УДК 81.139

Суворина Е.В.
Московский городской педагогический университет

ПОСТРОЕНИЕ ЛЕКСИЧЕСКОГО ПРОФИЛЯ СЛОВА ПО РАЗЛИЧНЫМ СТАТИСТИЧЕСКИМ КРИТЕРИЯМ (НА МАТЕРИАЛЕ БРИТАНСКОГО НАЦИОНАЛЬНОГО КОРПУСА)

E. Suvorina
Moscow City State University

THE UNSUPERVISED BUILD OF LEXICAL PROFILE USING VARIOUS STATISTICAL CRITERIA FROM BNC WEB SYSTEM

Аннотация. В статье на примере слов «emotion» и «feeling» показано, что лексический профиль, построенный с помощью различных статистических критериев правдоподобия, мало зависит от конкретного критерия, но отличен от результатов, полученных исключительно по частотным данным. Показано, что лексические профили, построенные методами правдоподобия, допускают лингвистическую интерпретацию. Впервые представлены результаты построения лексических профилей анализируемых слов с разбиением по возрастным когортам.

Ключевые слова: автоматизированная экспертиза лингвистического корпуса, сетевое программное обеспечение BNCweb, лексический профиль, статистический критерий, частотные данные, возрастные когорты.

Abstract. The article shows that the lexical profile suffers from the change of the likelihood statistical criteria negligibly but differs significantly from the one built by raw frequencies. Only the first one allows linguistic interpretation and analysis. The findings presented by the author in the article include lexical profiles for the words “emotion” and “feeling” with additional division into age cohorts.

Key words: unsupervised parsing of a linguistic corpus, the BNC web software, lexical profile, statistical criteria, raw data, age cohorts.

Отличительной чертой современных лингвистических исследований последнего десятилетия является обращение к различным статистическим критериям для проведения тщательного лингвистического анализа изучаемых единиц на основе выбранного языкового корпуса (см., напр., работы Захарова В.П., Хохловой М.В. и др. [1; 3; 5]). Основное условие успешной работы в таком направлении – наличие программного сетевого обеспечения для аннотированного лингвистического корпуса. Именно это условие позволяет лингвистам, имея минимальную пользовательскую подготовку, использовать данные, полученные по различным статистическим критериям (мерам) для составления лексического профиля (ЛП) целевого слова.

Сетевое программное обеспечение BNCweb [4], сопровождающее *Британский национальный корпус*, предлагает следующие статистические критерии: взаимная информация (mutual information), куб взаимной информации (MI3), T-критерий (T-score), Z-критерий (Z-score), логарифмическая функция правдоподобия (log-likelihood), коэффициент Дайса (Dice). Широкий выбор критериев и абсолютные шкалы, различающиеся на порядки величины, могут ввести в заблуждение. Ознакомление же со специальной литературой по математической статистике для лингвистов малопродуктивно. В настоящей статье мы продемонстрируем

на практике применение всех доступных в системе BNCWeb критериев для построения лексических профилей слов «emotion» и «feeling» и покажем их сходство и различие.

Построение лексических профилей слов проводим, ограничиваясь рассмотрением адъективных и глагольных коллокаций (статистически устойчивых словосочетаний), расположенных в диапазоне от 3 до 1 позиции слева от целевого слова. Из первичных получившихся списков слов (выдачи системы BNCWeb) мы отбираем первые 10, имеющие максимальные показатели. Повторяем процедуру для каждого предлагаемого статистического критерия, формируя шесть списков. Искомый профиль будет являться максимальным пересечением получившихся шести множеств – то есть состоять из словосочетаний, чаще всего повторяющихся в наибольшем количестве групп. Абсолютными количественными характеристиками каждого из критериев мы в этом подходе пренебрегаем.

Некоторые из полученных результатов приведены в таблицах 1-7 с разбиением на возрастные когорты. В таблицах по столбцам приведены коллокации, вычисленные по методикам взаимной информации, куба взаимной информации, Z- и T-критериям, логарифмической функции правдоподобия, коэффициенту Дайса. В последней колонке приведены слова, отобранные по «сырым» (raw) частотам встречаемости (rank group).

В таблицах 1-3 приведены данные построения лексического профиля слова «feeling» по трём возрастным когортам: 25-34 года, 35-44 года и 45-59 лет. При детальном рассмотрении полученных данных становится очевидным, что наполнение лексического профиля анализируемого слова меняется при переходе от одной возрастной группы к другой, что позволяет исследователю формировать более тонкую структуру профиля и проводить его лингвистическую интерпретацию.

Неизменными, независимо от возрастных когорт и статистических критериев, остаются лишь коллокации, которые составляют группу частотных употреблений. Для ЛП слова «feeling», например, такими являются следующие глагольные словосочетания: *to have a feeling, to get a feeling, to be a feeling.* Они присутствуют в каждой из представленных групп.

Проанализировав наполнение лексического профиля по различным статистическим критериям в пределах одной возрастной когорты, можно прийти к выводу о том, что максимальное совпадение данных присутствует в группах: *MI3, Z-score, Log* и *Dice*. Отдельное место занимает группа глагольных коллокаций, вычисленных по статистическому критерию взаимной информации (MI). Здесь присутствуют сочетания, довольно редко встречающиеся в корпусе. Такие данные заслуживают отдельного рассмотрения.

В таблицах 3-6 в качестве примера приведены данные сопоставления лексических профилей анализируемых слов («emotion» и «feeling») для возрастных когорт 45-59 лет (глагольные коллокации) и 35-44 года (адъективные коллокации).

Содержание профилей анализируемых слов абсолютно различно, что свидетельствует о разном концептуальном наполнении структур знания, стоящих за словами «feeling» и «emotion» в современном английском языке.

Данные таблиц 6, 7 на примере адъективных коллокаций демонстрируют различие в содержании лексического профиля слова «emotion» в зависимости от возрастной когорты.

Результаты группового анализа лексических профилей целевых слов подтверждают полученные ранее выводы об их принципиальном различии [2]. Расхождение между разными статистическими методами не столь велико, как можно было бы ожидать. Как правило, совпадают группы, рассчитанные по взаимной информации, Z-критерию, логарифмической функции правдоподобия и коэффициенту Дайса.

Таблица 1

Глагольные коллокации ЛП слова *feeling* (возрастная когорта 25-34 года)

MI	MI3	Z	T	Log	Dice	Rank
to stifle	to stifle	to stifle	to have (had)	to have (had)	to stifle	to have (had)
to originate	to have (had)	to originate	to get	to stifle	to escape	to be (was)
to glory	to originate	to glory	to have (have)	to get	to originate	to have (have)
to aggravate	to glory	to aggravate	to be (was)	to originate	to aggravate	to get (get)
to induce	to aggravate	to induce	to get (got)	to have (have)	to glory	to be (is)
to transmit	to induce	to transmit	to give	to glory	to induce	to get (got)
to mediate	to get	to have (had)	to stifle	to aggravate	to give	to be (been)
to dispel	to transmit	to mediate	to escape	to induce	to transmit	to know
to conjure	to mediate	to dispel	to keep	to escape	to mediate	to give
to recede	to have (have)	to conjure	to know	to transmit	to conjure	to like

Таблица 2

Глагольные коллокации ЛП слова *feeling* (возрастная когорта 35-44 года)

MI	MI3	Z	T	Log	Dice	Rank
to tun	to have (had)	to tun	to have (had)	to have (had)	to project	to have (had)
to expose	to tun	to project	to get (got)	to get (got)	to express	to be (was)
to squeak	to project	to have (had)	to have (have)	to project	to get (gets)	to have (have)
to mellow	to be (was)	to express	to be (was)	to express	to promote	to be (is)
to dissipate	to express	to expose	to get (get)	to get (gets)	to overcome	to get (got)
to project	to expose	to squeak	to know	to promote	to express	to get (get)
to savour	to squeak	to mellow	to get (gets)	to tun	to hate	to know
to evoke	to get	to dissipate	to give (gave)	to get (get)	to create	to get (gets)
to divert	to have (have)	to promote	to project	to have (have)	to increase	to give (gave)
to dispel	to mellow	to savour	to express	to overcome	to tun	to have (has)

Таблица 3

Глагольные коллокации ЛП слова feeling (возрастная когорта 45-59 лет)

MI	MI3	Z	T	Log	Dice	Rank
to sublimate	to have (had)	to sublimate	to have (had)	to have (had)	to experience	to have (had)
to swoop	to sublimate	to stifle	to have (have)	to get (get)	to overcome	to be (was)
to have	to stifle	to have	to get (get)	to get (got)	to experience	to be (is)
to nag	to get	to swoop	to get (got)	to give (gave)	(experienced)	to have (have)
to stifle	to have (have)	to have (had)	to give (gave)	to have (have)	to stifle	to get (get)
to pulse	to be (was)	to experience	to be (is)	to experience	to give (gave)	to get (got)
to relish	to be (is)	to nag	to be (was)	to stifle	to enjoy	to have (has)
to depart	to swoop	to overcome	to experience	to overcome	to reflect	to give (gave)
to immerse	to have	to relish	to overcome	to experience	to express	to be (be)
to evoke	to experience	to pulse	to experience	(experienced)	to escape	to know
			(experienced)	to sublimate	to encourage	

Таблица 4

Глагольные коллокации ЛП слова *emotion* (возрастная когорта 45-59 лет)

MI	MI3	Z	T	Log	Dice	Rank
to quaver	to quaver	to quaver	to trigger	to trigger	to trigger	to be (was)
to despise	to trigger	to trigger	to show	to quaver	to quaver	to show
to wobble	to despise	to despise	to quaver	to despise	to despise	to trigger
to counsel	to wobble	to wobble	to despise	to wobble	to wobble	to do
to bribe	to counsel	to counsel	to wobble	to counsel	to counsel	to have
to crackle	to bribe	to bribe	to counsel	to bribe	to bribe	to look
to betray	to crackle	to crackle	to bribe	to crackle	to crackle	toaccommodate
to rein	to betray	to betray	to crackle	to betray	to betray	to say
to trigger	to rein	to rein	to betray	to rein	to rein	to bribe
to excite	to excite	to excite	to rein	to show	to excite	to think

Таблица 5

Адъективные коллокации ЛП слова *feeling* (возрастная когорта 35-44 года)

MI	MI3	Z	T	Log	Dice	Rank
sapped	strangest	strangest	this	strong	uncomfortable	this
lacking	uncomfortable	sapped	strong	strangest	strangest	that
paranoiac	sated	lacking	that	uncomfortable	overwhelming	strong
sated	sapped	paranoiac	uncomfortable	this	unpleasant	same
closed-in	lacking	sated	strange	overwhelming	strong	uncomfortable
quivery	paranoiac	uncomfortable	same	unpleasant	instinctive	strange
all-enveloping	overwhelming	overwhelming	strangest	instinctive	horrible	strangest
unfeigned	strong	instinctive	overwhelming	strange	growing	overwhelming
woozy	this	closed-in	unpleasant	growing	nasty	growing

Таблица 6

Адъективные коллокации ЛП слова *emotion* (возрастная когорта 35-44 года)

MI	MI3	Z	T	Log	Dice	Rank
occurring	expressed	expressed	any	expressed	expressed	any
unspent	occurring	unspent	expressed	sudden	sudden	expressed
expressed	unspent	occurring	sudden	delicate	occurring	sudden
contained	sudden	sudden	such	unspent	unspent	such
isolating	contained	isolating	delicate	occurring	delicate	this
overcome	isolating	contained	human	any	isolating	delicate
freak	overcome	overcome	high	contained	contained	human
unaccustomed	freak	freak	occurring	isolating	overcome	that
populist	unaccustomed	unaccustomed	unspent	overcome	freak	some
vibrant	populist	populist	isolating	freak	unaccustomed	high

Таблица 7

Адъективные коллокации ЛП слова *emotion* (возрастная когорта 45-59 лет)

MI	MI3	Z	T	Log	Dice	Rank
indefinable	indefinable	indefinable	any	pleasurable	pleasurable	this
consoling	pleasurable	pleasurable	pleasurable	indefinable	indefinable	all
overpowering	consoling	consoling	powerful	consoling	consoling	some
subconscious	overpowering	overpowering	some	powerful	overpowering	any
pleasurable	subconscious	subconscious	strong	overpowering	subconscious	powerful
anguished	anguished	anguished	high	subconscious	anguished	such
ongoing	ongoing	ongoing	this	anguished	ongoing	pleasurable
oblivious	oblivious	oblivious	all	ongoing	oblivious	strong
devoid	devoid	devoid	such	oblivious	devoid	other
spontaneous	spontaneous	spontaneous			spontaneous	

Отдельного внимания заслуживает группа коллокатов, составленная по критерию взаимной информации (MI). Проведённый нами анализ подтверждает предположение, выдвинутое создателями BNCweb [4], о том, что этот метод переоценивает вклад от редко встречающихся слов; например, табл. 5: *vibrant (emotion), all-enveloping, woozy (feeling)* и т. д.

Показательно и поучительно, насколько списки слов, составленные по исходной частотности (RANK-group), отличаются от полученных расчётами правдоподобности коллокатов. Так, если принимать во внимание исключительно частотные характеристики, то можно прийти к ложным выводам при сопоставлении профилей анализируемых слов. Кроме того, набор лексем в группе RANK довольно однообразен (см., например, табл. 3, 7) и не представляет особого интереса для лингвистической интерпретации. Построение лексического профиля слова необходимо вести именно по выбранному исследователем статистическому критерию правдоподобия, отказавшись от анализа исключительно частотных данных.

Полученные результаты доказывают, что применение методов корпусной лингвистики в когнитивных исследованиях требует вычисления статистических характеристик правдоподобия совместной встречаемости слов в тексте (речи), а не абсолютных частот. Выбор же конкретного статистического критерия правдоподобия мало влияет на окончательные результаты.

ЛИТЕРАТУРА:

1. Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Материалы конференции по компьютерной лингвистике «Диалог 2010». М., 2010. С. 137-143.
2. Суворина Е.В. Использование сетевого программного обеспечения BNCweb при описании лингво-когнитивных особенностей слов emotion и feeling в современном английском языке // Вестник МГОУ. Серия «Лингвистика». № 1. 2011. С. 73-79.
3. Delgado Ana R. Spanish basic emotion words are consistently ordered // Qual Quant (43). (DOI 10/1007/s11135-007-9121-3), 2009. P. 509 517.
4. Hoffmann S., Evert S., Smith N., Lee D., Prytz Y. Corpus Linguistics with BNCweb – a Practical Guide. Frankfurt am Main: Peter Lang, 2008.288 p.
5. Janda L., Solovyev V. What constructional profiles reveal about synonymy: a case of study of Russian words for SADNESS and HAPPINESS // Cognitive Linguistics. Vol. 20(2). 2009. P. 367-393.