УДК 004.912

Р.О. БАГРІЙ
Хмельницький національний університет

СИСТЕМА ПРОГНОЗУВАННЯ СЛІВ ДЛЯ АЛЬТЕРНАТИВНОЇ КОМУНІКАЦІЇ

У роботі проведено дослідження інтелектуалізації введення інформації за допомогою системи прискореного введення тексту в цифрові пристрої з метою побудови моделі корпусу розмовної української мови та системи набору тексту, яка базується на цій моделі. Така система використовує меншу кількість команд для введення букв та прогнозує варіанти слів, базуючись на даних корпусу слів та словосполучень для спілкування. Експериментально показано для побудованого корпусу достатньо ефективне введення тексту за допомогою 4-х та 6-и клавіш-команд.

Ключові слова: альтернативна комунікація, формування корпусу слів, N-грами, прогнозування.

R.O. BAGRIY
Khmelnytsky National University

WORD PREDICTION SYSTEM FOR ALTERNATIVE COMMUNICATION

This paper investigates the intellectualization of text input using a system for accelerated input of texts into digital devices with a view to constructing a model of a corpus of the Ukrainian spoken language and a text typing system based on this model. Such a system uses a smaller number of commands to input letters and predicts variants of words on the basis of the corpus of words and word combinations for communication. The described procedure for collecting texts containing such dialogues and an algorithm of actions for the formation of the educational corpus of words. The described in detail a statistical language model, which is proposed to be applied to the prediction of words. It is experimentally shown that the input of texts using four and six command keys is rather efficient for the constructed corpus.

Keywords: alternative communication, formation of a corpus of words, N-grams, prediction.

Вступ. Концепція додаткової та альтернативної комунікації (Augmentative and Alternative Communication – AAC) [1] полягає в наданні людям з обмеженими можливостями спілкування засобів для комунікації з зовнішнім світом. Основна проблема будь-якого засобу AAC полягає в малій кількості доступних керуючих сигналів, які необхідно зв'язати з великим набором символів комунікації.

В сучасних умовах розвитку IT-пристроїв спостерігається тенденція їх мінімізації за розмірами, включаючи перспективні розробки пристроїв-чіпів, що можуть бути імплантованими у тіло людини. Виходячи з цього, стає можливим застосування цих пристроїв не тільки у стаціонарних умовах, а й під час переміщень. Використання таких пристроїв людьми з обмеженими можливостями спілкування дозволяє реалізувати для них альтернативну комунікацію, яка не буде прив’язана до якогось певного місця.

Альтернативна комунікація передбачає підходи, методи та технології для перетворення даних, отриманих за допомогою AAC пристроїв в повноцінну інформацію. Таке перетворення пропонується реалізувати шляхом організації введення тексту меншою кількістю комунікаційних одиниць та створення програмного забезпечення для надання можливості спілкування.

Авторами запропоновано інформаційну технологію альтернативної комунікації [2, 3], апаратно-програмна реалізація якої повинна забезпечити комунікацію максимально можливими способами. Важливою частиною запропонованої інформаційної технології є інтелектуалізація введення інформації.

Метою роботи є дослідження інтелектуалізації введення інформації за допомогою системи прогнозування, що автоматично пропонує наступні слова, які найбільш часто зустрічаються після вже введених слів у реченні. Така система прогнозування тексту оптимізує продуктивність введення даних та зводить до мінімуму взаємодії з користувачем та базується на даних корпусу слів, що адаптований до потрібного виду комунікації.

Моделювання системи прогнозування тексту для спілкування. Алгоритми прогнозування здатні автоматично завершувати введення тексту, що дозволяє оптимізувати час введення. Для прогнозування потрібно знайти баланс між швидкістю введення і функціональністю. Чим більший словник і чим більше технологій введення використовується, тим повільніше стає час відгуку, але підвищується якість результатів.

Для дослідження і моделювання системи прогнозування тексту для спілкування пропонується наступний підхід, представлений параметрами:

1) формування множини (корпусу) слів (словосполучень) української мови, обмеженого словами для повсякденного спілкування;

2) застосування статистичної моделі мови для прогнозування наступних слів в словосполученні.

Для формування корпусу слів пропонується експертний підхід, що дозволив би за джерелами з контенту україномовних сайтів, періодичної преси, словників-розмовників тощо, підібрати слова та словосполучення, що використовуються в повсякденному спілкуванні. Для оцінки ймовірностей наступних слів у словосполученні, за цим корпусом, пропонується застосувати одну з статистичних моделей мови.

Для системи прогнозування тексту необхідно сформувати навчальний корпус адаптований до потрібного виду комунікації. Вибір навчального корпусу відіграє важливу роль в розробці будь-якої системи прогнозування тексту. Статистичні мовні моделі необхідно навчати за великим набором текстів для отримання достовірних оцінок ймовірності. Також, чим більше навчальний корпус схожий на даний вид

комунікації, тим більш точними будуть оцінки ймовірності.

Для української мови існує кілька відомих корпусів. Це Національний корпус української мови (НКУМ) [4], який зберігає тексти письмового та усного (розмовного) варіантів національної мови. Також Українським мовно-інформаційним фондом (УМІФ) створений Український національний лінгвістичний корпус – УНЛК [5], який налічує понад 43 млн. слововживань. Також створено корпус української мови на лінгвістичному порталі mova.info [6], який надає користувачам можливість пошуку слів з кількох підкорпусів, а саме художня проза, наукові, поетичні, фольклорні, законодавчі та публіцистичні тексти.

Існуючі мовні корпуси в більшості зберігають тільки письмові тексти різних стилів, а підкорпус розмовного стилю часто відображає тільки монологи чи стенограми засідань. Це пов'язано з тим, що збори текстів розмовного стилю вимагають як великої кількості інтерв'юерів і респондентів для запису аудіоматеріалу, так і трудомісткості розшифровки зібраних записів.

Інший варіант – це створення корпусу текстів мови спілкування з контенту розташованого в мережі Інтернет, який використовують при комунікації (листуванні) в соціальних мережах, чатах, по електронній пошті, на форумах тощо. Цей спосіб збору текстів досить швидкий, але має кілька недоліків. По-перше, під час Інтернет-комунікації співрозмовники не бачать один одного, по-друге, в ході написання часто використовуються скорочення, які в реальному спілкування не зустрічаються, по-третє, різний рівень знань учасників розмови призводить до великої кількості помилок у текстах та використання запозичених слів з інших мов. Ці недоліки вимагають значної ручної перевірки текстів, що є досить трудомістким для великих обсягів даних.

Таким чином, використання існуючих корпусів для реалізації комунікації між людьми неможливо і виникає необхідність сформувати обмежений підкорпус розмовної української мови.

Для розв’язання цієї задачі запропоновано використати діалоги на побутові теми, що використовуються в словниках-розмовниках для вивчення іноземних мов. Такі діалоги моделюють розмови між людьми, які бачать один одного, охоплюють найбільш можливі побутові ситуації та використовують обмежений набір слів і фраз. Для системи комунікації людей з обмеженими можливостями подібні властивості діалогу підходять як не можна краще.

Для збору текстів пропонується використати розмовники, підручники, посібники та інші Інтернет-ресурси, які містять розмовні діалоги. Отримані діалоги, для подальшого формування моделі, пропонується розбити на базові та тестові набори, з метою визначення достатньої наповненості корпусу для задачі прогнозування слів і словосполучень при введенні тексту.

Для побудови статистичної моделі мови необхідно обробити отриманий набір текстів. На сьогоднішній день існує безліч способів розбиття електронного тексту на окремо значущі одиниці для подальшої комп'ютерної обробки цих одиниць.

Для розв’язання поставленої задачі пропонується застосувати алгоритми обробки природної мови (Natural Language Processing, NLP) [7]. В цих алгоритмах прийнято використати такі поняття: типи (types) – різні слова (послідовності слів) в тексті, токени (tokens) – все слова (послідовності слів) в тексті. Токенізація (tokenization) – процес розбиття даного тексту на одиниці, що називаються токенами.

Одним із способів формалізації в комп'ютерних науках є регулярні вирази [8] – мова для формування шаблонів, що використовуються для текстового пошуку рядків. Формально, регулярний вираз є алгебраїчним позначенням наборів рядків. Регулярні вирази полегшують обробку великих обсягів текстових даних за допомогою невеликої кількості шаблонів для пошуку. Регулярна функція для вираження пошуку, що застосовується до корпусу, повертає всі фрагменти тексту, що відповідають шаблону.

Всі отримані фрагменти тексту пропонується розбити на речення та видалити всі символи, що не відносяться до українського алфавіту. Далі, отримані текстові набори, пропонується:

- токенізувати – відокремити осмислені елементи (слова, фрази, символи) – токени;
- перевести всі букви в нижній регістр та видалити знаки пунктуації.

Таким чином, при підготовці статистичної мовної моделі буде враховуватися тільки набір символів, що відноситься до української мови.

Алгоритм токенізації отриманого набору текстів пропонується реалізувати за допомогою набору регулярних та лямбда виразів. Для видалення символів, що не відносяться до українського алфавіту пропонується застосувати наступні основні шаблони (табл. 1).

Таблиця 1

Шаблони регулярного виразу

Шаблон регулярного виразу	Значення
"[0-9]"	видалення цифр
"[A-Z,a-z]"	видалення букв англійського алфавіту
"[\"\\[\\]\',/=:;&(){}%*|<>\\+_@$#©«»]"	видалення спец. символів

Наступний крок полягає в обробці та розбитті отриманих текстів за допомогою лямбда-виразів (табл. 2).

Таблиця 2

Лямбда-вирази обробки тексту

Функції роботи зі строками	Значення
.Select(x => x.ToLower())	переведення в нижній регістр
.Select(x => x.Trim())	обрізка зайвих пробілів
.Select(x => x.Split())	розбиття речень на слова
.ToList()	формування списку речень

Для задач прогнозування тексту застосовуються статистичні моделі мови. В області комунікації ААС для прогнозування послідовності слів частіше всього використовується N-грамна модель. За допомогою інструментарію text2ngram [9] із списку речень генеруються N-грами у вигляді таблиць бази даних, що мають структуру зображену на рис. 1.

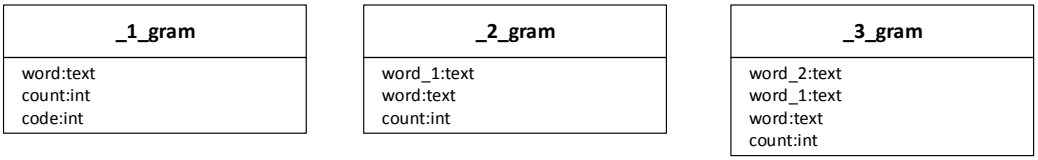


Рис. 1. Структура бази даних для збереження N-грам

Таблиці містять поля для збереження слів (word, word_1, word_2) та кількості їх попадань (count) в навчальних текстових наборах. Для таблиці юніграм також зберігається код слова (code), що обчислюється залежно від поточної розкладки символів на клавішах віртуальної клавіатури [3].

Метою побудови N-грам моделей є визначення ймовірності використання заданого слова чи фрази (словосполучення). Найбільш часто використовуються N-грамні моделі де N дорівнює 2 або 3, і називається вони відповідно біграмами і триграмами. Юніграми є виродженим випадком, коли припускається, що ймовірність кожного слова повністю незалежна від минулої історії. Чим більше N, тим краще точність прогнозування, але і тим більше використовується системних ресурсів. На рисунку 2 показано приклад розбиття набору текстів на юніграми, біграми та триграми.

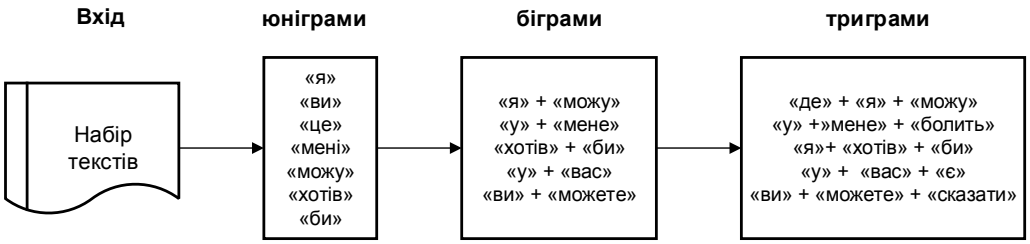


Рис. 2. Розбиття тексту на N-грами

N-грамні моделі визначають ймовірності використання заданої фрази (словосполучення), формально, як ймовірність виникнення послідовності слів у деякому корпусі (наборі текстів). Для оцінки цих ймовірностей використовують самий простий та найбільш інтуїтивний спосіб оцінки ймовірностей – метод максимальної подібності MLE (maximum likelihood estimation) [7].

Оцінка максимальної подібності полягає у визначенні параметрів, які максимізують ймовірність цієї подібності для заданих слів. Таким чином, MLE оцінка параметрів моделі N-грами може бути отримана як нормалізована кількість від корпусу, що використовується. Корпус є набір текстів, який статистично репрезентативний для моделювання мови. Наприклад, можна оцінити біграм-ймовірність слова w_n, враховуючи попереднє слово w_{n-1}, обраховуючи входження біграм $C(w_n, w_{n-1})$ та нормуючи сумою всіх біграм, які містять перше слово w_n:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n)}{\sum_w C(w_{n-1}, w)}. \quad (1)$$

Так як кількість всіх біграм, що починаються зі слова w_{n-1} повинна дорівнювати кількості юніграм для цього слова w_{n-1}, (1) спрощується:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})}. \quad (2)$$

В загальному випадку N-грам-моделі формула для оцінки параметрів MLE буде такою:

$$P\left(w_n \middle| w_{n-N+1}^{n-1}\right) = \frac{C\left(w_{n-N+1}^{n-1}, w_n\right)}{C\left(w_{n-N+1}^{n-1}\right)}. \quad (3)$$

Моделі N-грам в основному використовуються для прогнозування слів, тому що вони є ефективними і простими у використанні. До того ж статистика частот, яка використовується для розрахунку MLE оцінки, може бути отримана безпосередньо з текстів без особливих зусиль.

Класичний підхід до прогнозування по введеним буквам ускладнюється тим, що при використанні запропонованої технології прискореного введення на вхід подаються слова, зашифровані кодом, відповідним обраному порядку слідування букв на клавішах віртуальної клавіатури [3]. В цьому випадку одному коду може відповідати велика кількість слів-кандидатів. Для подолання цієї проблеми запропоновано використати статистичну модель мови Katz's backoff [7], що базується на корпусі слів.

Основною ідеєю Katz's backoff моделі є оцінка умовної ймовірності слова шляхом «відступу» до *n*-грами меншого порядку у випадку, коли використання більш високого порядку не знайдена у навчальному корпусі. Таким чином, модель з самої повною інформацією використовується для забезпечення найкращих результатів.

Зокрема, для найвищого порядку пропонується використати триграмну модель, і відступи виконувати рекурсивно до біграм та юніграм. Формула для триграм-моделі має наступний вигляд:

$$P_{BO}(w_n \mid w_{n-2}, w_{n-1}) = \begin{cases} P^*(w_n \mid w_{n-2}, w_{n-1}) & якщо\ C(w_{n-2}, w_{n-1}, w_n) > 0 \\ \alpha(w_{n-2}, w_{n-1}) P_{BO}(w_n \mid w_{n-1}) & якщо\ C(w_{n-2}, w_{n-1}) > 0 \\ P_{BO}(w_n \mid w_{n-1}) & інакше \end{cases}, \quad (4)$$

де

$$P_{BO}(w_n \mid w_{n-1}) = \begin{cases} P^*(w_n \mid w_{n-1}) & якщо\ C(w_{n-1}, w_n) > 0 \\ \alpha(w_{n-1}) P^*(w_n) & інакше \end{cases}, \quad (5)$$

$$P^*(w_n \mid w_{n-2}, w_{n-1}) = \frac{C^*(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}. \quad (6)$$

Тут P^* – згладжена ймовірність, обчислена за оцінкою Good-Turing [7] і α – коефіцієнти відступу. Коефіцієнт α пропонується прийняти рівним 0.4, як такий, що застосовується в алгоритмі Stupid backoff [7] та показує наближену до інших методів якість прогнозування та спрощує обчислення.

Ймовірність слів, що прогнозуються, обчислюються як оцінка моделі мови для поточного контексту. Згідно з рівнянням 4, якщо речення, що вводиться, складається з трьох або більше слів, для прогнозування використовується модель для триграм. Для розрахунку триграм-ймовірності використовуються два попередніх слова. Наприклад, для фрази "у мене болить" ймовірність для останнього слова буде записана як $P(болить \mid у\ мене)$.

Для словосполучення з двох слів або у випадку, коли словосполучення не знайдено серед триграм, використовується модель для біграм (5). У цьому випадку для розрахунку ймовірності береться тільки попереднє слово – $P(болить \mid мене)$.

Модель юніграм використовується у випадку, коли прогнозоване слово є першим в реченні (тобто відсутні попередні слова) або словосполучення не знайдено серед біграм або триграм. Імовірність слова буде дорівнювати його нормалізованій частоті – $P(болить)$.

Після оцінки ймовірностей всіх слів-кандидатів, прогнозовані слова впорядковуються, слова з найвищими ймовірностями відображаються користувачу для остаточного вибору.

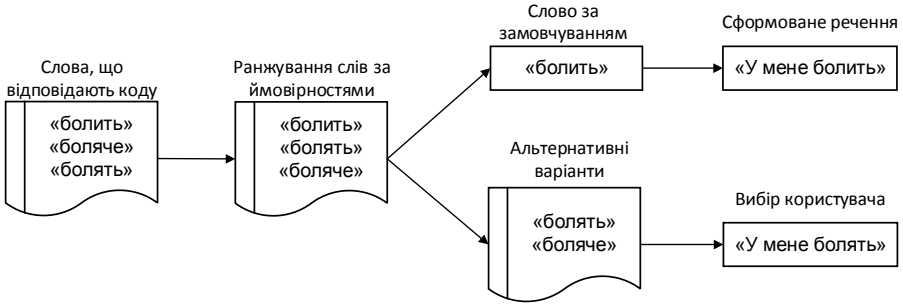


Рис. 3. Алгоритм формування речення

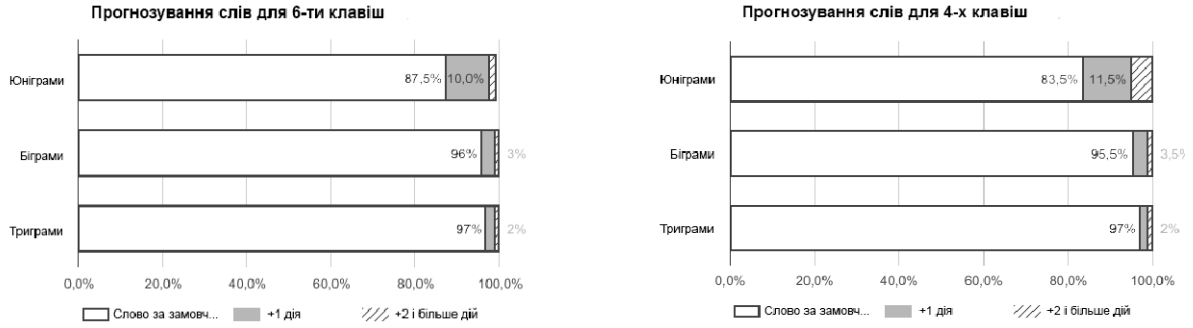
Для покращення процесу введення тексту необхідно, щоб якомога більше слів, що вводяться, прогнозувалися як слова "за замовчуванням", тобто відповідали вимогам користувача. Тоді це не

потребуватиме додаткових дій від нього. Якщо потрібне слово знаходиться нижче в списку прогнозованих слів, то його вибір вимагає одну додаткову дію за кожну позицію у списку (рис. 3).

Слід зазначити, що тільки слова, які є в словнику можуть бути прогнозовані. Якщо слово відсутнє в словнику, то для його введення необхідно затратити набагато більше часу, так як його потрібно ввести повністю по буквах.

Для визначення якості прогнозування було проведено ряд експериментів для кожної з моделей. Тестування було проведено для 6-и клавіш клавіатурного порядку слідування букв і 4-х клавіш частотної послідовності слідування букв з урахуванням голосних / приголосних [3]. На момент проведення тестування приблизна кількість типів юніграм дорівнювала 15000, біграм – 65000, триграм – 93000.

Початкове тестування полягало в прогнозуванні того ж тексту, на основі якого були сформовані N-грам-моделі. Загальна кількість фраз для прогнозування склала понад 18000.



а) **б)**

Рис. 4. Прогнозування слів за різними моделями: а) для 6-и клавіш розподілу; б) для 6-и клавіш розподілу

Юніграм-модель показала низьку якість прогнозування – 87.5% для 6-и клавіш (рис. 4, а) і 83,5% для 4 клавіш (рис. 4, б) розподілу. В інших випадках необхідно застосовувати одну додаткову дію для вибору слова, а іноді дві і більше таких дій. Для біграм-моделі цей показник значно більше і практично однаковий для обох випадків – 95.5–96%. Цього цілком достатньо для комфортної роботи. Триграм-модель ще трохи покращує прогнозування, і якість майже не залежить від кількості клавіш віртуальної клавіатури.

Наступні тестування полягали в прогнозуванні довільного тексту. Обсяг тексту склав близько 3000 фраз з загальною кількістю понад 15000 слів. Для 6-и блоків вірно було прогнозовано 90% слів, відомих N-грам-моделі. Для 4-х блоків, якість прогнозування склала 89%. Зниження точності прогнозування можна пояснити тим, що в довільному тексті завжди присутній певний відсоток нових типів біграм та триграм. Також, у разі відсутності слова в N-грам-моделі, наступне за ним слово може бути прогнозовано тільки за юніграм-моделлю.

Висновки. У статті для реалізації інформаційної технології альтернативних підходів до спілкування, запропонована система прогнозування, що автоматично пропонує наступні слова, які найбільш часто зустрічаються після вже введених слів у реченні. Надана модель комунікації для повсякденного спілкування, що передбачає використання розмовних діалогів на побутові теми. Описаний принцип збору текстів, що містять такі діалоги та алгоритм дій для формування навчального корпусу слів. Детально описана статистична модель мови, що пропонується застосувати для прогнозування слів, що дозволить прогнозувати слова, які є в словнику, з ймовірністю 89–90% для довільного тексту.

Подальші дослідження спрямовані на реалізацію запропонованого способу альтернативного спілкування за допомогою стандартних гаджетів (планшети, телефони) з метою використання для організації діалогів з людьми, у яких тимчасово відсутній або ускладнений канал основний вербальної комунікації. Оскільки запропонований підхід є загальним, то будуть проведені дослідження можливості його використання для інших мов.

Література

1. Augmentative and Alternative Communication (AAC) [Електронний ресурс]. – Режим доступу : http://www.asha.org/public/speech/disorders/AAC/.

2. Кривонос Ю.Г. Новые средства альтернативной коммуникации для людей с ограниченными возможностями / Ю.Г. Кривонос, Ю.В. Крак, А.В. Бармак, Р.А. Багрий // Кибернетика и системный анализ. – 2016. – Том 52, № 5. – С. 3–13.

3. Крак Ю.В. Система ввода текста для альтернативной коммуникации / Ю.В. Крак, А.В. Бармак, Р.А. Багрий // Проблемы управления и информатики. – 2017. – № 3. – С. 5–13.

4. Демська-Кульчицька О. Національний корпус української мови: концептуальний аспект / О. Демська-Кульчицька // Лексикографічний бюлетень : зб. наук. пр. – К. : Ін-т української мови НАН України, 2006. – Вип. 13. – С. 5–9.

5. Сидорчук Н. М. Организация данных и функциональная структура лексикографической системы «Украинский национальный лингвистический корпус» / Н. М. Сидорчук // Математичні машини і системи. –

2006. – № 2. – С. 126–135.

6. Дарчук Н. Дослідницький корпус української мови: основні засади і перспективи / Н. Дарчук // Вісник Київського національного університету імені Т.Г. Шевченка. – 2010. – № 21. – С. 45–49.

7. Jurafsky D. Speech and Language Processing / D. Jurafsky, J. H. Martin. – 2nd ed. – New Jersey : Prentice Hall, 2008. – 1024 p.

8. Bird S., Klein E., Loper E. Natural Language Processing with Python. – O'Reilly, 2009. – 504 p.

9. Xueqiang Lu. Statistical Substring Reduction in Linear Time / Xueqiang Lu, Le Zhang and Junfeng Hu. – Natural Language Processing. – IJCNLP, 2004. – P. 320–327.

References

1. Augmentative and Alternative Communication (AAC) [Elektronnyi resurs]. – Rezhym dostupu : http://www.asha.org/public/speech/disorders/AAC/.

2. Kryvonos Yu.H. Novye sredstva alternatyvnoi kommunykatsyy dlia liudei s ohranychennymy vozmozhnostiamy / Yu.H. Kryvonos, Yu.V. Krak, A.V. Barmak, R.A. Bahryi // Kybernetyka y systemnyi analyz. – 2016. – Tom 52, # 5. – S. 3–13.

3. Krak Yu.V. Systema vvoda teksta dlia alternatyvnoi kommunykatsyy / Yu.V. Krak, A.V. Barmak, R.A. Bahryi // Problemy upravlenyia y ynformatyky. – 2017. – # 3. – S. 5–13.

4. Demska-Kulchytska O. Natsionalnyi korpus ukrainskoi movy: kontseptualnyi aspekt / O. Demska-Kulchytska // Leksykohrafichnyi biuleten : zb. nauk. pr. – K. : In-t ukrainskoi movy NAN Ukrainy, 2006. – Vyp. 13. – S. 5–9.

5. Sydorchuk N. M. Orhanyzatsyia dannykh y funktsyonalnaia struktura leksykohrafycheskoi systemy «Ukraynskyi natsyonalnyi lynhvystycheskyi korpus» / N. M. Sydorchuk // Matematychni mashyny i systemy. – 2006. – # 2. – S. 126–135.

6. Darchuk N. Doslidnytskyi korpus ukrainskoi movy: osnovni zasady i perspektyvy / N. Darchuk // Visnyk Kyivskoho natsionalnoho universytetu imeni T.H. Shevchenka. – 2010. – # 21. – S. 45–49.

7. Jurafsky D. Speech and Language Processing / D. Jurafsky, J. H. Martin. – 2nd ed. – New Jersey : Prentice Hall, 2008. – 1024 p.

8. Bird S., Klein E., Loper E. Natural Language Processing with Python. – OReilly, 2009. – 504 p.

9. Xueqiang Lu. Statistical Substring Reduction in Linear Time / Xueqiang Lu, Le Zhang and Junfeng Hu. – Natural Language Processing. – IJCNLP, 2004. – R. 320–327.

Отримана/Received : 2.8.2017 р. Надрукована/Printed :14.9.2017 р.
Рецензент: д.т.н., проф. Сорокатий Р.В.

Повні вимоги до оформлення рукопису http://vestnik.ho.com.ua/rules/

Рекомендовано до друку рішенням вченої ради Хмельницького національного університету, протокол протокол № 1 від 31.08.2017 р.

Підп. до друку 15.09.2017 р. Ум.друк.арк. 26,42 Обл.-вид.арк. 29,21
Формат 30х42/4, папір офсетний. Друк різографією.
Наклад 100, зам. № ____

Тиражування здійснено з оригінал-макету, виготовленого
редакцією журналу "Вісник Хмельницького національного університету"
редакційно-видавничим центром Хмельницького національного університету
29016, м. Хмельницький, вул. Інститутська, 7/1. тел (0382) 72-83-63