Фіксуються також однослівні синтагми, що є самодостатніми у смисловому відношенні, а це, безумовно, і сприяє трансформації прийменниково-іменникових сполучень, які у нашому випадку входять у структурні моделі, побудовані за схемами ***в*** + *іменник чоловічого роду однини у місцевому відмінку*, ***без*** + *іменник множини у родовому відмінку*, ***без*** + *іменник чоловічого роду однини у родовому відмінку*, ***до*** + *іменник чоловічого роду однини у родовому відмінку*, у прислівниковий клас одиниць: ***в захваті, без жартів, без ладу, до вподоби***: *— О, яка великодушність, графе! Який широкий жест! Я, далебі,* ***в захваті...*** (Аліков Ю., Капустян В.); *Тому що Львів — магічне місто. Окей,* ***без жартів*** (В. Винниченко); *Біля одної праворуч, спертий на лікті, напів сидить, напів лежить Довбуш. Посередині на мураві розтаборилась як-небудь,* ***без ладу,*** *ватага опришків* (Б. Антонич); *Але їм це, здається,* ***до вподоби*** (А. Азімов).

Відзначимо, що нові смисли у прийменниково-іменникових конструкцій виникають у тих випадках, коли вони здатні утворювати окремі синтагми, адже смислове навантаження, зокрема у двослівних синтагмах, переважно падає на другу частину, якою є повнозначне слово, семантика котрого і набуває певних зрушень. Звичайно, такі значеннєви зсуви не відбуваються раптово, потрібен певний час. Це, наприклад, засвідчують дані дослідження, здійснені М. М. Пещак, яка зафіксувала, що двослівні прийменниково-іменникові синтагми ще з XIV ст. об'єднували в собі ті сполучення прийменника й іменника, з яких у сучасній українській мові утворилися прислівникові еквіваленти, зокрема прийменника ***безо*** та іменника в родовому відмінку, прийменників ***на, въ (во)*** та іменників у знахідному відмінку тощо [2, с. 89].

Проте, щоб встановити, як впливає на формування прислівникового статусу синтагматичне оточення або позиція прислівникового еквівалента у синтагмі, необхідним у подальших дослідженнях буде з'ясування й особливостей взаємодії усіх видів зв'язку її елементів, якими є граматичні, структурні, інтонаційні та смислові. Цьому, беззаперечно, сприяла б синтагматична розмітка текстів корпусу, яка б дозволила не вручну, а автоматично виявити усі наявні синтагми з еквівалентами слова. На сьогодні, на жаль, синтаксична розмітка корпусів є поки що екзотикою. Проте робота у цьому напрямі буде серйозним проривом у корпусних технологіях, а також встановленні й оформленні нових лінгвістичних фактів. Оскільки досвід показує, що найефективнішим способом представлення лінгвістичних знань є лексикографічні системи у вигляді словників, то при просуванні у цьому напрямі, зокрема і створенні словника синтагм, варто очікувати значного прориву у галузі комп'ютерної лексикографії.

Література

1. Лучик А. А. Прислівникові еквіваленти слова в українській мові / А. А. Лучик. — Katowice: Wydwo US, 2009. — 169 с.; Лучик А. А. Еквіваленти слова як предмет мовознавчих досліджень / А. А. Лучик. // Вісник Донецького університету. Серія Б.: Гуманітарні науки. — Донецьк: Донеччина, 2001. — С. 36–42.
2. Пещак М. М. Стиль ділових документів XIУ ст. (структура тексту) / М. М. Пещак. — К.: Наукова думка, 1979. — 268 с.
3. Рогожникова Р. П. Словарь эквивалентов слова: наречные, служебные, модальные единства / Р. П. Рогожникова. — М.: Рус яз., 1991. — 254 с.
4. Соссюр Ф. де. Курс загальної лінгвістики / Ф. де Соссюр. — К.: «Основи», 1998. — 324 с..
5. Широков В. А. Комп'ютерна лексикографія / В. А. Широков. — К.: Наукова думка, 2011. — 352 с.
6. Українська мова. Енциклопедія / Редкол.: Русанівський В. М., Тараненко О. О., Зяблюк М. П. та ін. — К.: Вид-во «Укр. енцикл.» ім. М. П. Бажана, 2007. — 856 с.

Стратегії й методи вдосконалення автоматичного морфологічного анотування Корпусу української мови

Маргарита Лангенбах

к. філол. н., асистент кафедри української мови та прикладної лінгвістики, Інститут філології Київського національного університету імені Тараса Шевченка, Україна, E-mail:labacompli@gmail.com

The article reviews the typical problems of the dictionary-based part-of- speech tagging. The main attention is focused on the non-recognized words. The experiment is based on the textual samples derived from the Corpus of the Ukrainian Language. The examples were classified by the specific features. The article suggests the strategy of increasing the efficiency of the dictionary-based part-of- speech tagger.

Ключові слова — автоматичний морфологічний аналіз, машинна морфологія, АГАТ, графемний аналіз, словниковий морфологічний аналіз.

Використання автоматичних морфологічних аналізаторів природних мов вже досить поширене у світовій практиці, їх розробка спирається на серйозне теоретичне і практичне підґрунтя, проте для жодної мови світу досі не вдалося укласти цілком досконалу систему граматичного кодування тексту. Як зазначає К.Меннінг, попри всі успіхи в цій сфері, межу точності 97–98% поки що не подолано. Та й ці цифри, за його словами, є до певної міри ідеалізованими [11].

Отже, поліпшення якості машинного морфаналізу сьогодні лишається актуальним питанням у галузі комп'ютерної лінгвістики.

Увага до цієї теми зумовлена необхідністю розвитку комп'ютерного інструментарію для опрацювання текстів української мови. Збільшення обсягу україномовного матеріалу в мережі Інтернет, поступова переорієнтація сучасної науки (зокрема й мовознавства) на роботу з великими масивами даних, а також екстралінгвістичні (суспільно-політичні) чинники зумовлюють потребу в сучасних ефективних програмах для текстового аналізу.

Проблема якості автоматичного морфологічного аналізу у теоретичній літературі зазвичай висвітлюється в аспектах загального огляду та оцінки ефективності різних стратегій [6], [4], [1] або аналізу помилкової розмітки текстів, пов'язаної, зокрема, з мовною омонімією [2]. Натомість до розгляду переважно не беруться випадки ігнорування аналізаторами певних лексем. Наша стаття присвячена саме цьому аспектові автоматичного опрацювання текстів на прикладі аналізу системи комп'ютерної граматичної анотації текстів АГАТ [3].

Метою нашого дослідження було розглянути способи підвищення ефективності роботи автоматичного морфологічного аналізатора АГАТ через зменшення кількості неопрацьованих лексем.

Досягнення цієї мети передбачало виконання таких завдань:

– аналіз причин неопрацювання аналізатором певних лексем;

– класифікація типових проблемних випадків;

– добір методів та методик, що дозволили б підвищити ефективність роботи морфологічного аналізатора, розробка загальної стратегії вдосконалення системи.

Об'єктом дослідження були лексеми, з певних причин не опрацьовані автоматичним морфологічним аналізатором, **предметом** – їх особливості, що призводять до виникнення в системі помилок такого типу.

Матеріалом дослідження слугував Корпус текстів української мови, що розробляється лабораторією комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка (доступний за посиланням http://mova.info/corpus.aspx). Обсяг вибірки становив 1,2 млн. слововживань.

Розв'язання проблеми передбачало, перш за все, окреслення її меж та пошук відповідей на такі питання:

1. Наскільки ефективний на сьогодні морфологічний аналізатор? Чи залежить його ефективність від характеру аналізованих текстів (функціональний стиль, жанр, тематика)?

2. Якщо для різних текстів система демонструє різну якість опрацювання матеріалу, то якими чинниками це пояснюється?

3. Які явища є типово проблемними для автоматичного морфологічного опрацювання?

4. Які способи вирішення виявлених проблем?

Помилкам кодування омонімічних форм у науковій літературі приділено чимало уваги, проте існує ще один аспект – певний відсоток слів у ході опрацювання морфологічним модулем не отримують граматичного коду взагалі. Стандартно проблемними об'єктами для морфологічних аналізаторів є нехарактерні для мови символи, зокрема слова, написані іншою абеткою (***Curiosity*** *зробив нові приголомшливі знімки*). Крім того, морфологічні аналізатори словникового типу можуть не опрацьовувати такі групи лексики:

1. рідковживані лексеми, зокрема галузеві терміни, діалектизми та застарілі слова, власні назви (*фітофаг, віргінільний, скарамангія, Вальденфельс, Ерзурум*) тощо;
2. неологізми (у т. ч. авторські) та іншомовні слова, не зафіксовані у словнику (*знеприміт-ниться, твітер, флешмоб*);
3. словоформи в нестандартному записі:
   а) скорочення (*гром., ред., ст., МАКтє, ЛСП*);
   б) друкарські помилки;
   в) цифровий запис фрагментів словоформ (*12-ох, 11-томний, 5-кратний*);
   г)складні слова, написані через дефіс (*дворянсько-поміщицькі, два-три, бак-акумулятор*);
   д) приклади т. зв. "авторської орфографії" (переважно характерні для художніх текстів: *до-о-овгий, сссті́й, багатиє, везють*).

Неповне або некоректне морфологічне опрацювання тексту призводить до помилок на подальших етапах роботи автоматичної граматики. Так, через відсутність граматичних позначок для деяких словоформ унеможливлюється контекстне коригування омонімічних кодів у зв'язаних із ними текстових одиниць. Прогалини у морфологічній розмітці зумовлюють некоректну роботу синтаксичного аналізатора: через пропуск зв'язків у словосполученнях і реченнях будуються неповні синтаксичні схеми [5].

З наведеної інформації очевидно, що пропуск елементів під час морфологічного кодування тексту є серйозною проблемою, яка помітно впливає на якість граматичного аналізу в цілому. Вирішувати це завдання можна рухаючись у кількох напрямках.

Зокрема, однією зі стратегій є поліпшення доморфологічного аналізу. На цьому етапі роботи машина повинна, по-перше, коректно визначити межі слів (якими типово є пробіли, проте є і чимало винятків з цього правила); по-друге, виявити в тексті елементи, потенційно складні для подальших модулів

опрацювання тексту (морфологічного, синтаксичного, семантичного тощо).

Інша стратегія передбачає подолання словникової обмеженості аналізатора. Використання словників дає змогу досягти більшої ефективності опрацювання порівняно з несловниковими підходами, проте приводить до того, що слова, не зафіксовані у словнику, ігноруються системою. Для того щоб зменшити кількість неопрацьованих лексем, необхідно навчити аналізатор певним чином реагувати на позасловникові елементи.

Для аналізу основних проблем у роботі морфологічного модуля системи АГАТ було сформовано тестову вибірку розміром 1,2 млн слововживань (на матеріалі Корпусу української мови). Вибірка складалася з трьох частин, що репрезентували різні функціональні стилі (художня література, публіцистика та наукові тексти) і містили по 4000 текстових фрагментів, кожен обсягом 1000 слововживань. Після застосування до вибірки процедури автоматичного морфологічного аналізу було укладено реєстр неопрацьованих одиниць. Його обсяг становив 39841 слововживання (3,32% від загального розміру вибірки).

Кількість та склад неопрацьованих лексем для підвибірок дещо відрізнялися (див. Рис. 1).

Сумарно найменшу кількість проігнорованих програмою лексем було зафіксовано в публіцистичній підвибірці, що можна пояснити порівняно нейтральним характером лексики, яку добирають для текстів цього стилю, а також дотриманням стандартів орфографії. Найбільша кількість неопрацьованих

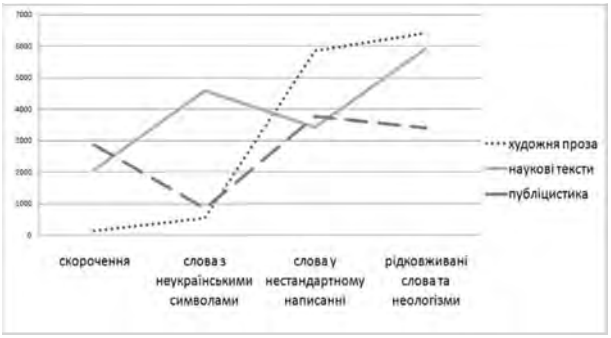

Рис. 1. Діаграма розподілу типів помилок за трьома підвибірками

одиниць виявилася в наукових текстах. Кількісний розподіл матеріалу за виділеними типами дає можливість детальніше проаналізувати, які саме одиниці спричиняють помилки в роботі програми з текстами певних стилів. Так, у наукових і публіцистичних текстах зафіксовано велику кількість скорочень, натомість у художніх текстах їх значно менше. Водночас у художній прозі набагато вищим є відсоток слів з нестандартною орфографією, тоді як для публіцистики такого типу помилки є найменш характерними. Це підтверджує нашу гіпотезу про те, що стандартизованість написання слів у публіцистичних текстах є причиною вищої якості морфологічного аналізу на матеріалах цього стилю.

На наступному етапі роботи необхідно було виявити типові ознаки, що дозволили б охарактеризувати кожну з наведених категорій помилок та розробити стратегію їхнього опрацювання. Ручне поповнення словника для більшості типів нерозпізнаних лексем не дало б вагомого результату, оскільки, за законом переваги [8: 21] співвідношення асортименту подібних одиниць у мові із частотою їхнього вживання зумовлює неефективність такого методу (це призвело б до збільшення обсягу бази за рахунок одиниць, потреба в яких незначна, а потенційна кількість таких елементів майже безмежна). Ілюстрацією цієї тези може бути список абревіатур, отриманих із нашої експериментальної вибірки: з 609 унікальних одиниць 142 одиниці не зафіксовані навіть у найсучаснішому і найдинамічнішому джерелі з цієї тематики – інтернет-словнику абревіатур www.ukrskor.info (*МЕТ* – маркетинг елітних товарів, *МІКУ* – музей історії коштовностей України, *МКЗН* – міжнародний кодекс зоологічної номенклатури, *НКРЕКП* – Національна комісія з регулювання у сферах енергетики і комунальних послуг, *ЗНС* – загони народної самооборони тощо). До того ж, словниковий підхід актуалізував би питання мовного статусу ненормативної орфографії (друкарських помилок і авторського правопису). Тому було вирішено звернутися до інших способів визначення морфологічних характеристик словоформ, зокрема графічного та графемного.

Графемний аналіз дає можливість ідентифікувати класи проблемних словоформ, що містять нехарактерні для української мови літерні послідовності. Так, технічно нескладним завданням є відсіювання словоформ, що включають неукраїнські символи. Порівняння графемного складу слова із символьним набором української мови уможливлює виявлення таких одиниць зі 100% точністю[1]. Також цілком очевидні ознаки мають деякі типи рідковживаної лексики, зокрема абревіатури: збіг трьох і більше голосних, чотирьох і більше приголосних, довжина словоформи у скороченнях типу *Гб*, *ст.*, *нм.*

За графічними особливостями можна ідентифікувати такі типи неопрацьованих одиниць: власні назви (написання з великої літери); ініціальні абревіатури (написання усього слова великими літерами); складні слова та словоформи із буквено-цифровим записом (написання через дефіс). Для подібних випадків було укладено формалізовані

[1] Проте такий підхід не дозволяє виявити іншомовні слова, що містять лише спільні з українською мовою символи.

моделі та розроблено правила їх застосування (зокрема, на базі регулярних виразів).

Проте підхід, що базується на правилах, хоча й долає до певної міри обмеження словникового аналізу, має меншу точність. Так, наприклад, умови "збіг 4-х і більше приголосних" та "довжина не більше ніж 6 символів" задовольняє, скажімо, слово *вогнлх 'вогнях'*, яке є друкарською помилкою. Також розроблені шаблони не завжди дозволяють надати точну й повну морфологічну інформацію про слово. Зокрема, слова, кваліфіковані як власні назви, та іншомовні слова (за винятком римських цифр) умовно позначаються кодом іменника (що є статистично обґрунтованим, але, тим не менш, продукує певний відсоток помилок), проте визначити їхню відмінково-часово-родову форму без застосування додаткових процедур (контекстного або імовірнісного аналізу, графемного аналізу за методом квазіфлексій тощо) неможливо. У деяких випадках правила не забезпечують і точної ідентифікації частиномовної належності. Наприклад, словам із цифровим записом надається омонімічний код "іменник–прикметник–числівник"; словоформи, записані через дефіс, програма у процесі аналізу розбиває на окремі частини і на виході приписує слову ланцюжок кодів усіх їхніх складників. Уточнення цієї морфологічної інформації має здійснюватися на наступному етапі роботи модуля – в ході контекстного аналізу [4]. Окремої уваги в цьому класі одиниць потребують складні слова типу *науково-технічний, шапка-вушанка, пекучо-беззахисне*, оскільки проблемним є загальне питання "розуміння" їх машиною – як двох різних слів чи як однієї лексеми, що суттєво впливає на побудову морфологічної парадигми таких слів та використання їх у подальшій роботі – на наступних етапах опрацювання тексту, під час пошуку за морфологічно розміченим корпусом і т.п.

Крім того, вагомим чинником, що впливає на якість опрацювання проблемних одиниць, є порядок застосування розроблених правил. Так, діагностика словоформ із буквено-цифровим записом і неукраїнськими символами доцільна на етапі доморфологічного аналізу, тоді як опрацювання власних назв і абревіатур через "шумність" алгоритмів вимагає попередньо відсіяти всі словникові одиниці, що можуть бути помилково інтерпретовані. Це, з одного боку, пришвидшить процес роботи допоміжних алгоритмів (оскільки кількість аналізованих одиниць суттєво зменшиться після застосування АМА), з іншого ж, покращить якість опрацювання (див. Рис. 2).

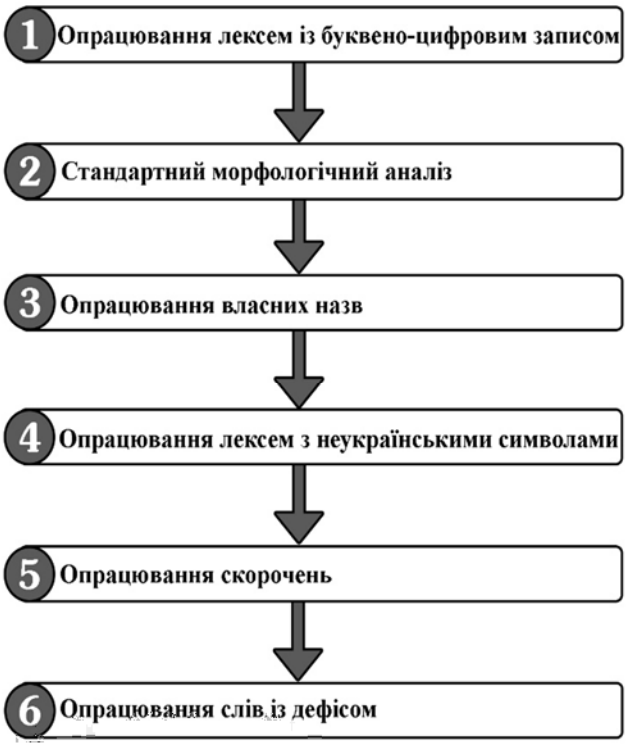

Рис. 2. Схема роботи морфологічного аналізатора із застосуванням допоміжних процедур

Результати тестування розроблених нами доповнень до системи АМА АГАТ демонструє Табл. 1.

Як видно, ефективність роботи допоміжних модулів коливається залежно від типу неопрацьованих словоформ та функціонального стилю текстів. Найкращих результатів вдалося досягти у виявленні одиниць із цифровими фрагментами та слів, написаних неукраїнською абеткою, також досить високі показники продуктивності продемонстрували правила діагностики абревіатур та власних назв. Натомість поки що не вироблено успішної стратегії аналізу випадків нетипової орфографії, рідковживаних слів та неологізмів, які не є власними назвами. Опрацювання таких класів словоформ потребує проведення подальших досліджень та використання складніших алгоритмів аналізу. Також проблемним лишається питання автоматичного кодування складних слів, записаних через дефіс, та друкарських помилок (зокрема помилок розпізнавання тексту).

Різне кількісне співвідношення типів неопрацьованої лексики в межах різних функціональних стилів зумовлює неоднаковий результат роботи допоміжних процедур на відповідних текстових зразках.

ТАБЛ. 1.

РЕЗУЛЬТАТИ ОПРАЦЮВАННЯ ТЕКСТОВОГО МАТЕРІАЛУ (ЗА ПІДВИБІРКАМИ)

	Публіцистика			Художні			Наукові		
	опра-цьова-ні	неопра-цьо-вані	к-ть помилок (%)	опра-цьова-ні	неопра-цьо-вані	к-ть помилок (%)	опра-цьова-ні	неопра-цьо-вані	к-ть помилок (%)
буквено-цифровий запис	381	0	25 *(6,56)*	36	0	2 *(5,56)*	280	0	30 *(10,71)*
скорочення	2374	52	145 *(6,11)*	119	3	18 *(15,13)*	1135	7	23 *(2,03)*
власні назви	2263		108 *(4,77)*	4416		872 *(19,75)*	3007		433 *(14,40)*
неукраїнські символи	592	0	0	779	353	0	3711		0
дефісний запис	924	46	8 *(0,87)*	718	130	33 *(4,60)*	1123	129	75 *(6,68)*

Так, значна частка скорочень, слів із неукраїнськими символами і буквено-цифровим записом у публіцистичному та науковому підкорпусах забезпечують вищу ефективність роботи на текстах цих стилів. На противагу їм, художній стиль містить велику кількість рідковживаної лексики та зразків авторської орфографії, які важко піддаються опрацюванню.

Висновок

Проведений експеримент засвідчив, що, незважаючи на якісні переваги використання словникового морфологічного аналізу, є певні групи лексики і певні категорії текстового матеріалу, що вимагають застосування інших, несловникових методів. Такими, зокрема, є тексти, що містять значний відсоток лексем у нестандартному записі та рідковживаних слів. Відповідно, оптимізація роботи морфологічного аналізатора передбачає часткову відмову від уніфікованого підходу. Шляхом комбінування різнопланових методик видається можливим підвищити ступінь покриття матеріалу аналізатором в середньому на 1,5%. Ефективність застосованої стратегії виявилася неоднаковою для текстового матеріалу різних функціональних стилів. Найкращих результатів вдалося досягти для публіцистики, дещо менших – для наукових текстів; найгірше піддаються опрацюванню тексти художньої літератури.

Крім того, експеримент засвідчив, що кожен клас лексики, "проблемної" для системи, вимагає специфічної стратегії опрацювання. Так, власні назви, абревіатури й лексеми, що містять у своєму складі цифри, можна ідентифікувати за графічними ознаками. Слова з нетиповим для української мови графемним складом (збігом великої кількості голосних або приголосних) досить вдало опрацьовуються методом графемного аналізу. Однак запропоновані методи не завжди дозволяють отримати точний результат.

Загалом же, попри очевидно позитивний результат роботи варто визнати, що робота над деякими типами помилок потребує детальнішого розгляду та залучення складніших методик опрацювання текстового матеріалу.

Література

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков / О. И. Бабина, Н. Ю. Дюмин // Вестник Южно-Уральского государственного университета. Серия "Лингвистика". №25(284), выпуск 15. – Челябинск, 2012. – С. 38-44.
2. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті / Г. Б. Буньо // Наукові записки [Національного університету "Острозька академія"]. Серія : Філологічна. – 2014. – Вип. 49. – С. 12-16.
3. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи : (монографія) / Н. П. Дарчук. – Київ : Освіта України, 2013. – 543 с.
4. Дарчук Н. Морфологічне анотування Корпусу української мови / Н. П. Дарчук // Комп’ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К.: КНЛУ, 2012. – С 16-19.
5. Лангенбах М. Автоматичний синтаксичний аналіз речення за принципами граматики залежностей / М. О. Лангенбах // Науковий вісник Східноєвропейського національного університету імені Лесі Українки. Серія : Філологічні науки. Мовознавство. – Луцьк, 2015. – № 3. – С. 249-254.
6. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів / Н. Міщенко //

Наукові записки. Серія "Філологічні науки". Випуск 95(2). – К., 2011. – С. 538-542.
7. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы) [Електронний ресурс] / И. М. Ножов. – М., 2003. – 140 с. – Режим доступу: http://aot.ru/docs/Nozhov/msot.pdf.
8. Перебийніс В. Математична лінгвістика. Навчальний посібник / В. І. Перебийніс. – К.: Вид. центр КНЛУ, 2014. – 125 с.
9. Brooks A.Unsupervised Part-of-Speech Tagging: An Introduction [Електронний ресурс] / A. Brooks,
10. M. Stees. – Режим доступу: https://ou.monmouthcollege.edu/_resources/pdf/academics/mjur/2014/Unsupervised-Part-of-Speech-Tagging.pdf
11. Blevins J. Word-based morphology [Електронний ресурс] / J. P. Blevins // Journal of Linguistics. – Режим доступу: https://www.researchgate.net/publication/231781491_Word-based_morphology
12. Manning C. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? / C. D. Manning // Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6608. – Springer, Berlin, Heidelberg. – P. 171-189.

Використання статистичних методів описової статистики у корпусній лінгвістиці

Зоя Кочуєва

к. т. н., професор, доцент кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: kochueva@kochuev.com

Валерій Дідусьов

магістр кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: valeradidusiov@gmail.com

The purpose of this study: the development of software to determine the criteria of Pearson Hi^2 on the basis of both scientific and literary texts. In the first part of the study it is conducted the analysis of contemporary corpus linguistics. Statistical studies and their views of corpus linguistics are also described. Held his own description of the body of fiction and non-fiction texts. In the second part of the paper it is focused on statistical methods in corpus linguistics, namely the descriptive statistics. The description of the operation of the developed software.

Ключові слова — комп'ютерні науки, лінгвістичні аспекти, корпус, критерії, генеральна сукупність, вибірка.

I. Вступ

Наукові дослідження в рамках різних лінгвістичних напрямків мають об'єктом свого дослідження текст або зібрання текстів і мають на увазі спочатку підбір матеріалу, а потім аналіз і обробку великих текстових обсягів з метою виявлення деяких мовних закономірностей. Традиційні лінгвістичні методи аналізу тексту дозволяють виконати всі перераховані вище завдання, але їх невисока ефективність обумовлює все більш часте використання методів комп'ютерного аналізу тексту, який дозволяє скоротити роботу лінгвіста, при цьому значно збільшити обсяг оброблюваних даних, а також уникнути неточності і помилки в підрахунках. Таким чином, комп'ютерний аналіз тексту уможливлює встановлення мовних закономірностей, заснованих не так на теоретичних, як на емпіричних даних [4].

У межах досліджень із корпусної лінгвістики під «корпусом» розуміють «уніфікований, структурований і розмічений масив мовних даних в електронному вигляді, призначений для певних філологічних і гуманітарних досліджень». Перевага застосування корпусного аналізу тексту полягає в великій мірі в об'єктивності дослідження: такі функції як підбір, розмітка, аналіз текстів та виявлення відповідностей виконуються автоматично. Таким чином, завданням дослідника є не аналіз матеріалу, а обробка отриманих даних, виведення мовних закономірностей і підведення підсумків.

Розвиток корпусної лінгвістики і зростання уваги до статистичних методів обробки мовного матеріалу за останнє десятиліття привели до розробки цілого ряду методик, пов'язаних з обробкою паралельних або близьких текстів різними мовами. Актуальною задачей є застосування до таких текстів методів статистики та лінійного програмування, які радикально скорочують трудомісткість робіт. У рамках даного дослідження увага приділяється застосуванню методів описової статистики, а саме методу для знаходження значення критерію Хі^2 для підтвердження гіпотези щодо відношення тексту до відповідного корпусу.