2020 г. университетом Сан-Паолу (Бразилия)[12]; вебинары 2021-2022 гг., организованные на базе Школы языков и культур университета Квинсленда (Австралия), по теме *International Perspectives on Corpus Technology for Language Learning*[13].

Обратимся теперь к некоторым готовым ООР для использования подходов корпусной лингводидактики в обучении ИЯ.

Интересен опыт Нины Вяткиной, которая обратила внимание, что руководства пользователя к таким корпусам, как НКРЯ и COCA, очень далеки от знакомых учителю планов уроков с использованием корпусов. Она отметила, что инструкции для преподавателей должны базироваться на педагогических принципах, подразумевающих введение материала, сопровождение, вовлечение, и идти от простого к сложному, от большей поддержки со стороны преподавателя к заданиям, выполняемым самостоятельно. На материале корпуса немецкого языка DWDS Н. Вяткина разрабатывает задания для учащихся с разным уровнем владения немецким языком (от начинающих до продвинутых), предполагающих разный формат работы (индивидуальную работу, работу в парах и малых группах, введение материала преподавателем). Основным фокусом является расширение словарного запаса, и

---

[7] Разработан в Ланкастерском университете Т. МакЭнери и размещен на платформе FutureLearn: https://www.futurelearn.com/courses/corpus-linguistics.

[8] Разработан в ВШЭ А.И. Левинзоном и размещен на Национальной платформе «Открытое образование»: https://openedu.ru/course/hse/CORPUS/?session=fall_2020.

[9] Small private online course: маленький частный онлайн курс.

[10] Разработан П. Кроссвейтом в университете Квинсленда и размещен на платформе EdEx: https://edge.edx.org/courses/course-v1:UQx+SLATx+2019/about. Задания разработаны на основе открытого корпуса BAWE из коллекции корпусов Sketch Engine. Апробация курса описана в статье [11].

[11] Разработан А. Ленко-Шиманска для студентов Варшавского университета и находится в свободном доступе на сайте университета: https://moodle.ils.uw.edu.pl/course/view.php?id=101. Апробация курса описана А. Ленко-Шимански в статьях [7].

[12] Ссылка на записи вебинаров в честь 50-летия кафедры LAEL с презентациями докладчиков: http://corpuslg.org/lael_english/webinars/.

[13] Ссылка на записи вебинаров с описанием и презентациями докладчиков: https://languages-cultures.uq.edu.au/event/7334/international-perspectives-corpus-technology-language-learning-seminar-series.

DOI: 10.36078/1688391447

задания могут включать освоение новых лексико-грамматических характеристик слов с помощью примеров из корпуса, верификация примеров употребления слов/словосочетаний в учебниках примерами из корпуса. Примеры заданий, приведенные в статье[14], соответствуют разным инструментам корпуса DWDS: базовый поиск, временная ось, визуализация коллокаций через облако слов и др.[12].

Другим блестящим примером создания ООР для преподавания ИЯ является проект *CORPUS FOR SCHOOLS: Teaching English Language with Corpus Linguistics*, создаваемый в Ланкастерском университете под руководством Д. Габласовой. Учебные материалы привязаны к большим выборкам устной речи BNC 1990-х и 2014, представляющих неформальные беседы носителей британского варианта английского языка, и включают раздаточные упражнения. Уроки разработаны как для носителей английского языка, готовящихся сдать итоговый экзамен по курсу средней школы (A-level exam), так и для изучающих английский язык как второй или иностранный[15].

Еще одним примером привязки планов уроков к корпусу является мультимедийный ресурс *English Central*. Он создан с использованием корпусных технологий и содержит более 14000 видео и 250 млн строк аудиозаписей для развития навыков аудирования и говорения. Контент содержит тысячи отдельных видео с указанным уровнем сложности и более 100 мини-курсов в категориях: *Academic English, Travel English, Business English, Career English, Useful Expressions* и др. Для каждого видео урока имеется интерактивный План урока (Lesson Plan), который предлагает задания на прослушивание фрагментов видео, составление собственных предложений со словами, которые тренировались ранее, на понимание видеосюжета в формате множественного выбора. В последнем задании предлагается четыре вопроса для обсуждения: первый просит кратко сформулировать основную идею видеосюжета, а три других вопроса побуждают высказать собственное мнение на основе информации видео урока.

Не менее успешной практикой по интегрированию корпусной лингвистики в обучение ИЯ является платформа *The Corpus-Aided Platform for Language Teachers* (CAP), созданная на кафедре лингвистики и современных языков в университете Гонконга под руководством К. А. Ма[16]. На платформе представлены материалы для повышения квалификации/самообразования преподавателей в области корпусной лингвистики и лингводидактики, небольшие учебные видеоролики продолжительностью 5-7 минут для грамматики, произношения, развития письменных навыков, работе с параллельным корпусом и др. В разделе *Teaching*

[14] В полном объеме курс на базе DWDS доступен на сайте проекта языкового ресурсного центра Канзасского университета: https://corpora.ku.edu .
[15] BNClab tutorial 2018: https://www.youtube.com/watch?v=28EFVcak99Q&t=18s.
[16] The Corpus-Aided Platform for Language Teachers (CAP): https://corpus.eduhk.hk/cap/.

DOI: 10.36078/1688391447

Activities помещены примеры учебных материалов и детально разработанных планов уроков продолжительностью 80-120 мин. по категориям: обучение лексике, грамматике, произношению, задания на основе параллельного англо-китайского корпуса. Применение платформы CAP на курсах повышения квалификации преподавателей английского языка описано в статье [13].

В контексте использования корпусных технологий при создании ООР весьма любопытен опыт Э. Ле Фолл, которая вместе со своими студентами-магистрантами, будущими преподавателями иностранного языка, разработала учебные материалы и задания на основе материалов открытых онлайн корпусов по разным изучаемым темам и разместила их на своем сайте[17]. Материалы представляют четкие инструкции по поиску в разных корпусах и разработаны для конкретной категории учащихся: в 1-й части для начальной и средней школы, во 2-й – для старших классов средней школы, в 3-й – для использования разработанных материалов, в 4-й – для английского для специальных целей и профессионально-ориентированного языка [14]. Все материалы созданы на основе доступных онлайн корпусов и корпусных инструментов, список которых приведен в приложении пособия[18]. По мнению разработчиков, это должно способствовать более широкому применению подходов КЛ в повседневной педагогической практике преподавателей ИЯ.

Некоторым аналогом этого проекта является размещенные в разделе сайта НКРЯ Studorium результаты проекта *Своевольные смыслы*[19], направленного на изучение и описание изменений в значении и употреблении 30 лексем русского языка на протяжении XIX-XX вв., выполненный студентами ВШЭ под руководством Н.Р. Добрушиной и М.А. Даниель в 2010-2012 гг.

На основе проанализированного материала можно сделать следующие выводы.

Во-первых, не все доступные онлайн корпуса и другие корпусные ресурсы являются ООР в строгом смысле этого слова, т.е. обладают базовыми характеристиками ООР, описанными в литературе. Однако, свободный доступ к ним делает их ценными инструментами для самообразования преподавателей ИЯ, желающих повысить свой профессиональный уровень и понять, как можно использовать подходы корпусной лингвистики в своей практике, а также как разрабатывать учебные материалы на основе корпусных данных.

Во-вторых, анализ учебных материалов/планов уроков, созданных как ООР, показал, что ряд вопросов/идей/высказанных предложений требует дальнейшего исследования. Так, очень привлекательны идеи Н. Вяткиной о разработке учебных материалов в привязке к определенному корпусу, разработки Д. Габласовой по

[17] Сайт Элен Ле Фолл: https://elenlefoll.pressbooks.com/.
[18] Приложение пособия Э. Ле Фолл: https://elenlefoll.pressbooks.com/back-matter/appendix.
[19] Своевольные смыслы: опыт микроисторического исследования лексики 19 – 21 веков URL: https://studiorum.ruscorpora.ru/meanings/.

DOI: 10.36078/1688391447

устному компоненту BNC на сайте BNClab, планы мини-уроков к каждому видеофрагменту на сайте *English Central*.

В-третьих, требует дальнейшего исследования вопрос о том, что эффективнее и методически целесообразнее: двухстраничный план мини-урока с заданиями или план полноценного урока продолжительностью 80-120 минут (как у Ма) объемом в десятки страниц?

В-четвертых, в созданных ООР недооценивается основной учебник как ресурс/источник для разработки заданий на основе подходов корпусной лингвистики, хотя разработка заданий с использованием подходов корпусной лингвистики для изучения конкретной темы на основе определенного учебника также может быть весьма эффективной и привести к ряду планов мини-уроков, в качестве дополнения к конкретному учебнику (по модели, реализованной создателями *English Central*).

В-пятых, совершенно очевидно, что создание качественных ООР с использованием корпусов – очень трудоемкий процесс. Идея вовлечения студентов, будущих педагогов, к их разработке в рамках проектной деятельности заслуживает всяческого внимания, поддержки и внедрения в разных учебных заведениях, особенно в условиях, когда Российские стандарты уделяют первостепенное внимание проектной деятельности студентов.

Таким образом, тренд на открытое образование, науку и информацию, всемерно поддерживаемый ООН и ЮНЕСКО, может помочь более широкому внедрению методов корпусной лингводидактики в повседневную педагогическую практику преподавателей иностранных языков.

Использованная литература

1 Madalli D. P. UNESCO Concepts of Openness and Open Access. Paris. 2015. 67 p.

2 Blyth C.S., Thoms J.J. Introduction: Second Language Education as an Open Knowledge Ecology // Open Education and Second Language Learning and Teaching: The Rise of a New Knowledge Ecology (C.S. Blyth, J. J. Thomas – eds.). Bristol, Blue Ridge Summit: Multilingual Matters. 2021. P. 1-22. URL: https://www.degruyter.com/document/doi/10.21832/9781800411005-002/html (дата обращения:10.05.2023).

3 Kosmas P., Parmaxi A., Perifanou M., Economides A.A. Open Educational Resources for Language Education: Towards the Development of an e-Toolkit // Learning and Collaboration Technologies: New Challenges and Learning Experiences. HCII 2021. Lecture Notes in Computer Science (P. Zaphiris, A. Ioannou – eds.). 2021. Vol №12784. P. 65–79. https://doi.org/10.1007/978-3-030-77889-7_5.

4 Boulton A., Cobb T. Corpus use in language learning: A meta-analysis // Language learning. 2017. Vol. 67, №2. P. 348–393.

DOI: 10.36078/1688391447

5 Chambers A. Towards the corpus revolution? Bridging the research–practice gap // Language Teaching. 2019. Vol. 52, №4. P. 460–475. https://doi.org/10.1017/S0261444819000089.
6 Захаров В.П., Коган М.С. Использование корпусов в изучении и преподавании языка в России: достижения, проблемы, перспективы // O‘zbek milliy va ta’limiy korpuslarini yaratishning nazariy hamda amaliy masalalari. Vol. 1, №. 01. 2021. С. 15–19.
7 Leńko-Szymańska A. Training teachers in data driven learning: Tackling the challenge // Language Learning & Technology. 2017. Vol. 21, №3. P. 217–241.
8 Dmitriev A., Kogan M. The Role of Corpus Linguistics in the Training of Specialists in the Field of Computer Language Teaching. In: Integrating Engineering Education and Humanities for Global Intercultural Perspectives, Proceedings of the Conference “Integrating Engineering Education and Humanities for Global Intercultural Perspectives”, 25-27 March 2020, St. Petersburg, Russia, 511–520. (2020) doi:10.1007/978-3-030-47415-7_54.
9 Вдовина Е.К., Дмитриев А.В., Коган М.С. Теоретико-прикладное значение корпусов в компьютерной лингводидактике // Litera №1, 2020. С. 200–216. DOI: 10.25136/2409-8698.2020.1.32219.
10 Дмитриев А.В., Коган М.С. Потенциал корпусной лингвистики в подготовке специалистов в области компьютерной лингводидактики // Научно-технические ведомости СПбГПУ. Гуманитарные и общественные науки. Т. 10, №4, 2019. С. 69–85. DOI: 0.18721/JHSS.10407.
11 Crosthwaite P. Taking DDL online: Designing, implementing and evaluating a SPOC on data-driven learning for tertiary L2 writing //Australian Review of Applied Linguistics. 2020. Vol.43 №2. P. 169-195. https://doi.org/10.1075/aral.00031.cro.
12 Vyatkina N. Language corpora for L2 vocabulary learning: Data-driven learning across the curriculum // Understanding vocabulary learning and teaching: Implications for language program development (P. Ecke, S. Rott – eds.). Boston, MA: Cengage Learning. 2018. P. 121–145.
13 Ma Q., Yuan R. (E.), Cheung L.M. E., Yang J. Teacher paths for developing corpus-based language pedagogy: a case study // Computer Assisted Language Learning. 2022. https://doi.org/10.1080/09588221.2022.2040537
14 Le Foll E (ed.). Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for teachers using online resources. Osnabrück: Pressbooks. 2021 URL: https://elenlefoll.pressbooks.com (дата обращения:10.05.2023).

 DOI: 10.36078/1688391447