УДК 81'276.6

Неврева М. Н.,
кандидат филологических наук, доцент,
доцент кафедры иностранных языков
Одесского национального политехнического университета

Лебедева Е. В.,
старший преподаватель кафедры иностранных языков
Одесского национального политехнического университета

Гвоздь О. В.,
старший преподаватель кафедры иностранных языков
Одесского национального политехнического университета

Ершова Ю. А.,
старший преподаватель кафедры иностранных языков
Одесского национального политехнического университета

АНГЛИЙСКАЯ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ ТЕХНИЧЕСКОЙ СПЕЦИАЛЬНОСТИ «ХИМИЧЕСКОЕ МАШИНОСТРОЕНИЕ» (ЧАСТОТНЫЙ СЛОВАРЬ)

Аннотация. Статья описывает процедуру формирования вероятностно-статистической модели (частотного словаря) английской специальности «Химическое машиностроение», включенной в научно-технический дискурс. Наличие такого типа модели позволяет проводить всесторонний анализ любых текстовых единиц специальности «Химическое машиностроение» и получать объективные и надежные результаты.

Ключевые слова: частота, надежность, текстовый корпус, семантическое пространство, экспертная оценка.

Постановка проблемы. Развитие компьютерных технологий способствовало не только изменению самого качества социальной деятельности человека, но и увеличило исследовательскую активность в самых разнообразных областях научного знания. Это затронуло и такую, казалось бы, очень далекую от технических процессов науку, как лингвистика, и дало мощный толчок для широкомасштабного исследовательского поиска у целого поколения ученых-лингвистов.

Отметим наиболее важные преимущества, которые приобрела лингвистика, введя компьютер в различные виды лингвистического анализа. Это, во-первых, позволило в реальном времени получать результаты, требующие обработки таких массивов текстов, с которыми обычный исследователь справиться просто не в состоянии. Во-вторых, применение компьютеров позволило избежать той субъективности и неполноты, которыми часто страдали традиционные описания.

Таким образом, компьютерные технологии, позволяя создавать текстовые корпуса, не просто ускоряют темпы исследования языка и многократно повышают их эффективность, достоверность и проверяемость, но также и дают возможность решать такие задачи, которые лингвистика предыдущих эпох практически не ставила в силу их трудоемкости или невыполнимости. Это касается, прежде всего, формирования вероятностно-статистических моделей (или иначе – частотных словарей), которые охватывают все статистические, грамматические и лексические особенности текстов той или иной области знания. В этом случае корпусный характер словарей, а также грамматики и лексики, повышает их надежность и проверяемость.

Анализ последних исследований и публикаций. В конце прошлого века наблюдался настоящий всплеск появления частотных словарей по самым разнообразным тематикам, как техническим [1; 2; 3; 5; 6; 7; 8; 9; 10; 11], так и гуманитарным [4]. Такое внимание к этой проблеме можно объяснить необходимостью создания надежной словарной базы для будущего компьютерного перевода текстов. В настоящее время, когда такая база уже создана и введена в эксплуатацию, на основе вероятностно-статистических моделей (частотных словарей) можно не только делать корректный перевод текста по любой специальности, но и анализировать речевые единицы, функционирующие в текстовых корпусах.

Наличие лексикографических ресурсов, которые предоставляют статистические данные о речевых единицах, дает возможность исследователям учитывать количественные параметры, которые оказывают важное влияние на качественные преобразования в развитии языка.

В связи с рассмотрением проблемы формирования частотных словарей нельзя не упомянуть также и лингвистов, которые, занимаясь проблемами лингвостатистики, математической лингвистики и корпусной лингвистики [12; 13; 14; 15; 16; 17; 18; 19], дают необходимые рекомендации для исследователей, рассматривающих языковые объекты с позиций как лингвистики, так и статистики.

Цель статьи – описать последовательность создания вероятностно-статистической модели (английского частотного словаря) одной из специальностей, входящих в научно-технический дискурс – «Химическое машиностроение».

Изложение основного материала. Наиболее эффективным приемом отбора лексического материала для лингвистических исследований является извлечение объекта исследования из генеральной совокупности текстов конкретной

предметной области. Такой прием, сторонником которого был еще Л.В. Щерба [20, с. 265], требует обследования больших объемов текстового материала, на основании которого можно получить достоверные и надежные результаты.

Ввиду того, что каждая предметная область достаточно сложна, она обычно недоступна прямому наблюдению и копированию. Поэтому строится аналог (модель) предметной области – семантическое пространство, которое представляет собой поле рассуждений, отражающее участок объективной действительности. Когда речь идет о научно-техническом тексте, то семантическое пространство моделирует семантику определенной области знаний (в нашем случае – это английский подъязык «Химическое машиностроение», далее – ХМ). Разбивка семантического пространства зависит от семантических позиций исследователя.

Самым сложным вопросом для лингвистов, зачастую достаточно далеких от технических областей знания в целом и английской специальности «Химическое машиностроение» в частности, является именно формирование семантического пространства, которое, как уже было сказано, фактически симулирует предметную область любого типа.

Известно, что любая техническая область знания содержит в себе несколько взаимосвязанных друг с другом разделов, без которых она не может функционировать. Чтобы сформировать список разделов, применялись следующие методы: экспертная оценка специалистов, данные журналов предыдущих изданий, которые демонстрируют самые последние данные о той или иной специальности. Такие методы помогают представить долевое распределение участков семантического пространства.

Итак, было определено, что семантическое пространство английской области знания «Химическое машиностроение» состоит из четырех разделов, имеющих определенное процентное соотношение, которое также учитывалось после опроса специалистов:

Процессы и аппаратура для химических технологий – 30%

Конструирование машин для химического производства – 35 %

Коррозия химической аппаратуры – 25%

Общая химическая технология – 10%.

Определение доли каждого из разделов в общем объеме текстового корпуса имеет большое значение, поскольку в соответствии с этими долями подсчитывалось количество словоупотреблений, включенных в каждый раздел.

Компилирование текстового корпуса производилось по следующим принципам:

– строгая отнесенность текстов корпуса к определенной отраслевой литературе. В нашей статье это – научно-технические тексты по химическому машиностроению;

– хронологическая ограниченность текстового материала, из которого формировался корпус;

– законченность текстов статей, входящих в корпус, независимо от длины в словоупотреблениях. Нередко при составлении корпуса его формируют не из законченных текстов, а из отрывков с определенным количеством словоформ, например – 1000. Такую методику предложил известный ученый Н.Д. Андреев [12, с. 62]. Однако большинство лингвистов не придерживается этого способа компилирования текстового корпуса, поскольку при таком подходе часть важной информации о составе лексикона специальности и стратификации лексических групп теряется.

– достаточность объема текстового корпуса для получения статистически надежного исходного материала.

В соответствии с указанными принципами формирования текстового корпуса материалом, из которого получены объективные сведения о специальности ХМ, послужили тексты из английских и американских журналов за последние 10 лет: Chemical Engineering, Chemical Processing, Chemical Engineering Progress и др. Для компиляции корпуса использовался метод сплошного обследования текстов. Объем был ограничен длинной 200 тыс. текстовых единиц, которая позволила получить практически полный инвентарь лексических единиц, употребляемых в подъязыке, т.е. является достаточным для создания репрезентативного (надежного) частотного словаря.

Далее, для формирования содержания будущей вероятностно-статистической модели английского подъязыка «Химическое машиностроение» подсчитывалось количество словоупотреблений в виде словоформ, встречающихся в текстах. Под словоупотреблениями понималась любая последовательность букв, ограниченная двумя пробелами. К учитываемым единицам были отнесены все самостоятельные и служебные слова, числительные, написанные словами, общепринятые сокращения, а к не учитываемым – имена собственные, математические символы, формулы, иноязычные вставки. В дальнейшем все словоформы объединялись в словарную единицу.

Анализ отобранных текстов предусматривал решение следующих задач:

1. Составление алфавитного рангового списка всех словоформ текста.

2. Составление частотного списка, в котором все словоформы располагались в порядке убывания их абсолютных частот.

3. Сведение всех словоформ полученного частотного словаря в основные словарные единицы.

4. Вычисление статистических параметров по каждому слову и выявление некоторых общих лингвистических и информационных закономерностей исследуемого корпуса.

При формировании алфавитных ранговых списков применен способ маркировки, позволяющий разграничивать лексико-грамматическую и грамматическую омографию. Система маркировки предусматривает использование кодов-индексов, выраженных в латинском алфавите. Например, словоформы различались на уровне класса слов: pump – существительное и pump – глагол; to – предлог и to – частица; личные и неличные формы глагола: worked – Past Indefinite и worked – Past Participle; разные функции употребления глаголов to have, to be, should, would.

После получения списков каждая словоформа фиксировалась отдельно с учетом маркировки. Когда все словоформы были занесены в базу данных компьютера, разные формы одного слова объединялись, их частоты обобщались, и составлялся вторичный список слов в порядке убывания абсолютной частоты каждой единицы. Эти данные послужили исходным материалом для английского частотного словаря специальности «Химическое машиностроение». Все слова частотного списка снабжались следующими статистическими характеристиками: абсолютная частота, с которой речевая единица функционирует в текстах специальности; абсолютная накопленная частота, которая суммирует все предыдущие частоты; относительная частота, позволяющая производить сравнительный анализ с единицами текстовых корпусов, которые имеют другой объем текстового материала; относительная накопленная частота.

Статистические параметры по каждому слову вычислялись следующим образом.

1. F – количество словоупотреблений слова во всей выборке, например, system – 811;

2. F* – сумма абсолютных частот в процессе накопления, например, system 811 + 69 841, т.е. 70 652;

3. f – отношение абсолютной частоты данного слова к длине всей выборки, например, слово system (F = 811); чтобы вычислить относительную частоту, 811 : 200 000 = 0,00405; величина вычислялась до пятого знака включительно;

4. f* – сумма всех предыдущих относительных частот плюс относительная частота данного слова, например, f* слова system 0,00405 + 0,34921= 0,35326.

Полученный английский частотный словарь специальности ХМ содержит 6589 разных слов, которыми любой из текстов корпуса ХМ покрывается на 95%; 5% текста покрываются неучтенными словами: именами собственными, математическими символами, формулами и т.п.

Наличие английской вероятностно-статистической модели специальности ХМ позволяет провести предварительный анализ составляющих текстового корпуса. Все словоформы, которые вошли в словарь, были рассмотрены с точки зрения их количественных характеристик. Таблица, приведенная ниже, демонстрирует полученные данные и процентное соотношение по количеству у разных словоформ в текстовом корпусе «Химическое машиностроение».

Словоформы, функционирующие в текстовом корпусе ХМ

№№ п/п	Словоформа	Количество единиц	Доля от общего количества словоформ
1.	Существительные	3080	46,7%
2.	Прилагательные	1358	20,6%
3.	Глагол	1181	17,9%
4.	Наречие	482	7,3%
5.	Причастие прошедшего времени	261	4%
6.	Причастие настоящего времени	82	1,3%
7.	Герундий	39	0,6%
8.	Числительное	36	0,54%
9.	Предлог	41	0,6%
10.	Местоимение	25	0,4%
11.	Артикль	2	0,03%
12.	Частица	2	0,03%
Всего		6589	

В таблице вся номенклатура словоформы расположена по убыванию их абсолютных частот, что представляет иерархический порядок речевых единиц, который существует в исследуемом текстовом корпусе. Факт учета словоформ дает возможность более подробно описывать формы и функции любой части речи.

Выводы. Результаты проведенного исследования позволяют сделать следующие выводы.

1. Английская вероятностно-статистическая модель специальности «Химическое машиностроение» обладает всеми характеристиками надежности, поскольку при формировании текстового корпуса этой специальности были соблюдены все условия, необходимые при компилировании репрезентативного текстового корпуса.

2. Частотный список учитывал не только статистические, но и грамматические и лексические параметры входящих в него единиц.

3. При формировании частотного словаря специальности ХМ был использован список маркеров, которые снимали грамматическую и лексическую омонимию и указывали функцию (если это касалось глагола или глагольной словоформы) или семантическое значение (у других частей речи).

4. Наличие частотного словаря позволяет представить реальную иерархию составляющих английского текстового корпуса специальности ХМ, что дает возможность проводить всесторонний исследовательский анализ и получить корректные результаты.

Описанная в статье вероятностно-статистическая модель может служить основой для дальнейших исследований единиц текстового корпуса этой специальности, а также дает возможность проводить сравнительный анализ с единицами текстовых корпусов других специальностей научного дискурса. В настоящее время изучается словообразовательная типология имен существительных этого подъязыка.

Литература:

1. Алексеев П.М., Турыгина Л.А. Частотный англо-русский словарь-минимум газетной лексики. М.: Воениздат., 1974. 280 с.
2. Алексеев П.М. Частотный англо-русский словарь-минимум по электронике. М.: Воениздат, 1971. 302 с.
3. Гурова Н.В. Частотный словарь английского подъязыка металлургии. Л.: Ленингр. пед. ин-т., 1973. 102 с.
4. Басовская Г.В., Вербицкий А.Л. Лексико-терминологические материалы для чтения текстов по психологии на английском языке. Л.: Ленингр. пед. ин-т., 1980. 81 с.
5. Сутягина Л.М. Учебно-терминологические материалы для чтения текстов по математике на английском языке. Л.: Ленингр. пед. ин-т., 1982. 86 с.
6. Убин И.И. Англо-русский частотный словарь по электронике. М.: Moscow, M: ВПШ, 1977. 217 с.
7. Томасевич Н.П. Методические указания по работе с английской специальной лексикой для студентов специальности «Автомобилестроение» (частотный словарь-минимум). Одесса: изд-во ОПИ, 1983. 80 с.
8. Неврева М.Н. Методические указания по работе с английской специальной лексикой для студентов специальности «Химическое машиностроение» (частотный словарь-минимум). Одесса: Изд-во ОПИ, 1984. 80 с.
9. Кочеткова В.К. Частотный французско-русский словарь-минимум по электронике. М.: Воениздат, 1975. 158 с.
10. Дьяченко Г.Ф., Фалькова В.Ю. Методические указания по работе с английской специальной лексикой для студентов специальности «Акустика и ультразвуковая техника» (частотный словарь-минимум). Одесса: Изд-во ОПИ, 1985. 60 с.
11. Нелюбин Л.Л. Частотный англо-русский военный словарь-минимум (подъязык штабных документов Армии США). М.: Воениздат,1974. 224 с.
12. Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л.: Наука, 1967. 404 с.
13. Алексеев П.М. Статистическая лексикография (типология, составление и применение частотных словарей). Л.: Ленингр. гос. пед. ин-т им. А.И. Герцена, 1975. 120 с.
14. Береснев С.Д. Что такое научный функционально-речевой стиль. Иностранные языки в школе. 1961. № 6. С. 89–101.
15. Береснев С.Д. Исследование лексики немецких научно-технических текстов с позиции получателя речи: автореф. дис. … д-ра. филол. наук. Л.: АН СССР. Ин-т языкознания. Ленингр. отд-ние, 1974. 35 с.
16. Пиотровский Р.Г., Ястребова С.В. Выступление на совещании по лингвистическим проблемам научно-технической терминологии.

Лингвистические проблемы научно-технической терминологии. М.: Наука, 1970. С. 212–217.
17. Захаров В.П. Корпусная лингвистика: уч.-метод. пособие. С-Петербург: СПбГУ, 2005. 48 с.
18. Piotrovsky Rajmund G., de Gruyter Walter. Quantitative Linguistics An International Handbook. 2005. 1027 p. edited by Reinhard Köhler, Gabriel Altmann.
19. Krishnamurthy Ramesh Corpus Lexicography. Birmingham: Aston University. Elsevier Encyclopedia of Language and Linguistics. 2nd Edition. URL: https://www.researchgate.net/publication/291110989_Corpus_Lexicography
20. Щерба Л.В. Языковая система и речевая деятельность. Л.: Наука, 1974. 425 с.

Неврева М. М., Лебедева О. В., Гвоздь О. В., Єршова Ю. А. Англійська ймовірносно-статистична модель технічної спеціальності «Хімічне машинобудування» (частотний словник)

Анотація. Стаття описує процедуру формування ймовірнісно-статистичної моделі (частотного словника) англійської спеціальності «Хімічне машинобудування», яку включено в науково-технічний дискурс. Наявність такого типу моделі дозволяє проводити всебічний аналіз будь-яких текстових одиниць спеціальності «Хімічне машинобудування» і отримувати об'єктивні та надійні результати.

Ключові слова: надійність, семантичний простір, текстовий корпус, частота, експертна оцінка.

Nevreva M., Lebeleva E., Gvozd J., Ershova Yu. The English probabilistic-statistical model of the technical specialty «Chemical Engineering» (frequency dictionary)

Summary. The article describes the procedure of the probabilistic-statistical model (frequency dictionary) formation of the English specialty «Chemical Engineering» included in the scientific and technical discourse. The presence of this type of model allows for a comprehensive analysis of any text units of the specialty «Chemical Engineering» and to obtain objective and reliable results.

Key words: expert assessment, frequency, reliability, text corpus, semantic space.