AUTOMATIZATION OF COMPOSITION OF TEACHING VOCABULARY-MINIMUM ON THE BASIS OF EXTRACTION OF TERMS FROM THE CERTAIN ARRAY OF TEXTS: ANALYSIS OF CONDITION AND WAYS OF SOLUTION

V. Serbin, Candidate of Technical science, Head of the laboratory
Yu. Smirnova, Candidate of Education, Associate Professor
Almaty Institute of Power Engineering and Telecommunications,
Kazakhstan

The author provides the analysis of the problem of automatic compiling of frequency dictionary for teaching LSP. The mechanism of extraction of terms from specialized professional texts is presented.

Keywords: electronic dictionary, terms, LSP, scientific and technical text.

Conference participants, National championship in scientific analytics, Open European and Asian research analytics championship

АВТОМАТИЗАЦИЯ СОСТАВЛЕНИЯ УЧЕБНОГО СЛОВАРЯ-МИНИМУМА НА ОСНОВЕ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ ОПРЕДЕЛЁННОГО МАССИВА ТЕКСТОВ: АНАЛИЗ СОСТОЯНИЯ И ПУТИ РЕШЕНИЯ

Сербин В.В., канд. техн. наук, руководитель лаборатории
Смирнова Ю.Г., канд. пед. наук, доцент
Алматинский университет энергетики и связи,
Казахстан

В статье представлен анализ проблемы автоматизации частотного учебного словаря-минимума. Предложен механизм извлечения терминов из определённого массива текстов.

Ключевые слова: электронный словарь, термины, ЯСЦ, научно-технический текст.

Участники конференции, Национального первенства по научной аналитике, Открытого Европейско-Азиатского первенства по научной аналитике

Непременным атрибутом жизни современного человека является регулярное использование словаря: будь то электронный словарь в смартфоне, тезаурус текстового редактора «Microsoft Word» или программа проверки орфографии, ориентированная на использование орфографического словника, работа с которой ощущается даже тогда, когда пишем e-mail или sms. Словарь стал незаменимым аксессуаром, когда мы находимся в заграничной поездке. И безусловно, неоспорима ключевая роль словаря при освоении того или иного языка. Традиционно лексиконом №1 в этом процессе становится словарь-минимум, скомпилированный по частотному принципу, причем ориентируются на «нижний порог» употребительности, то есть подтверждение единицы тремя источниками[1]. В целом же словари могут дифференцироваться по разным основаниям: по характеру отображаемой информации (лингвистические и энциклопедические), по типу носителя (от древних словарей на глиняных табличках до современных электронных), − характер оснований классификаций зависит преимущественно от вкуса и личных предпочтений лексикографа. Строгая классификация словарей, по определению В.М. Лейчика, вообще вряд ли возможна[2]. Нередко дискуссию вызывает само определение словаря в контексте того или иного лингвистического направления[3], а также характер построения современных лингвистических словарей и их типология, особенно жаркие споры разгораются вокруг отличительных признаков толкового словаря от тезауруса или глоссария[4], а также вокруг концепции идеального словаря[5]. В этой статье нас будут интересовать прежде всего современные учебные словари-минимумы научно-технических терминов. Это весьма актуальное для сегодняшнего дня направление – научно-техническая терминография. И рост интереса к ней связан преимущественно со следующими двумя факторами:

1. взрывообразное развитие современной научно-технической терминологии, обусловленное развитием новых технологий (компьютеры, электроника, связь etc.);

2. заметное (более чем в 10 раз) численное преобладание терминологических словарей по сравнению с другими видами лексиконов[6].

Первый фактор влечёт за собой ряд следствий методического характера. Так, научно-технические тексты, отобранные для учебных целей и предназначенные для занятий по русскому языку в студенческой аудитории, довольно быстро перестают быть актуальными и уже не мотивируют учащихся к изучению языка как инструмента профессиональной коммуникации, поскольку исчезает элемент новизны в текстах, соответственно снижается интерес к предмету. Учебный словарь-минимум (терминологический, двуязычный, энциклопедического типа, комплексный или просто словник – любая из этих разновидностей) традиционно составляется по частотному принципу на основании учебных текстов, таким образом, он устаревает тоже. При изменении набора текстов, используемых в терминографических целях, закономерно меняется содержание массива терминологических единиц и происходит их частотное перераспределение, то есть меняется содержание словаря-минимума, иногда радикальным образом.

Второй фактор демонстрирует необходимость наличия такого рода словарей. Но, с другой стороны, процесс компилирования словаря-минимума, отбора для него лексических единиц – трудоёмкое и долгое занятие[7], которое практикующему русисту приходится делать довольно часто. Поэтому актуализируется проблема автоматизиро-

1 Кудашев И.С. Проектирование переводческих словарей специальной лексики. – Хельсинки, 2007. – С. 172.
2 Лейчик В.М. Опыт построения классификации терминологических словарей // Теория и практика научно-технической лексикографии. – М., 1988. – С.4.
3 Pius ten Hacken. What is a Dictionary? A View from Chomskyan Linguistics // International Journal of Lexicography. Volume 22, issue 4, 2009. – P.399-421.
4 Stricker S. Glossary-Vocabulary-Dictionary and the Question of their Differentiation // Sprachwissenshaft. Volume 36, issue 2-3, 2011. – P.115-144.
5 Abecassis M. The Ideology of the Perfect Dictionary: How Efficient Can a Dictionary Be? // Lexicos. Volume 18, 2008. – P.1-14.
6 Гринёв-Гриневич С.В. Введение в терминографию. Как быстро и легко составить словарь. – М., 2009. – С.10.
7 Susniene D., Vibrickaite R. Toward a Systematic Dictionary: Compiling a Glossary of Terms // International Conference on Nation and Language "Nation and Language: Modern Aspects of Socio-Linguistic Development Proceedings", 2008. – P.110-114.

ванного поиска (выборки) терминов из конкретного набора актуальных научно-технических текстов.

На сегодняшний день существует неавтоматизированный гибридный метод отбора терминов из аутентичных текстов, в котором частично используются лексическая классификация, анализ ключевых слов, терминоизвлечение и систематическое классифицирование[8]. Тем не менее этот метод не решает проблемы автоматизированного отбора терминов, поскольку лексическая классификация терминов в автоматизированных системах, как будет показано далее, в силу объективных причин затруднена.

Мы предлагаем разработанный механизм поиска терминов из научно-технического текста, отличающийся системой исправления и распознавания запросов, а также системой ранжирования и смешивания результатов, т.е. наборов терминов, полученных из разных текстов. Этот механизм, как и большинство современных онлайновых лингвистических ресурсов (Машинный фонд, Национальный корпус русского языка и т.д.), предполагается использовать в Интернете.

Алгоритм поисковой системы состоит из 3 систем. Это следующие системы: Система распознавания и исправления запросов, Система кеширования результатов, Система получения и смешивания результатов, поиск. На рисунке 1 представлена общая модель автоматизации составления учебного словаря-минимума на основе извлечения терминов из определённого массива текстов.

При смешивании результатов соблюдаются следующие требования:

– результаты со всех трех поисковых систем являются равноправными;

– учет, уничтожение дубликатов и ранжирование результатов производится по доменному имени “site.kz”, а не по полному url-адресу, т.к. поисковые системы запоминают конкретные страницы веб-сайтов с разными переменными и префиксами;

– результаты с поисковых систем берутся пакетами по 10-20-30 и т.д. адресов, но с Google – 8-16-24 и т.д. Ограничение Google API.

Все получаемые с поисковых систем результаты условно делятся на три группы и при смешивании каждая группа следует за предыдущей по порядку:

– адрес сайта (доменное имя) встречается в трех поисковых системах;

– адрес сайта (доменное имя) встречается в двух поисковых системах;

– адрес сайта (доменное имя) встречается в одной поисковой системе.

В каждой группе результаты выстраиваются по принципу:

– чем меньше сумма мест каждого отдельного результата в каждой поисковой системе, тем выше этот адрес при ранжировании в смешанных результатах.

Результаты в первой группе. Если страница сайта “a” во всех трех поисковых системах заняла первые места, т.е. её сумма мест будет равна 3, что является максимально низкой суммой мест, то и при ранжировании в смешанных результатах она займет первое место. Если страница сайта “b” во всех трех поисковых системах заняла вторые места, сумма мест будет равна 6 и при ранжировании в смешанных результатах она займет второе место. Если страница сайта “c” во всех трех поисковых системах заняла третьи места, сумма мест будет равна 9 и при ранжировании в смешанных результатах она займет третье место.

Результаты во второй группе, которая следует за первой. Если страница сайта “a” в двух любых поисковых системах (при условии отсутствия в третьем поисковике) заняла первые места, т.е. её сумма мест будет равна 2, что является максимально низкой суммой мест, то и при ранжировании в смешанных результатах она займет первое место. Если страница сайта “b” в двух любых поисковых системах (при условии отсутствия в третьем поисковике) заняла вторые места, сумма мест будет равна 4 и при ранжировании в смешанных результатах она займет второе место и т.д.

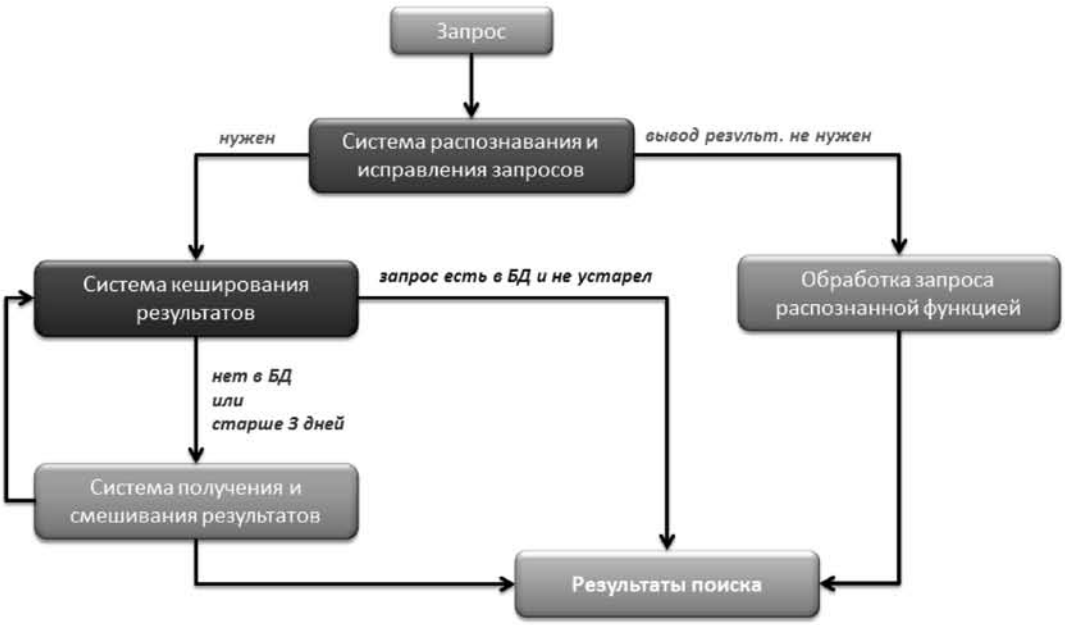

Рис. 1. Общая модель автоматизации составления учебного словаря-минимума на основе извлечения терминов из определённого массива текстов

8 Deny Armos Kwary. A Hybrid Method for Determining Technical Vocabulary // System. Volume 39, issue 2, 2011. – P.175-185.

Таким образом, при ранжировании групп и адресов сайтов (рис. 2), а также уничтожении дубликатов при смешивании результатов от всех поисковых системах, на отдельной странице выстраивается наиболее релевантная по отношению к запросу пользователя последовательность поисковых результатов.

Таким образом, разработан механизм поиска, отличающейся системой исправления и распознавания запросов, системой ранжирования и смешивания результатов.

Распознавание лексических значений неоднозначных единиц производится человеком автоматически, на основе некоторых имплицитных правил. Поскольку отправитель и адресат сообщения в одинаковой степени владеют этими правилами, лексическая неоднозначность не препятствует их общению. Компьютер же подобными правилами декодирования не обладает, т.е. их необходимо задать[9], то есть необходимо сформулировать запрос. И здесь возникает сложность, состоящая в принципиальной невозможности формулировки технического задания как такового. Связано это со следующими причинами.

1. Влияние человеческого фактора. Подавляющее большинство текстов (в том числе научно-технических) являются продуктом работы человеческого мозга. Поэтому в технических текстах не исключены ошибки в употреблении терминов, например, на почве паронимии: *спектр – спекл, кюри – кюрий, гало – галоид, ангидрит – ангидрид, квадрат – квадрант* и т.п., эти ошибки можно распознать, опираясь на контекст и располагая соответствующими знаниями в данной области, что в настоящее время под силу только человеку.

2. Отсутствие идеального термина. Так, термин – определяемое, т.е. субъект, обычно обозначается аббревиатурой *Dfd* (от латинского *Definiendum*), языковое выражение значения термина – определяющее, т.е. предикат, обозначается *Dfs* (от латинского *Definiens*). Установление тождества субъекта и предиката обычно обозначается формулой *Dfd* ≡ *Dfs* (≡ – знак дефиниционного тождества). Для автоматизированного извлечения термина из текста целесообразно использовать метод качественно-количественного логического анализа. В его основе лежит понимание содержания понятия как суммы всех качеств и отношений соответствующего этому понятию конкретного или абстрактного предмета. Формально это выглядит так[10]:

$$B = \sum_{i=k}^{n} \sum_{i=1}^{m} K_i (B) \Lambda \sum K (C), \qquad (1)$$

где B – максимально полное понятие о предмете;

$K_1, K_2, \ldots K_n$ – качества; $C_1, C_2, \ldots C_n$ – разнопорядковые стороны, свойства, отношения и т.п.;

$K(B)$ – качества объекта в целом;

$K(C)$ – качества его сторон, свойств, отношений и т.п.;

Λ – знак конъюнкции.

Термины, имеющие все представленные в этом выражении переменные, то есть безупречную характеристику, сравнительно редки. Более того, в научно-техническом дискурсе встречаются термины-девиации: многозначные и термины свободного использования, абсолютные терминологические синонимы-дублеты, нестандартизованные и устаревшие термины, а также профессионально-просторечные варианты терминов. Кроме того, категории терминов неодинаково распределяются текстах различных подстилей и жанров научно-технического стиля (табл. 1). Так, например, в монографиях, диссертациях, научных статьях и тезисах, проектных и конструкторских документах, в описаниях изобретений ect. будут содержаться нестандартизованные термины[11], которые невозможно выделить автоматически как раз в силу их нестандартизованности, то есть незафиксированности в тезаурусной системе компьютера.

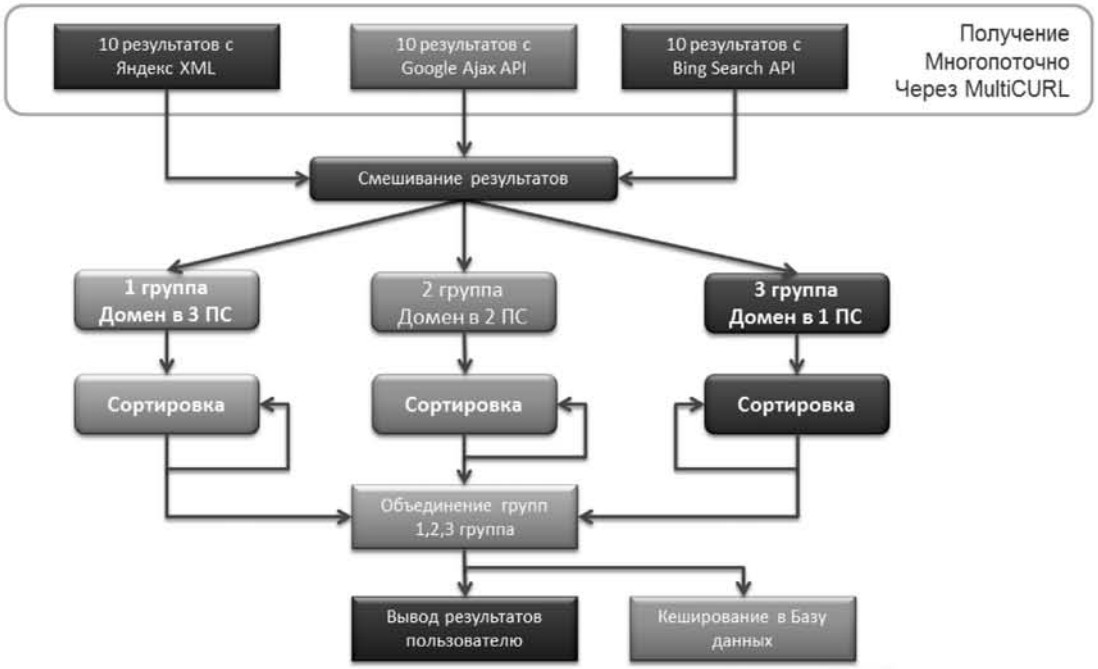

Рис. 2. Алгоритм смешивания результатов поиска

9 Трибис Л.И. Об одной модели распознавания лексических значений неоднозначных слов. В кн.: Статистика речи и автоматический анализ текста. – М., 1972. – С.131.
10 Квитко И.С., Лейчик В.М., Кабанцев Г.Г. Терминоведческие проблемы редактирования. – Львов, 1986. – С.47.
11 Там же. С.75.

Таблица 1

Матрица распределения терминов в подстилях и жанрах научно-технического стиля

Термины	Собственно научный подстиль					Техн. подстиль			Техн.-экон. Подстиль			Научно-деловой подстиль			Учебно-научный подстиль		Научно-справочный подстиль						Научно-популярный подстиль		
	Монографии	Научные статьи и тезисы	Диссертации	Обзоры и обзорные статьи	Описание открытий	Проектные и констр. документы	Технологические документы	Описание изобретений	Плановые документы	Статистические документы	Классификаторы	Организационные документы	Распределительные документы	Справочн.-инф. документы	Учебники и учебные пособия	Учебные и справочные словари	Тезаурусы	Энциклопедии и справочники	Рефераты	Аннотации	Промышленные каталоги	Инструкции	Рекламные материалы	Научно-популярные произведения	Массово-производств. издания
Стандарт.			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+
Рекоменд.	+	+	+	+	+			+						+		+	+	+	+	+		+	+	+	+
Нестандарт.	+	+	+		+			+										+	+				+	+	+

Кроме того, весьма сложной задачей даже для эксперта в области терминоведения является классифицирование слов в особой функции по их категориям: общенаучная лексика, общеспециальная нетерминологическая лексика, общеспециальная терминология, общеотраслевая терминология, терминология частных областей.

3. Терминированность, обусловленная контекстом, который не всегда может быть распознан автоматическими системами (ср.: *вода – тяжёлая вода*).

4. Отсутствие максимально глубокой семантической разметки термина, развивающего свои лингвистические характеристики, особенно это заметно в обновляющихся терминологиях (компьютерные технологии, электроника, связь ect.).

Корпусные данные (www.ruscorpora.ru) по терминам и данные, предоставляемые различными терминологическими банками, к сожалению, опираются на собственные текстотеки, которые могут использоваться в учебном процессе лишь отчасти, не в полной мере, а имеющиеся частотные словари (Алексеев П.М. Частотный англо-русский словарь по электронике. – М., 1971; Тер-Мисакянц З.Т. Частотный словарь математической лексики. – Ереван, 1973; Денисов П.Н., Морковкин В.В., Сафьян Ю.А. Комплексный частотный словарь русской научной и технической лексики. – М., 1978) не являются словарями-минимумами, более того – лексический состав многих из них требует обновления. Единственным, пожалуй, исключением на сегодня является частотный англо-русский словарь по оптоэлектронике и лазерной технике Щаповой И.А. (2011).

Таким образом, в области автоматизации компилирования учебного словаря-минимума технических терминов на основе закрытого массива текстов больше проблем, чем готовых оптимальных решений, несмотря на стремительный взлёт компьютерной лингвистики в последнее десятилетие.