УДК 004:37

Я. А. Єрмолова, І. С. Кравчук

Харківський національний університет імені В. Н. Каразіна

Частотні словники: укладання й застосування

Єрмолова Я. А., Кравчук І. С. Частотні словники: укладання й застосування. Стаття присвячена проблемам укладання частотних списків різних мовних одиниць і сферам їх застосування. Детально розглядається процес лематизації як попередній етап укладання частотних словників. Проведено узагальнення основних методів лематизації. Описано два алгоритми укладання частотних списків. Схарактеризовані сфери використання частотних словників.
Ключові слова: частотні словники, методика викладання мов, стилеметрія, мовна картина світу, соціоніка.

Ермолова Я. А., Кравчук И. С. Частотные словари: составление и применение. Статья посвящена проблемам составления частотных списков различных языковых единиц и сферам их применения. Детально рассматривается процесс лемматизации как предварительный этап составления частотных словарей. Проведено обобщение основных методов лемматизации. Описаны два алгоритма составления частотных списков. Охарактеризованы сферы использования частотных словарей.
Ключевые слова: частотные словари, методика преподавания языков, стилеметрия, языковая картина мира, соционика.

Yermolova J. A., Kravchuk I. S. Frequency Dictionaries: Compilation and Use. This paper is devoted to the listing issues of the various language units and to the domains of these lists application. There is considered in details the lemmatization process to be as the preliminary stage for frequency dictionary design. Generalization of the basic lemmatization methods is made. Two algorithms of frequency words listing are described. The domains for frequency dictionaries use are characterized.
Keywords: frequency dictionaries, language teaching methods, stylemetrics, the linguistic world view, socionics.

Одним з найважливіших завдань комп'ютерної лексикографії є укладання частотних словників, тобто списків мовних одиниць, які трапляються в даному масиві текстів, і відповідних частот вживання цих одиниць. В залежності від мети дослідження одиницями підрахунку можуть бути літери, склади, морфеми, словоформи, лексеми, словосполучення, граматичні або семантичні класи слів. Частоти можуть бути абсолютними або частотними.

Незважаючи на існування у вільному доступі в Інтернеті комп'ютерних програм укладання частотних словників, у багатьох випадках виникає потреба у побудові власної програми, яка б реалізувала вказане завдання. Така потреба виникає завжди тоді, коли укладання частотних словників є не самостійним завданням, а частиною іншого алгоритму, наприклад, вимірювання відстаней між текстами й тощо.

Одиницями частотних підрахунків найчастіше є словоформи і лексеми. Частотний словник словоформ уважається простішим для укладання і менш досконалим, оскільки він не вимагає лематизації. Але існують такі завдання, коли є необхідним саме словник словоформ. Наприклад, при автоматичному визначенні відстаней між текстами порівнюються саме словоформи, а не лексеми, що дає можливість враховувати не тільки лексичну, але й граматичну схожість і відмінність порівнюваних текстів.

Вхідними даними для алгоритму укладання словника є текст: або первісний, або закодований символами класів слів чи словоформ. Якщо планується укладання словника лексем, то власне укладанню передує етап лематизації [8:129–135], тобто заміни словоформ їх машинними основами. Останні у деяких випадках можуть відрізнятися від канонічних основ, наприклад, рос. *граф-* для словоформи *графлю*.

Лематизація може бути здійснена двома способами: виходячи 1) зі списку основ або 2) зі списку квазизакінчень, тобто кінцевих послідовностей літер, які з достатньою вірогідністю прогнозують необхідність відсічення даного закінчення в даній словоформі. Ці два способи лематизації можна позначити як

сегментацію зліва – направо і справа – наліво. Перевагою другого способу є можливість сегментації нових словоформ, які з'являються в аналізованих текстах.

У комп'ютерній лінгвістиці було запропоновано кілька різних методів представлення знань про квазизакінчення. Причому відношення між цими методами розглядалися як взаємовиключні. Насправді ж вони є різними графічними способами подання однієї й тієї ж інформації і знаходяться у відношенні не взаємовиключення (строгої диз'юнкції), а у відношенні доповнювальності (кон'юнкції).

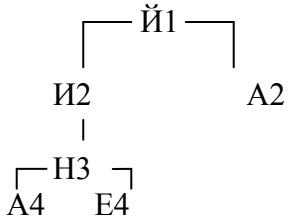
Пояснімо сказане на прикладі. Нехай буде дано два мікросписки словоформ російської мови, що закінчуються літерою *-й*:

(1а) *сара-й, послани-й, окончани-й, укреплени-й*;

(1б) *больш-ой, сер-ый, син-ий, син-ей*.

У списку (1а) містяться форми із закінченням *-й*, а у списку (1б) – форми із іншими закінченнями, хоча вони і містять ту ж саму кінцеву літеру *-й*. Складімо правило для відсічення закінчення *-й*. Для цього треба виділити у формах списку (1а) такі мінімальні послідовності кінцевих літер, які відсутні у списку (1б): *-ай, -аний, -ений.* Отримані квазизакінчення використовуються наступним чином: якщо аналізована словоформа містить одне з вищенаведених квазизакінчень, то у цій словоформі відсікається закінчення *–й*. Само собою зрозуміло, що повнота й адекватність складеного таким чином списку квазизакінчень істотно залежить від корпусу тих словоформ, на підставі яких цей список складався. Але перевагою описаної процедури є можливість програмного складання і подальшого поповнення списку квазизакінчень.

Отримані списки можна представити у вигляді або об'єктних (дихотомічних), або ознакових графів [1:222–225]. У вузлах останніх містяться не самі об'єкти, а їх ознаки, у нашому випадку літери квазизакінчень, нумеровані справа – наліво:



Як бачимо, у цьому графі, на відміну від списку, немає повторних входжень однакових літер, що зменшує обсяг збережуваної інформації. Крім того, деревовидне представлення зменшує кількість порівнянь при пошуку елемента в масиві. Наведене дерево можна подати в лінійній формі у вигляді виразу алгебри висловлень [9:127–151]:

$$F(-Й)=Й1\wedge(А2\vee(И2\wedge Н3\wedge(А4\vee Е4))),$$

де *F* – функція відсічення (factoring) закінчення *–Й*. Літери, які знаходяться на одному рівні галуження дерева, пов'язані між собою логічною зв'язкою диз'юнкції, а літери різних рівнів – зв'язкою кон'юнкції. Для сегментації заданої словоформи обчислюються значення відповідних правил відтинання закінчень. Певне закінчення відтинається лише тоді, коли значення відповідного йому виразу алгебри висловлень дорівнює 1 (істині).

Граф квазизакінчень можна подати також у вигляді таблиці з використанням посилань – адрес зв'язку (АЗ):

№ з/п	Букви	АЗ1	АЗ2
1	*Й1*	2	
2	*А2*	КГ	3
3	*И2*	4	
4	*Н3*	5	
5	*А4*	КГ	6
6	*Е4*	КГ	КГ

де КГ – кінець галуження дерева, АЗ1 – адреса зв'язку, яка використовується для переходу до літер наступного рівня, АЗ2 – адреса для переходу до літер того ж самого рівня [1:222–235; 5:266–268].

Після лематизації починається процес власне укладання частотного словника. Цей процес може відбуватися по-різному в залежності, по-перше, від структури вхідних даних, а по-друге, від наявності подвійного обсягу пам'яті, яка використовується при укладанні словника.

При наявності подвійного об'єму пам'яті частотний словник створюється у три кроки:

1. Представлення вхідних даних у вигляді масиву.

2. Вибір попарно різних елементів масиву методом включення елементів у вільний масив пам'яті. Збільшення абсолютної частоти відповідного елементу, якщо він уже є у масиві включення.

3. Заміна абсолютних частот елементів відносними частотами.

При відсутності додаткової пам'яті словник укладається у такій послідовності:

1. Представлення вхідних даних у вигляді масиву.

2. Вибір попарно різних елементів шляхом стирання повторних входжень елементів.

При знаходженні повторюваного елемента він стинається, а його абсолютна частота збільшується.

3. Компресія масиву шляхом елімінації стертих елементів.

4. Заміна абсолютних частот елементів відносними частотами.

5. Упорядкування масиву або за абеткою, або за частотами.

Головною сферою застосування частотних словників є викладання іноземних мов. Частотні словники сприяють оптимізації цього процесу. Було встановлено, що 1000 найбільш вживаних слів англійської мови покриває 80,5% слововживань у середньостатистичних текстах, 2000 слів – приблизно 86% слововживань, а 3000 – відповідно 90%. З цього випливає, що оптимально методика викладання мови мусить спиратися на частотний словник певної підмови.

Наступною важливою сферою застосування частотних словників є стилеметрія, тобто визначення приналежності даного тексту до одного з фіксованих класів текстів. До таких задач відносяться, зокрема, атрибуція писемних та усних текстів, визначення соціального і психологічного портрету особи за даними писемного та усного мовлення й тощо. Типові випадки таких задач описуються наступними ситуаціями:

1) Множинна невизначеність. Є множина текстів, треба встановити, скільки авторів їх писали і кому з них належить певний текст.

2) Порівняння зі зразком. Є приклад тексту (текстів) деякого автора X. Треба встановити, чи є він автором деякого іншого тексту (текстів).

3) Конкуренція зразків. Є зразки текстів авторів X, Y, Z … . Треба встановити, хто з них є автором текстів A(1), A(2),…, A(N).

Кількісні методики визначення авторства тексту спираються на стохастичні моделі породження мовлення. В основі цієї моделі лежать уявлення про те, що із зростанням обсягу тексту його мовні особливості стають сталими з імовірнісної точки зору, що дозволяє встановлювати авторство за стійкими формальними характеристиками тексту.

Сучасні методики статистичного аналізу цих характеристик спираються на ідеологію теорії розпізнавання образів, у відповідності до якої процес визначення авторства передбачає наступні процедури:

1) подання аналізованих текстів у вигляді вектора значень формальних параметрів стилю;

2) обчислення відстаней між аналізованими текстами;

3) прийняття рішення про авторство анонімного тексту на підставі обчислених відстаней між текстами.

Конкретні методики атрибуції обов'язково (імпліцитно чи експліцитно) містять вказані процедури, але істотно відрізняються «речовинним» наповненням цих процедур.

Як параметри стилю для статистичного аналізу можуть бути використані одиниці різних ярусів: графеми, лексеми, функціональні класи слів, синтаксичні конструкції. У першому випадку використовуються *n*-грамні послідовності графем. Такі ознаки вимагають текстів великого обсягу, що у більшості прикладних ситуацій є нездійсненим.

Частотні списки словоформ можна використати як для атрибуції текстів, так і для встановлення соціонічного типу авторства [6]. При цьому як ознаки опису тексту розглядаються нормалізовані частот словоформ у достатньо великому тексті. Нормалізація здійснюється шляхом поділу частот словоформ в аналізованому тексті на середньомовну частоту, наприклад, за частотним словником Шарова [3]. Відстань між словами обчислюється як сума квадратів різниць між частотами однакових слів у словниках порівнюваних текстів.

Хоча частоти слів у багатьох випадках можуть виконувати діагностичну функцію, однак для їх отримання потрібні предметно й жанрово однорідні тексти великої довжини. На практиці ж найчастіше доводиться мати справу з достатньо короткими текстами довжиною приблизно у 500–1000 словоформ. У таких текстах більшість слів, крім службових, трапляється по одному разу. І тому висновки про тотожність чи відмінність авторства стають статистично необґрунтованими. У подібних випадках слід перекодувати текст, замінивши елементи тексту символами збільшених класів цих елементів. Це призводить до збільшення повторюваності одиниць представлення тексту і зменшує вплив предметних і жанрових особливостей тексту на його параметричний опис. Повторюваність одиниць опису тексту при цьому збільшується, що робить статистичні висновки більш достовірними.

У ролі згаданих збільшених одиниць опису тексту варто використовувати функціональні класи слів із зазначенням варіативних граматичних категорій. Для закодованих таким чином текстів будуються частотні списки *n*-місних ($n=1,\ldots,k$) послідовностей класів словоформ, після чого ці списки опрацьовуються за допомогою обраних вирішальних

правил. Зокрема, при диференціації авторства методом порівняння за зразком може використовуватись формула порівняння часток для встановлення суттєвості/несуттєвості частотних розходжень [2:14]. Так, подібним методом нами було встановлено, що інформативними, тобто диференціальними для текстів М. Шолохова і К. Симонова є частоти дієприкметників, прислівників, займенників і сполучників. Відповідна програма для встановлення суттєвості кількісних розходжень нами була реалізована за допомогою Excel.

Відносно нещодавно частотні словники стали використовуватись як дослідницький інструмент у філософії і літературознавстві. У цьому випадку мова розглядається як засіб відображення сукупної картини світу (акумулятивна функція мови).

Частотні словники показують, які значення є найбільш необхідними у людському спілкуванні, тобто являють собою своєрідну ієрархію цінностей у цій картині. Частотні словники різних мов мають багато спільного між собою, принаймні у верхній частині списку. Однак картина світу виражена в частотних словниках неявно, імпліцитно. Завдання філософії мови полягає в експліцитній інтерпретації цієї картини, у поясненні і тлумаченні того, про що свідчить сама мова. Наприклад, М. Епштейн уважає, що одне з найбільш частотних слів англійської мови, а саме визначений артикль *the* відповідає одному з центральних понять філософії – поняттю «буття» – і називає цю властивість артиклю властивістю «цетості», тобто відмінності будь-якої речі від усіх інших речей у світі [7].

Для російської мови, на думку того ж автора, визначальним , «артикулюючим» є не властивість «цетості», а інша властивість – «уміщуваність», яка передається найчастішим словом ***в***. Цей факт характеризує колективне мовне позасвідоме тих, хто розмовляє російською мовою. Суть цього «позасвідомого» характеризується тим, що річ визначається як існуюча не сама по собі, а як елемент відношення «уміщуваності».

Частотні словники можуть використовуватись не тільки для відтворення загальномовної картини світу, але й індивідуальної картини світу автора [4]. Найбільш показовими для цього виявляються повнозначні слова: дієслова, прикметники, іменники. Наявність чи відсутність певних семантичних груп слів дозволяють схарактеризувати картину авторського світовідчуття та її динаміку. Сукупна частотність і різноманітність слів у кожній тематичній групі дає можливість побачити, наскільки дана сфера життя значима, складна і деталізована в авторському світовідчутті. Частотні словники характеризують не тільки елементи дійсності, що оточують автора, але і його ставлення до них.

Наведені приклади застосування частотних словників (сюди можна додати також контент-аналіз) свідчать про появу нового точного, об’єктивного методу лінгвістичних досліджень, який можна успішно застосовувати при вивченні різних рівнів мови і розв’язувати різноманітні лінгвістичні, літературознавчі, філософські, політологічні задачі.

Література

1. Дмитриева М. В. Элементы современного программирования / М. В. Дмитриева, А. А. Кубенский. — СПб. : Изд-во С.-Петербургского университета, 1991. — 272 с.

2. Ермоленко Г. В. Анонимные произведения и их авторы / Г. В. Ермоленко. — Минск : Изд-во «Университетское», 1988. — 118 с.

3. Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики [Электронный ресурс] / О. Н. Ляшевская, С. А. Шаров // Режим доступа : http://www.dict.ruslang.ru/freq.php

4. Самойлова И. Ю. Динамическая картина мира И. Бродского: лингвистический аспект / И. Ю. Самойлова / Монография. — Гродно, 2007. — 191 с.

5. Флорес И. Структуры и управление данными / И. Флорес. — М. : Финансы и статистка, 1982. — 319 с.

6. Хрулев О. Применение частотного анализа в соционике [Электронный ресурс] / Хрулев Олег // Режим доступа: http://www.socionic.ru/index.php/-3/2869-2010-10-26-00-03-38

7. Эпштейн М. Н. Частотный словарь как философская картина мира [Электронный ресурс] / М. Н. Эпштейн // Режим доступа : www.topos.ru/article/3194

8. Dierks Karin. Automatic Stylistic Analysis of Lyrical Texts / Karin Dierks // — Literary and Linguistic Computing. — Vol. 1. — No. 3. — 1986. — P. 129—135.

9. Oettinger A. G. Automatic language translation / A. G. Oettinger. — Cambridge (Mass.), Harvard University Press, 1960. — 380 p.