прыпынку і злучнікі. На наступным этапе адбудзецца аналіз алгарытмаў на аснове тэсціравання існуючых лінгваакустычных рэсурсаў, якія пакрываюць усю разнастайнасць сінтаксічных канструкцый беларускай літаратурнай мовы. Гэта дазволіць выявіць недакладнасці ў прадстаўленым прататыпе, якія будуць выпраўлены ў найбліжэйшы час.

БІБЛІЯГРАФІЧНЫЯ СПАСЫЛКІ

1. Зяноўка, Я.С. Перспектывы развіцця аўтаматычнай апрацоўкі вуснага маўлення / Я.С. Зяноўка, М.В. Супрунчук, Ю.С. Гецэвіч // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития / Белорусский гос. ун-т ; редкол.: Н. А. Куркович (гл. ред.) и др.. – Минск : БГУ, 2022. – С. 252-257.
2. Зяноўка, Я.С. Прынцыпы аўтаматычнага вызначэння інтанацыйных партрэтаў беларускага маўлення / Я. С. Зяноўка // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития : материалы IV Междунар. науч.-практ. конф., Минск, 19–20 марта 2020 г. / редкол.: О.Г. Прохоренко (отв. ред.) [и др.]. – Минск : БГУ, 2020. – С. 472-477.
3. Zianouka, Y. Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module / Y. Zianouka, Y. Hetsevich, D. Latyshevich, Z. Dzenisiuk // 15th International Conference, NooJ 2021 / France ; ed. Magali Bigey, Annabel Richeton, Max Silberztein, Izabella Thomas. – Besançon : Springer, Cham, 2022. – P. 231-242.
4. Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. – 2022. Рэжым доступу: http://ssrlab.by/. – Дата доступу: 11.07.2017.
5. Інтанацыйны працэсар [Электронны рэсурс]. – 2023. Рэжым доступу : https://corpus.by/IntonationalProcessor/?lang=be. – Дата доступу : 29.01.2023.
6. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by // [Электронны рэсурс]. – 2022. Рэжым доступу : http://corpus.by/ . – Дата доступу : 12.01.2023.

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ЧАСТОТНОГО СЛОВАРЯ

AUTOMATIC CONSTRUCTION OF FREQUENCY DICTIONARY

Ю.Ю. Красовская

J.J. Krasowskaja

Белорусский государственный университет,
Минск, Беларусь, *hisradaar@gmail.com*
Belarusian State University,
Minsk, Belarus, *hisradaar@gmail.com*

Частотный словарь включает в себя слова или словоформы какого-либо текста или корпуса текстов и частоту их встречаемости в исследуемом корпусе текстов. Составление частотного словаря посредством программного обеспечения позволяет автоматизировать процесс. Автоматизация с применением лемматизации позволит упростить долгий про-

цесс ручного построения частотного словаря для корпуса текстов некоторой предметной области.

Ключевые слова: цифровые технологии; частотный словарь; лемматизация; частотность; корпус текстов.

The frequency dictionary includes words or word forms of any text or text corpus and the frequency of their occurrence in the language of the text corpus. Compiling a frequency dictionary using software allows one to automate the process. Automation using lemmatization is a rather simplified process of manually compiling a frequency dictionary for texts of one subject area.

Keywords: digital technologies; frequency list; lemmatization; frequency; texts corpus.

Частотный словарь – вид словаря, одноязычного или многоязычного, в котором лексические единицы характеризуются с точки зрения степени их употребительности в совокупности текстов, представительных либо для языка в целом, либо для отдельного функционального стиля, либо для одного автора. Это модель особым образом преобразованного текста, модель рангового распределения частот употребления языковых единиц в тексте. Для создания частотных словарей используются статистические закономерности.

Частотный анализ основывается на выдвигаемом предположении о существовании нетривиального статистического распределения отдельных символов и их последовательностей, который даже при замене символов будет сохранен как в процессе шифрования, так и в процессе дешифровки. Такому анализу может быть подвергнут как открытый текст, так и зашифрованный. Благодаря методу частотного анализа русским математиком А.А. Марковым в результате исследования романа А.С. Пушкина «Евгений Онегин» было доказано, что появление букв в открытом тексте нельзя считать независимым друг от друга. Также А.А. Марковым отмечена закономерность чередования гласных и согласных букв в русском языке [4, с. 159]. Благодаря частотному анализу можно определить наиболее продуктивные способы словообразования. В русском языке таковым выступает аффиксация, в то время как в английском – словосложение и конверсия.

Частотные словари содержат статистические данные, широко используемые для решения различных языковых задач, таких как: определение динамичных средств словообразования и анализа языка, усовершенствование вопросов орфографии и графики, связанных с учётом статистических сведений о словарном составе (необходимо принимать во внимание возможные характеристики сочетаний графем, которые реализованы в лингвистических единицах типами буквосочетаний), применение практической транслитерации и транскрипции, решение

вопросов автоматизации печатного дела, автоматического чтения и распознавания буквенного текста.

Составление частного словаря вручную – процесс зачастую рутинный и монотонный, однако совсем не сложный. Для оптимизации этой задачи было решено автоматизировать процесс настолько, насколько это возможно.

Для написания программного обеспечения для составления частотного словаря был выбран язык Python. Данный язык был выбран ввиду нескольких причин:

1. Python является интерпретируемым языком программирования, который не компилируется. То есть до запуска он представляет собой обычный текстовый файл. Соответственно, это решает проблему кроссплатформенности.
2. Сам язык логичен и хорошо спроектирован. Кода в нём меньше, чем при использовании других языков программирования, поэтому разработка осуществляется быстрее.
3. Для языка программирования Python имеется ряд библиотек, в том числе встроенных, которые значительно облегчают выполнение вычислений и исследований.
4. Python удобен для обработки текста на естественном языке (Natural Language Processing, NLP).

Впрочем, считается, что выполнение программ на языке Python недостаточно быстрое. Однако программа, написанная для составления частотного словаря, занимает крайне мало памяти, посему данное замечание становится неактуальным.

Первым делом импортируется ряд библиотек, необходимых для работы программы. Библиотека NLTK (Natural Language Toolkit), предназначенная для обработки текстов на естественном языке, и модуль `pymorphy2`, служащий морфологическим анализатором, предварительно загружены отдельно.

Листинг 1

Фрагмент кода программы, в которой пользователю самому предлагается выбрать файл для частотного анализа

```
Tk().withdraw()
filename = askopenfilename()
print(filename)

f = open(filename, "r", encoding="utf-8")
text_orig = f.read()
type(text_orig)
len(text_orig)
```

Далее при помощи модуля tkinter, работающего с графическим интерфейсом, пользователю предлагается выбрать файл, который будет подвергнут частотному анализу. Исследованию может быть подвергнут любой текст на русском языке, сохраненный в формате .txt (листинг 1).

Листинг 2

Текст программного кода, ответственного за приведение исходного текста в вид, пригодный для последующей обработки

```
text_orig = text_orig.lower()
    string.punctuation
    type(string.punctuation)
    spec_chars = string.punctuation + '\n\xa0«»\t—–…'
    text_orig = "".join([ch for ch in text_orig if ch not
    in spec_chars])
    import re
    text_orig = re.sub('\n', '', text_orig)

    def remove_chars_from_text(text_orig, chars):
        return "".join([ch for ch in text_orig if ch not
        in chars])

    text_orig = remove_chars_from_text(text_orig,
    spec_chars)
    text_orig = remove_chars_from_text(text_orig,
    string.digits
```

Затем текст приводится в вид, подходящий для обработки: прописные буквы заменяются на строчные, удаляются знаки препинания (листинг 2).

При помощи модуля pymorphy2 производится лемматизация слов, то есть приведение их к начальной форме (листинг 3).

Листинг 3

Фрагмент кода, осуществляющий лемматизацию

```
    def lemmatize(text_orig):
        words = text_orig.split() #
        res = list()
        for word in words:
            p = morph.parse(word)[0]
            res.append(p.normal_form)
        return res
    mylist=lemmatize(text_orig)
```

После проведения лемматизации слова в начальной форме записываются в отдельный файл. Подсчитывается их общее количество (листинг 4).

Листинг 4

Часть кода, отвечающая за запись обработанного текста в отдельный файл и подсчёт общего количества слов

```
d = open('text_2.txt', "w", encoding="utf-8")
mylist = map (lambda x: x + ' ', mylist)
d.writelines (mylist)
d.close()

g = open('text_2.txt', "r", encoding="utf-8")
text=g.read()
type(text)
len(text)

from nltk import word_tokenize
text_tokens = word_tokenize(text)
print('Общее количество слов в
тексте:',type(text_tokens), len(text_tokens))
```

При помощи модуля `nltk` подсчитывается частотность слов. Результаты, отсортированные по убыванию частот, записываются в отдельный файл slovar.txt построчно. Это и является частотным списком (листинг 5).

Листинг 5

Код, подсчитывающий частоты и записывающий результаты в отдельный файл

```
from nltk.probability import FreqDist
fdist = FreqDist(text)

print('Самые частотные слова:',fdist.most_common(25))

sw = fdist.most_common()

with open("slovar.txt", "w+") as slovar:
    for  s in sw:
        slovar.write(' '.join(str(sw)  for  sw  in  s)  +
'\n')
```

Финальным этапом данного исследования является создание частотного словаря на основе списка, полученного прежде. Для удобства полученный список стоит перенести в Excel-таблицу следующим образом: Данные – Получить данные – Из файла – Из текстового файла.

Конечный вариант частотного словаря представлен в формате xlsx-таблицы, которая может быть отсортирована как по алфавиту, так и по убыванию или возрастанию частот.

Программное обеспечение написано таким образом, что способно провести частотный анализ любого текстового файла в формате txt. Таким образом, код может быть применен для дальнейших исследований в области частотного анализа.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Алексеев, П. М. Квантитативная типология текста: дис. … докт. филол. наук / М.П. Алексеев – Л., 1977. – 337 с.
2. Арапов, М.В. Квантитативная лингвистика / М.В. Арапов. – М.: Наука, 1988. – 183 с.
3. Долинский, В. А. Квантитативная лингвистика в исследовании текста / В.А. Долинский – Алфавит: Строение повествовательного текста. Синтагматика. Парадигматика. Смоленск: СГПУ, 2004. – С. 283–324
4. Марков А.А. Пример статистическаго изследованія над текстом «Евгенія Онегина», иллюстрирующій связь испытаній в цепь» / А.А.Марков – Известия Императорской Академии Наук. VI серія. Т. 7. Вып. 3. 1913. – С. 153-162.

ОСОБЕННОСТИ МЕЖЛИЧНОСТНОГО ОБЩЕНИЯ В ИНТЕРНЕТ-ПРОСТРАНСТВЕ

PECULIARITIES OF INTERPERSONAL COMMUNICATION IN THE INTERNET SPACE

Н.В. Лучкина[1)], С.А. Мирзоева[2)], И.Ю. Проценко[3)]

N.V. Luchkina[1)], S.A. Mirzoeva[2)], I.U. Protsenko[3)]

[1)]Ростовский государственный медицинский университет,
Ростов-на-Дону, Россия, *luchkina7@mail.ru*
[2)]Ростовский государственный медицинский университет,
Ростов-на-Дону, Россия, *samirzoeva@yandex.ru*
[3)]Ростовский государственный медицинский университет,
Ростов-на-Дону, Россия, *ir.protsenko@yandex.ru*
[1)]Rostov State Medical University
Rostov-on-Don, Russia, *luchkina7@mail.ru*
[2)]Rostov State Medical University
Rostov-on-Don, Russia, *samirzoeva@yandex.ru*
[3)]Rostov State Medical University
Rostov-on-Don, Russia, *ir.protsenko@yandex.ru*

Быстрое распространение межличностного виртуального общения породило необходимость комплексного изучения данного явления, чреватого разного рода рисками. Обобщая эмпирический опыт коммуникации в интернет-пространстве, можно прийти к заключению, что в противоборство с человеком вступает его виртуальная сущность, его цифровой двойник. Исследование коммуникации человека в интернет-среде нуждается в новой научной парадигме, объединяющей данные социолингвистики, психолингвистики, лингвокультурологии, лингвистики текста, теории коммуникации, лингвоэкологии и пр.

Ключевые слова: межличностное общение; интернет-пространство; риск.