УДК 81'33+811.411.21

ОПРАЦЮВАННЯ КОРПУСУ ДЛЯ ЧАСТОТНОГО СЛОВНИКА АРАБСЬКОЇ МОВИ

Бабак Богдан Георгійович,
студ.
Київський національний університет імені Тараса Шевченка

У статті розглянуто етапи створення першого у світі частотного словника арабської мови. Розповідається про складнощі, які при цьому виникали, головною з яких було "рафінування" словника, тобто виокремлення словоформ. Робота з рафінування ще повністю не завершена, але виконано найбільш дидактично важливу її частину – першу тисячу найчастотніших слів, з перекладом їх українською та англійською мовами.

Ключові слова: *арабська мова, частотний словник, корпус, текстовий масив*

Сприйняття мови неможливе без словника. При сприйнятті мови основною оперативною одиницею виступає СЛОВО. Із цього випливає, зокрема, що кожне слово сприйманого тексту повинне бути ототожнене з відповідною одиницею внутрішнього словника слухача (або читача). Природно вважати, що вже із самого початку пошук обмежений деякими підобластями словника. Відповідно до більшості сучасних теорій сприйняття мови, власне фонетичний аналіз тексту, що звучить в типовому випадку дає лише деяку ЧАСТКОВУ інформацію про можливий фонологічний вигляд слова, і такого роду інформації відповідає не одне, а певна МНОЖИНА слів словника; отже, виникає завдання (а) виділити відповідну множину по тим або інших параметрах й (б) у межах обкресленої множини (якщо воно виділено адекватно) зробити "відсіювання" всіх слів, крім того єдиного, яке й відповідає щонайкраще даному слову розпізнаваного тексту. Одна зі стратегій "відсіювання" – виключення низькочастотних слів. Звідси випливає, що словник для сприйняття мови – це частотний словник. Саме створення комп'ютерної версії частотного словника арабської мови, як яскравого прикладу частотного списку східної мови і було нашим головним ***завданням***.

Робота зі складання частотного словника вимагає насамперед терпіння, посидючості й певної лінгвістичної кваліфікації. Можна обмежитися чисто формальним визначенням текстового слова, слововживання як ланцюжка букв від пробілу до пробілу (як це прийнято зазвичай в статистичній лексикографії). Відповідно слово (точніше, словоформа) словника даного тексту буде зовні повністю збігатися зі слововживанням. Але слово (словоформа) – це одна з багатьох одиниць тексту, тобто одиниця його словника, а слововживання – це одиниця самого тексту, одна із всіх його одиниць. Тут довелося докладно пояснювати звичні для лексикографа-частотника терміни, які нерідко бентежать філолога звичайного. Отже, якщо обмежитися тільки зовнішнім виглядом одиниці частотного словника, що прийнята до уваги в тексті, то начебто немає ніяких проблем. Так що зробити хоч який-небудь частотний словник в принципі не так вже й важко. Складніше зробити гарний частотний словник. Дійсно, підраховувати в тексті слова при такому підході просто навіть і з записником, а комп'ютером й поготів. Наприклад, Дж.Чивер уже в 1964 р. у своєму "Скандалі в сімействі Уогапотів" наділив одного з героїв роману, програміста військового НДІ, стомленого рутинними обов'язками, ідеєю використати комп'ютерний комплекс для виготовлення частотного словника віршів Кітса, що той і зробив цілком успішно. Зазначені навіть довжина всіх текстів і число різних

слів (певно, словоформ у термінах "від пробілу до пробілу"), дорівнює 15 357 слововживанням й 8 503 словоформам відповідно.

Залишається лише вирішити, який саме текст (тексти) піддати аналізу, якою повинна бути довжина аналізованого тексту (або сумарна довжина, обсяг текстів) у слововживаннях, а далі підуть питання, які вирішити без лінгвістичної підготовки було б нереально. Справа в тому, що одиницею майбутнього словника може бути не тільки словоформа безвідносно до її морфологічних і семантичних ознак у тексті: тоді доведеться враховувати лексико-граматичну (частини мови) і граматичну (граматичні категорії) омонімію (омографію). Такою одиницею може виявитися словникове слово, тобто слово в його вихідній, словниковій формі. Це може бути й словосполучення – і доведеться вирішувати проблеми класифікації словосполучень. Це може бути термін, односкладовий або складовій. Не встигнувши визначити власну точку зору по тому або іншому питанню, доведеться зіштовхнутися з іншими, щодо яких у літературі немає однакових рекомендацій. Зрозуміло, крім базової філолого-лінгвістичної підготовки, майбутньому укладачеві частотного словника буде потрібно згадати елементарну математику для початкової школи, а зрідка й дещо з матеріалів курсу середньої школи.

ЕТАПИ СТВОРЕННЯ ЧАСТОТНОГО СЛОВНИКА АРАБСЬКОЇ МОВИ.

Будь-який довготривалий процес може бути природно або штучно розподілений на певні етапи. Створення частотного словника арабської мови забрало у автора приблизно 2 роки. Цей період часу звичайно не можна порівнювати з попереднім досвідом дослідників, коли цілі колективи дослідників вимушені були працювати десятиліттями над обробкою величезних текстових масивів. Ця праця була б неможливою без технічної підтримки Рудого Б.А., та його авторської комп'ютерної програми *Text Analyser* [http://www.langs.com.ua].

Створення текстового масиву.

Корпус для частотного словника містить усього лише 120 тис. слів. Невеликий обсяг пояснюється тим, що створення словника розпочиналося у рамках написання студентської курсової роботи. Сучасні частотні словники мають корпуси деяких мов (англійська, російська, німецька та ін.) обсягом понад 40 млн. слів. Але обсягу у 120 тис. слів теж цілком досить для об'єктивної інформації про частотні слова мови, особливо верхньої частини списку.

Саме тут яскраве відображення знайшли менталітетні, психолінгвістичні особливості етносу. Яскравим прикладом тому – ранг і місце слів "Аллах", чи пари слів "він-вона".

Автор використовував мережу Internet в якості поля для відбору інформації. Цей вибір не випадковий – Internet наразі є полем де діалекти арабської мови почасти ігноруються плюс мережа Internet як ніяке інше місце дасть нам саме сучасний, актуальний зріз мови.

Існують певні правила по відбору інформації. Загальний текст повинен складатися з уривків різних за тематикою довжиною не більше 500 слів. При складанні масиву загальною довжиною в 200.000 слів дана пропорція дасть нам 400 уривків різного тематичного забарвлення.

Щонайменше 50% цих текстів повинні бути з художньої літератури (поезії і прози), інші – з медицини, технічних текстів, журналістики, юриспруденції, психології тощо. Ці особливості (характерні при створенні) чітко проілюстровано в наступних діаграмах і таблицях (*див. Додаток 1* (статистика).

Обробка текстового масиву за допомогою програми *Text Analyser*.

Програма *Text Analyser* відрізняється від своїх аналогів здатністю працювати над великими обсягами матеіалу – це й робить її незамінною при складанні частотних словників. Програма має простий інтерфейс і при правильному ознайомленні дозволяє перетворитися на "обізнаного" користувача вже після нетривалого тренування.

Після завантаження до програми тексту вона автоматично починає його обробку. Залежно від розміру оброблюваного тексту та потужності комп'ютера аналіз тексту займає від кількох годин до кількох діб. В результаті ми отримуємо "сирий", необроблений список частотності в форматі Excel (*див. Додаток 2* (зразок "сирого" списку).

Аналіз попереднього списку частотності.

Отриманий "сирий" список потребує подальшої обробки та глибокого аналізу. Власне на даному етапі подальший розвиток процесу залежить від граматичних особливостей мови тексту, що аналізується. Потрібно визначити нюанси на які необхідно звернути увагу, беручі до уваги зазначені особливості. Ці моменти знадобляться нам на наступному етапі.

Всі відомі нам частотні словники російської мови побудовані на обробці масивів письмових (друкованих) текстів. Почасти із цієї причини, коли тотожність слова багато в чому опирається на збіг формальний, графічний, недостатньо враховується семантика. У результаті виявляються зміщеними, перекрученими й частотні характеристики; наприклад, якщо слова зі словосполучення "один одного" укладач частотного словника включає в загальну статистику вживання слова "один", то навряд чи це виправдано: з огляду на семантику, ми повинні визнати, що в складі сполучення це вже *інші* слова, а точніше, що самостійною словниковою одиницею виступає сполучення в цілому.

"Рафінування" словника, виокремлення словоформ.

Даний етап є однім з найголовніших та найбільш кропітких, оскільки всю роботу потрібно виконувати власноруч.

У всіх існуючих словниках слова подані лише у своїх основних формах: іменники у формі одн., ім.п., дієслова у формі інфінітива й т.д. Деякі зі словників подають інформацію про частотність словоформ, але звичайно роблять це недостатньо послідовно, не вичерпним образом.

Частотності різних словоформ одного й того самого слова свідомо не збігаються. Розроблювач же моделі сприйняття мови повинен ураховувати, що в реальному перцептивному процесі розпізнаванню підлягає саме конкретна словоформа, "занурена" у текст: на базі аналізу початкової ділянки експонента словоформи формується безліч слів з ідентичним початком, причому початкова ділянка словоформи не обов'язково тотожна початковій ділянці словникової форми. Саме словоформі належить конкретна ритмічна структура – також надзвичайно важливий параметр для перцептивного відбору слів. Нарешті, у підсумковому поданні розпізнаного висловлення знов-таки слова представлені відповідними словоформами.

Існує безліч робіт, у яких демонструється важливість частотності в процесі сприйнятті мови. Але нам не відомі роботи, де використалася б частотність *словоформ* – навпаки, всі автори практично ігнорують частотність окремих словоформ, звертаючись винятково до лексем. Якщо отримані ними результати не вважати артефактами, доводиться припустити, що носію мови якимось чином доступна інформація про співвідношення частотностей словоформ і словникової форми, тобто, фактично, лексеми. Причому такого роду перехід від словоформи до лексеми, звичайно, неможливо пояснити природним знанням відповідної парадигми, оскільки інформація про частотність повинна використатися ДО остаточної ідентифікації слова, інакше вона просто втрачає сенс.

Можливо, варто визнати реалістичність *кумулятивної* частотності для всіх слів (словоформ), однокорінних даному (даній). Тоді всім членам такого класу надавався б той самий індекс кумулятивної частотності.

При створенні арабського частотного словника укладачам у першу чергу довелося звернути увагу на арабський прийменник "و" ("і"), оскільки даний прийменник почасти пишеться дуже близько до наступного слова і тому програма не завжди могла його виділити. Проблему поглибив той факт, що велика кількість арабських слів починається літерою "و", яка є частиною слова, тому виділити її автоматично було неможливо.

Певну проблему також становили арабські присвійні займенники, котрі пишуться злитно з попереднім словом (в кінці слова). Таким чином їх виділення також перетворилося на екстра завдання для укладачів.

Необхідно було також виділити словоформи арабського дієслова, а також "ламану" множину арабського іменника.

Остаточно проаналізованими на даний момент залишаються лише перша тисяча слів частотного списку арабської мови.

Таким чином виникла можливість зробити аналіз отриманих результатів.

Перші 100 слів покривають 29,5% середньостатистичного арабського тексту. А 50% покриваються приблизно 900-ма словами (так зване число N_{50}). Для рафінованих частотних словників такі цифри є рекордно малими. Для порівняння: для основних індоєвропейських мов N_{50} = 100–230. Це означає унікальність арабської мови серед мов світу. Вона має особливо багато синонімів та винятків. Але, все ж і такий словник становить величезну дидактичну цінність (*див. Додаток 3* (аналіз першої 1000).

Аналіз остаточного списку частотності.

Отримавши цей список частотності, автору залишається лише зробити остаточні підрахунки й висновки. Так перша сотня слів частотного списку, виявляється, несе в собі особливості не тільки лінгвістичні, але скоріше навіть лінгвокраїнознавчі, чи навіть психологічні.

Частотні словники можуть слугувати яскравим прикладом того, як діалекти однієї і тієї ж мови можуть надзвичайно суттєво відрізнятися поміж собою (діалекти китайської чи арабської напр.)

Висновки. У доповіді розглянуто етапи створення першого у світі частотного словника арабської мови. Розповідається про складнощі, які при цьому виникали, головною з яких було "рафінування" словника, тобто виокремлення словоформ.

Робота з рафінування ще повністю не завершена, але виконано найбільш дидактично важливу її частину – першу тисячу найчастотніших слів, з перекладом їх українською та англійською мовами. Тепер постає необхідність у створенні підручника чи принаймні розмовника на основі цієї лексики.

В статье рассмотрены этапы создания первого в мире частотного словаря арабского языка, рассказывается о трудностях, которые при этом возникали, главной из которых было "рафинирование" словаря, то есть вычленение словоформ. Работа над рафинированием еще полностью не завершена, но выполнена дидактически наиболее важная ее часть – первая тысяча наиболее частотных слов с переводом на английский и украинский языки.

Ключевые слова: *арабский язык, частотный словарь, корпус, текстовый массив.*

In the article the author observes the steps of creation of the first frequency vocabulary of the Arabic language. The article deals with the problem of the vocabulary's refinement. It is postulated that the first thousand of Arabic words is refined and translated into English and Ukrainian.

Key words: *the Arabic language, frequency vocabulary, frame, text aray.*

Література:

1. *Алексеев П.М.* Статистическая лексикография (типология, составление и применение частотных словарей): учебное пособие / П.М. Алексеев. – Л.: ЛГПИ им. А.И. Герцена, 1975.

2. *Тищенко К. М.* Метатеорія мовознавства / К.М. Тищенко. – К.: Основи, 2000.

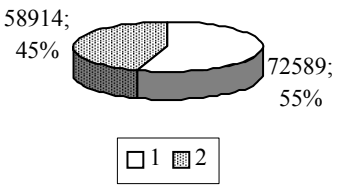
3. *Тищенко К.Н.* Лингвостатистические законы и содержание обучения языку / К.Н. Тищенко // Вестник Киевского ун-та. Вып. "Романо-германская филология". – К., 1985. – С. 3–8.

4. Частотный словарь русского языка / [под ред. Л.Н. Засориной]. – М., 1977.

Додаток 1. Статистика

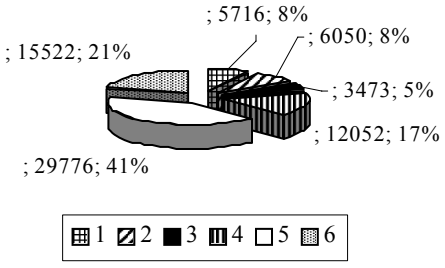
№	Назва	Кількість слів
1.	Художня література	72589
2.	Інше	58914
Всього		131503

Статистика використаних текстів (загальна)



№	Назва	Кількість слів
1.	Essays	5716
2.	Book reviews	6050
3.	Literature news	3473
4.	Literature articles	12052
5.	Prose(short stories)	29776
6.	Poems	15522
Всього		72589

Статистика використаних текстів (художня література)



Додаток 2. Зразок "сирого" списку

Rank X	*Word*	*f*	*Rank X*	*Word*	*f*
1	في	1600	13	على	457
2	في	1570	14	عم	339
3	،	1306	15	التي	336
4	من	1264	16	على	334
5	من	1249	17	عن	320
6	على	747	18	عن	319
7	أن	686	19	وأ	309
8	ال	674	20	ام	290
9	على	642	21	وأ	290
10	أن	573	22	ام	287
11	الله	555	23	التي	248
12	و	477	24	الذي	239

Додаток 3. Аналіз першої 1000

Українська	*English*	*Rank X*	**Слово**	*f*
і	and	1.	و	5626
в	in	2.	في	3398
з *(напрямок дії)*	from	3.	من	2513
на	at, in, to	4.	على	1471
	that	5.	أن	1468
ні	no	6.	ال	1029
до *(напрямок дії)*	to	7.	على	970
Аллах	God (Lord)	8.	الله	893
про	about	9.	عن	789
що	pron. what	10.	ام	753
котра	which (f)	11.	التي	675
чи	or	12.	وأ	623
був	was	13.	كان	553
цей	this	14.	اذه	524
сказав	said	15.	قال	498
котрий	which (m)	16.	الذي	489
він	he	17.	وه	481
цей	this	18.	ذلك	477
ця	this (f)	19.	هذه	456
з (кимось)	with	20.	عم	440
що	that	21.	إن	419
вона	she	22.	هي	410
ним	by him	23.	عليه	384
ні	not	24.	لم	378