МРНТИ 81-119;001.4

Онгарбаева М.С.[1], Таева Р.М.[2], Колесникова Т.П.[3], Зуева Н.Ю.[4]

[1]PhD докторант, [2]к.ф.н. доцент, [3] старший преподаватель, [4]к.ф.н. доцент
Казахского национального университет им. аль-Фараби, Казахстан, г. Алматы,
e-mail: trmeru16@mail.ru, rozt@mail.ru

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ (ДСА) В ИССЛЕДОВАНИИ СМЫСЛОВОЙ БЛИЗОСТИ СЛОВ ОДНОЙ ЛЕКСИКО-ТЕМАТИЧЕСКОЙ ГРУППЫ (ЛТГ)

В данной научно-исследовательской работе рассматривается вопрос об объективных методах исследования лексического богатства текстов, авторами был проведен количественный анализ распределения слов в тексте, исследованы их морфологические структуры, выявлены зависимости между частотой слова и характером текста. Целью научной данной статьи служит выявление степени семантической близости слов тематической группы «образование». Объектом исследования послужили 7 лексических единиц, относящихся к этой группе: education (образование), knowledge (знание), school (школа), teacher (учитель), learn (изучать), study (учиться), teach (обучать). Материалом послужили тексты художественной и научной прозы современного английского языка. Для исследования семантических связей слов группы методом ДСА все выписанные словосочетания из текстов были распределены по лексическим группам (ЛГ) с общим семантическим признаком. Был установлен средний порог частоты сочетаемости слов: для существительных – 20, для глаголов – 10.

В данной научной работе был применен метод дистрибутивно-статистического анализ (ДСА), для изучения распределения элементов в тексте, который характеризуется максимальной объективностью.

Ключевые слова: метод дистрибутивно-статистического анализа (ДСА), лексическая группа, количественный анализ, коэффициент корреляции, глаголы, существительные.

Ongarbayeva M.S.[1], Tayeva R.M.[2], Kolesnikova T.P.[3], Zuyeva N.Yu.[4]

[1]PhD Doctoral student, [2]Candidate of Philology, [3]Senior Lecturer, [4]Associate professor
Al-Farabi Kazakh National University, Kazakhstan, Almaty,
e-mail: trmeru16@mail.ru, rozt@mail.ru

Distributive statistical analysis (dsa) in the study of semantic proximity of words of one lexico-thematic group (ltg)

This research paper addresses the issue of objective methods for studying the lexical richness of texts, the authors carried out a quantitative analysis of the distribution of words in the text, investigated their morphological structures, revealed the relationship between the frequency of the word and the character of the text. The purpose of this scientific article is to identify the degree of semantic proximity of the words of the thematic group «education». The object of the study were 7 lexical units belonging to this group: education, knowledge, school, teacher, learn, study, teach. The material was chosen from the texts of the fiction and scientific prose of modern English. To study the semantic links of the words of the group using the DSA method, all the written phrases from the texts were distributed into lexical groups (LG) with a common semantic feature. The average frequency threshold of word compatibility was set: for nouns - 20, for verbs - 10.

In this scientific work, the method of distributive statistical analysis (DSA) was used to study the distribution of elements in the text, which is characterized by maximum objectivity.

Key words: method of distributive statistical analysis (DSA), lexical group, quantitative analysis, correlation coefficient, verbs, nouns.

© 2019 Al-Farabi Kazakh National University

Оңғарбаева М.С.[1], Таева Р.М.[2], Колесникова Т.П.[3], Зуева Н.Ю.[4]
әл-Фараби атындағы Қазақ ұлттық университетінің
[1]PhD докторанты, [2]доценті, ф. ғ. к., [3]аға оқытушысы, [4]доценті, ф. ғ. к.
Қазақстан, Алматы қ., e-mail: trmeru16@mail.ru, rozt@mail.ru

Бір лексика-тақырыптық топтағы (ЛТТ) сөздерінің мағыналық жақындығын зерттеуде дистрибутивті-статистикалық талдау (ДСТ) әдісі

Берілген ғылыми-зерттеу жұмысында мәтіндердің лексикалық байлығын зерттеудің объективті әдістері мәселесі қарастырылады, авторлар мәтіндегі сөздердің таралуына сандық талдау жүргізді, олардың морфологиялық құрылымы зерттелді, сөз жиілігі мен мәтін сипаты арасындағы тәуелділік анықталды. Ғылыми мақаланың мақсаты «білім» тақырыптық тобының сөздерінің семантикалық жақындығы дәрежесін анықтау болып табылады. Зерттеу объектісі осы топқа жататын 7 лексикалық бірлік болды: education (білім беру), knowledge (білім), school (мектеп), teacher (мұғалім), learn (үйрену), study (оқу), teach (оқыту). Қазіргі ағылшын Топ сөздерінің семантикалық байланыстарын ДСТ әдісімен зерттеу үшін мәтіндегі барлық сөз тіркестері жалпы семантикалық белгісі бар лексикалық топтарға (ЛТ) бөлінген. Сөздер тіркесімінің орташа шегі орнатылды: зат есімдер үшін – 20, етістіктер үшін – 10.

Бұл ғылыми жұмыста ең жоғары объективтілікпен сипатталатын мәтіндегі элементтердің таралуын зерттеу үшін дистрибутивті-статистикалық талдау (ДСА) әдісі қолданылды.

Түйін сөздер: дистрибутивті-статистикалық талдау әдісі (ДСТ), лексикалық топ, сандық талдау, корреляция коэффициенті, зат есімдер, етістіктер.

Введение

В настоящее время внимание лингвистов привлекает вопрос об объективных методах исследования лексического богатства текстов, количественному анализу подвергаются распределения слов в тексте, исследуется его морфологическая структура, выявляются зависимости между частотой слова и характером текста

Под лексическим значением слова обычно разумеют его предметно-вещественное содержание, оформленное по законам грамматики данного языка и являющееся элементом общей семантической системы словаря этого языка (Vinogradov V.V., 1977: 5). Так, авторы данной статьи обращались к англоязычным словарям.

Лингвистическая статистика – отрасль языкознания, занимающаяся изучением методов раскрытия закономерностей, свойственных большим совокупностям однородных объектов на основании их выборочного обследования. Свои важнейшие понятия лингвистическая статистика заимствует у математической статистики. Существенно обратить внимание на то, что просто количественный подсчет того или иного явления в нескольких или даже в большом числе текстов статистическим не является. Корректное применение статистики требует серьезного с ней ознакомления.

Основным методом применения статистики в сочетании с дистрибутивным анализом следует признать дистрибутивно-статистический анализ, как он представлен в трудах А.Я. Шайкевича и Ю.Д. Апресяна. Их методика имеет много общего с валентностным анализом, как он разработан Г. Хельбигом, а в Ленинграде – Б.М. Лейкиной...., термин «валентность» тоже означает сочетательную способность лингвистического элемента (Arnold I.V., 1991: 41).

Одним из широко применяемых методов для изучения распределения элементов в тексте является дистрибутивно-статистический анализ (ДСА), который характеризуется максимальной объективностью. Целью данной статьи служит выявление степени семантической близости слов тематической группы «образование». Объект исследования – 7 лексических единиц, относящихся к этой группе: education, knowledge, school, teacher, learn, study, teach. Материалом послужили тексты художественной и научной прозы современного английского языка общим объемом в 60 словоупотреблений.

Для исследования семантических связей слов группы методом ДСА все выписанные словосочетания из текстов распределяются по лексическим группам (ЛГ) с общим семантическим признаком. Был установлен средний порог частоты сочетаемости слов: для существительных – 20, для глаголов – 10.

Эксперимент

Существительными, имеющими такой порог частоты, являются teacher, school, education, knowledge, глаголами – teach, study, learn, выде-

ленными из корпуса словосочетаний с ядерным именным компонентом, являются следующие:

1. существительные – названия лиц: a teacher to smb.;
2. существительные – названия предметов: diploma of education;
3. существительные абстрактные: knowledge of love;
4. существительные: a teacher to Adele;
5. существительные, обозначающие единицы измерения: afternoon school;
6. прилагательные со значением качества: good education;
7. глаголы, обозначающие действие: to go to school;
8. существительные, обозначающие названия учреждений: a teacher at school;
9. глаголы, обозначающие речевую деятельность: to speak to the teacher;
10. глаголы, обозначаюшие чувства: to enjoy education;
11. глаголы, обозначаюшие существование: to be a teacher;
12. глаголы обладания: to acquire knowledge;
13. глаголы восприятия: to see the school.

Для глагола были выделены следующие ЛГ:

1. Существительные – названия лиц: to teach a person;
2. Существительные – абстрактные: to study labours;
3. Наречия, обозначаюшие качество: to teach well;
4. Глаголы, обозначающие чувство: to teach to love;
5. Существительные, обозначающие названия предметов: to study a book;
6. Собирательные существительные: to study the crowd;
7. Наречия, обозначающие количество: to learn much;
8. Глаголы, обозначающие речевую деятельность: to learn to speak;

Количественные данные о распределении ЛГ представлены в таблицах 1 и 2. В верхнем горизонтальном столбце даны номера ЛГ, с которыми сочетаются анализируемые слова. Цифры в вертикальных столбцах указывают на количество сочетаний.

Таблица 1 – ЛГ существительных

слова	1	2	3	4	5	6	7	8	9	10	11	12	13
teacher	40	35	28	49	20		40	30	51		29		
school	43	79	20		21	21		92	28		20		20
knowledge		60	49				30					91	
education	35	25	31	44			26			42			

Таблица 2 – ЛГ глаголов

слова	1	2	3	4	5	6	7	8
teacher	20	10			15		19	
study	38	19	18	21			10	19
learn	12	15			40	10	12	

Наличие одной и той ЛГ у двух или более анализируемых слов свидетельствует о наличии общего компонента в их значении. Выделенные ЛГ слов дистрибуции можно рассматривать как парадигматические компоненты значений, которые реализованы в систематическом плане целыми группами слов с общим компонентом.

Для измерения степени семантической связи пар анализируемых слов используем формулу коэффициента корреляции (2), с помощью которого устанавливается зависимость данного явления, т.е. значение рассматриваемых слов от каких-либо других факторов, например, от их дистрибуции. Формула коэффициента корреляции имеет вид:

$$r = \frac{\sum(x1 - x)(y1 - y)}{\sqrt{(x1 - x)2}\sum\sqrt{(y1 - y)2}}$$

где r – коэффициент корреляции (показатель тесноты связи по сравниваемым признакам), x – числовое выражение ЛГ слов дистрибуции второго языкового элемента, сравниваемого с первым, и x, y – средние величины частоты словоупотреблений выделенных лексических групп.

Коэффициент корреляции изменяется в пределах от – I до + I. Чем ближе значение коэффициента к одному из этих пределов, тем теснее связь между переменными. Если значение коэффициента выражается положительным числом, то это означает, что связь между двумя переменными прямо пропорциональна, т.е. при увеличении независимой переменной увеличивается зависимая переменная. При отрицательном значении коэффициента зависимость между двумя переменными обратно пропорциональна: при росте независимой переменной зависимая переменная уменьшается. Чем ближе значение коэффициента к нулю, тем меньше связь между переменными. Степень корреляции между величинами находит свое отражение в соответствии изменений количественных показателей этих величин. В нашем случае зависимыми переменными являются сравниваемые слова, а независимости – ЛГ словоупотреблений их дистрибуции.

Процедуру измерения степени семантической близости покажем на примере пары слов school-education (табл.). Графы таблицы заполняются параметрами, характеризующими исследуемые величины. Корреляционные таблицы дают возможность получить первое общее представление о характере связи, а также отчасти и о степени тесноты.

Первая графа таблицы заполняется цифрами, обозначающими ЛГ, характерными для всех исследуемых существительных. Графа x заполняется цифрами, отражающими частоту сочетаемости ЛГ со словом school, графа y – цифрами, отражающими частоту сочетаемости ЛГ со словом education. Остальные графы таблицы представляют собой компоненты, на которые раскладывается формула коэффициента корреляции для облегчения подсчетов.

Таблица 3

ЛГ	x	y	$x_1.x$	$y_1.y$	$(x_1.x)^2$	$(y_1.y)^2$	$(x_1.x)\,(y_1.y)$
1.	43	35	16,54	16,77	273,57	281,23	277,37
2.	79	25	52,54	6,77	2760,45	45,83	355,69
3.	20	31	-6,46	12,77	41,73	163,07	82,49
4.	21	0	-5,46	18,23	29,81	888,64	99,53
5.	21	0	-5,46	18,23	29,81	888,64	99,53
6.	0	26	26, 46	7,77	700,13	60,37	-205,209
7.	92	44	65,54	25,77	4295,49	664,09	1688,96
8.	0	0	26,46	18,23	700,13	888,64	482,36
9.	28	0	26,46	18,23	700,13	888,64	482,36
10.	0	42	26,46	23,77	700,13	565,01	-628,94
11.	20	34	-6,46	15,77	41,73	248,69	-101,87
12.	0	0	26,46	18,23	700,13	888,64	482,36
13.	0	0	-6,46	18,23	41,73	888,64	117,76
	344	237			1807	7224	2124

x = 26,46 y = 18,23

$$x = \frac{2124}{\sqrt{1807 * 7283}} = x = \frac{2124}{\sqrt{13160381}} x = \frac{2124}{3627} = 0,58$$

Результат и обсуждение

Коэффициент корреляции всех пар существительных выражается положительным числом: 1. school – education = 0,582. 2. teacher- knowledge = 0,5 3. teacher-school = 0,5 4. school-knowledge = 0,33 5. teacher-education = 0,03 6. knowledge-education = 0, 25.

Степень связи между анализируемыми глаголами: 1. Teach- learn = 0, 662. Teach-study = 0, 443. Study-learn = 0, 08.

В этом случае наблюдается обратно пропорциональная зависимость, что объясняется различием ЛГ слов дистрибуции study и learn.

На основании полученных коэффициентов корреляции строится схема связей между словами, относящихся к тематической группе «образование» (рис. 1., 2.). Принимает следующую градацию связи: r < 0,3 – слабая связь, 0, 3 <r<0,5-средняя, 0,5<r<0,7 – сильная. Сильная связь зафиксирована между парами существительных: school-education, teacher-knowledge, teacher-school.

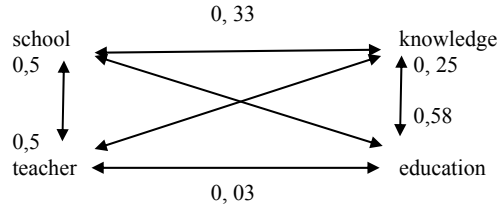

Рисунок 1 – Семантические связи существительных на основании коэффициента корреляции

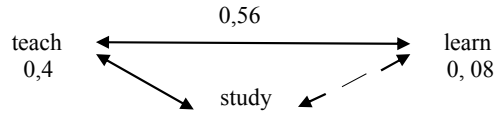

Рисунок 2 – Семантические связи глаголов на основании коэффициента корреляции

Сильная связь наблюдается также между парами глаголов teach-learn, teach-study.

Вторым этапом анализа является сравнение результатов дистрибутивно-стилистического анализа с данными толковых словарей. Эта необходимость вызывается прежде всего тем, что на основе дефиниций нельзя установить степень семантической близости слов, тогда как данные дистрибутивного анализа позволяют измерить величину этой связи. Рассмотрим семантические связи существительных, подвергнутых ДСА, которые были зафиксированы в толковых словарях /I-5 /.

school → teaching (3), learning (2), education (2), training (1), instruction (3), college (2), university (2), student (1), knowledge (2), examination (1), high school (1), classroom (1), degree (1), study ^{n} (1) (24 связи).

knowledge → learning (1), study (1), instruct (1).

teacher → instruction (1), school (1), insctruct (1), teach (1) (4 связи).

Education → learning (1), study ^{n} (1), teaching (1), instruction, scholastic (1), schooling (1), training (1).

Стрелка указывает на направление связи. Цифры в скобках указывают, сколько раз данное слово использовалось в толковании. Предполагаем: если пара слов имеет хотя бы одну общую связь, то слова этой пары семантически связаны между собой. Наибольшее количество связей зафиксировано у пары school – education (5). Второй по количеству общих связей является пара school-knowledge (3) и третьей – knowledge – education (2). Глаголы teach, study, learn имеют следующие связи согласно словарным толкованиям:

teach → instruction (1), lesson (1), student (1), pupil (1), instruct (1), knowledge (2), teacher (1), educate (1), school v (1) (10 связей); study → student (1), college (1), study ^{n} (1), learning (2), university (1), professor (1), master (1), educational (1). learn → knowledge (1), study ^{n} (1).

Заключение

У анализируемых пар глаголов были зафиксированы следующие общие связи: I. Teach – learn → knowledge; 2. Teach – study → student; 3. Study – learn → study ^{n}. Все глаголы имеют только по одной общей связи.

Семантические связи исследуемых пар существительных и глаголов представлены на рис. 3. Слова, заключенные в прямоугольники, являются общими в толкованиях каждой пары слов, подвергаемых анализу.

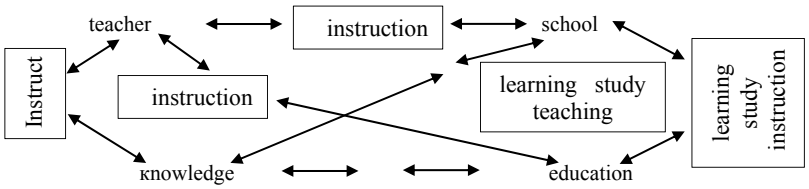

Рисунок 3 – Семантические связи глаголов на основании словарных толкований

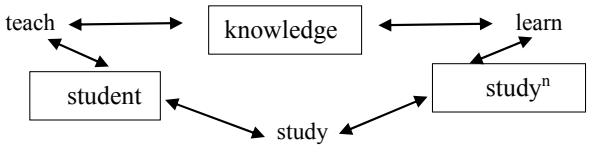

Рисунок 4 – Семантические связи существительных на основании словарных толкований

Таким образом, ДСА позволяет дать количественную интерпретацию степени семантической близости пар слов, подвергающихся исследованию. Данные о семантических связях слов, полученные в результате анализа словарных толкований, подтверждаются результатами ДСА. Эти результаты позволяют более четко представить положение анализируемых слов в лексической подсистеме, а также выделить из группы родственных по значению слов лексемы, наиболее полно воплощающие в себе семантические свойства всей исследуемой лексической группы, и определить величину этих свойств. Метод ДСА позволяет расширить исследование в области метафоризации метеорологической лексики.

Литература

Виноградов В.В. Основные типы лексических значений слова. Лексикология и лексикография: Избранные труды, М. 1977 г. – 160 с.

Арнольд И.В. Основы научных исследований в лингвистике: Учеб. пособие. – М.: Высш. шк., 1991 г. – 140 с.

Concise Oxford Russian Dictionary (ENG-RUS/RUS-ENG) Marcus Wheler and Boris Unbegaun, Oxford University Press, 1998 y., –1007 p.

Oxford Advanced learner`s dictionary New 9th edition, Oxford University Press, 2015 y., – 1820 p.

Oxford dictionary of English idioms 3rd edition edited by John Ayto, Oxford University press, 2009 y., – 408 p.

Oxford Collocations Dictionary for students of English, 2nd edition, Oxford University Press, 2009 y., – 963 p.

References

Vinogradov V.V. (1997). Osnovnye tipy lexicheskih zhnacheniy slova. [The main types of lexical meaning of the word]. (Vinogradov V.V Lexikologiya i lexikografiya: Izbrannye trydy. M., – 160 p. (In Russian)

Arnold I.V. (1991). Osnovy nauchnyh issledovaniu v lingvistike. [Fundamentals of research in linguistics] Uchebnoe posobie. – M.: Vyshaya shkola, – 140 p. (In Russian)

Concise Oxford Russian Dictionary (1998) (ENG-RUS/RUS-ENG), by Marcus Wheler and Boris Unbegaun, Oxford University Press, 1007 p. (In English)

Oxford Advanced learner`s dictionary (2015), New 9th edition, Oxford University Press, 1820 p. (In English)

Oxford dictionary of English idioms (2009), Third edition edited by John Ayto, Oxford University press, 408 p. (In English)

Oxford Collocations Dictionary for students of English (2009), 2nd edition, Oxford University Press, 963 p. (In English)