ИСПОЛЬЗОВАНИЕ КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ ДЛЯ ЭКСПЕРТИЗЫ ТЕРМИНОЛОГИЧЕСКОГО СЛОВАРЯ В ОБЛАСТИ ГОСУДАРСТВЕННОГО ФИНАНСОВОГО КОНТРОЛЯ

Лукашевич Н.В.[1,2], Салий А.Д. [2], Добров Б.В.[1,2]

[1] Научно-исследовательский вычислительный центр МГУ
[2] АНО Центр информационных исследований

{louk, sali, dobroff}@mail.cir.ru

Статья описывает решения, принятые при выполнении работы по договору со Счетной Палатой РФ, по уточнению терминологического словаря финансового контроля. Использовались компьютерные технологии, такие как автоматическое извлечение терминологических словосочетаний из текстов, а также анализ документов Счетной Палаты средствами специализированной ИПС. Компьютеризация работ позволила четче отграничить предметную область, выявить ее терминологический состав, проанализировать контексты употребления терминов

1. Введение

Разработка терминологических словарей в той или иной сфере деятельности является обычно длительной и сложной процедурой. Необходимо сформулировать принципы включения терминов в словарь, принципы организации словаря, составить список терминов для включения в словарь, снабдить термины дефинициями.

Специалисты Счетной Палаты Российской Федерации, в процессе сопровождения словаря (Воронин, Мешалкина 2003) в области государственного финансового контроля (далее Словарь), столкнулись с рядом проблем.

Основным назначением словаря является выявление и обеспечение дефинициями терминов в сфере государственного финансового контроля, осуществляемого Счетной Палатой РФ. Словарь предназначен для использования специалистами Счетной Палаты РФ для стандартизации используемой терминологии при подготовке документов.

Основной проблемой, возникшей после опубликования Словаря, оказалась проблема противоречивости требований пользователей – аудиторов Счетной палаты, других специалистов Счетной Палаты по поводу того, какие именно термины должны включаться в Словарь.

В результате возникла необходимость более четкой формулировки принципов составления Словаря, обоснования его словарного состава.

Основные цели работы по выполнению экспертизы Словаря были сформулированы следующим образом:

- определение состава Словаря терминов государственного финансового контроля
- контроль дефиниций Словаря терминов по контекстам употребления терминов
- разработка рекомендаций по ведению Словаря государственного финансового контроля.

Для выполнения работы использовались такие компьютерные технологии, как автоматическое извлечение терминологических словосочетаний из текстов, а также анализ документов Счетной Палаты средствами специализированной ИПС. Компьютеризация работ позволила четче отграничить предметную область, выявить ее терминологический состав, проанализировать контексты употребления терминов, точнее описать значения многозначных терминов.

2. Принципы формирования терминологического словаря предметной области «Государственный финансовый контроль»

Предметная область государственного финансового контроля имеет междисциплинарный характер, находясь на стыке таких крупных областей как экономика, право, аудит и бухгалтерия. Наиболее ярко эта междисциплинарность проявляется в том, что тексты, созданные в рамках данной области, содержат самые разнообразные термины из этих областей. Поэтому выявление границ предметной области является достаточно серьезной проблемой.

Целесообразно использовать следующую совокупность принципов, на основе которых можно установить границы понятийно-терминологической системы данной предметной области:

Принцип 1.

Термин должен соответствовать фиксированному набору семантических и тематических типов, разработанному на основе анализа нормативных актов, регулирующих деятельность в данной предметной области.

Принцип 2.

Описание терминов предметной области должно быть системным.

Принцип 3.

Важным фактором, влияющим на включение/невключение термина в состав терминосистемы предметной области, является частотность употребления этого термина или его текстовых вариантов в текстах предметной области.

Принцип 4.

Словарные входы словаря должны быть по возможности сформулированы однозначно. Неоднозначные терминологические выражения могут использоваться как отсылочные элементы к основным словарным статьям.

2.1. Основные семантические и тематические типы терминов Словаря

На основе анализа основных нормативных документов, касающихся деятельности Счетной палатой, была выделена совокупность семантико-предметных категорий терминов, которые относятся к данной предметной области. Например:

- термины, относящиеся к этапам, процедурам, участникам процесса государственного финансового контроля;
- термины, относящиеся к бюджетной системе и бюджетному процессу;
- термины, относящиеся к области приобретения, использования и распоряжения государственной собственностью и др.

Именно термины этих категорий должны включаться в Словарь.

2.2. Системность терминологического состава Словаря

Набор терминов Словаря должен образовывать терминологическую систему, то есть термины должны толковаться либо через общезначимую лексику, либо содержать только те термины, которые имеют определения в данном Словаре.

Другим проявлением принципа системности является то, что термины, используемые в документах предметной области и принадлежащие одному и тому же классу, должны трактоваться в рамках терминологической системы схожим образом. Например, если в Словарь включено определение для термина *внебюджетные средства*, то должно быть включено и определение термина *бюджетные средства*.

2.3. Частотность употребления термина в документах Счетной палаты как фактор его включения или не включения в Словарь

Частотность употребления термина в документах Счетной палаты является важным фактором для решения по включению или не включению определения термина в Словарь:

- если термин не употреблялся в открытых публикациях Счетной палаты, то он может быть включен в словарь только, если он необходим для поддержания принципа системности словаря, то есть если в словаре имеется один или более терминов, которые требуют в своем определении ссылки на данный термин;
- включение и описание значений многозначных терминов в Словаре должны базироваться на реальной употребимости этих значений в документах Счетной палаты;
- включение текстовых терминов (текстовых вариантов терминов) должно базироваться на их реальном употреблении в документах Счетной палаты РФ.

2.4. Однозначность словарных входов

Обычно каждому понятию области соответствует хотя бы один однозначно понимаемый термин, значением которого является это понятие (Суперанская и др. 2003). В реальных текстах предметной области для ссылки на понятие помимо основных терминов может использоваться множество разнообразных языковых выражений:

- синтактико-словообразовательные варианты: *получатель бюджетных средств – бюджетополучатель*;
- лексические варианты – *безакцептное списание*, *бесспорное списание*;
- многозначные выражения, в зависимости от контекста служащие отсылкой к разным понятиям области, например, слово *валюта* в разных контекстах может означать *национальная валюта* или *иностранная валюта*.

Выявление среди совокупности терминологических синонимов основного термина, наиболее точно и адекватно представляющего понятие, уже само по себе играет важную нормализующую и стандартизирующую роль (Ахманова, 1966).

3. Исходные данные и используемые компьютерные технологии

Для работы были представлены:

- текущая версия Словаря основных терминов и понятий, применяемых в Счетной Палате Российской Федерации при осуществлении контрольно-ревизионной, экспертно-аналитической и других видов деятельности,
- публикации Бюллетеня Счетной палаты РФ (1999-2004гг.) – около 800 отдельных статей.
- материалы Бюллетеня Счетной палаты РФ были автоматически обработаны следующим образом:
- пройдя процедуру автоматической текстовой обработки они были загружены в Университетскую Информационную систему Россия (www.cir.ru). Загрузка обеспечила возможность поиска текстов по словам и словосочетаниям, что дало возможность проверять реальную употребимость в данных материалах тех или иных терминов, а также выявлять контексты их употребления.

- была произведена обработка специальными процедурами автоматического извлечения терминоподобных словосочетаний, что дало возможность дополнительных проверок употребимости терминов в материалах, а также нахождения дополнительных терминов, входящих в состав предметной области.

Для выявления терминов было проведено сопоставление с терминами Общественно-политического тезауруса (Лукашевич, Добров; 2001). Также были применены два алгоритма выделения терминоподобных слов и словосочетаний (Добров, Лукашевич, Сыромятников; 2003).

Первый алгоритм выделяет существительные, прилагательные, согласованные пары и тройки прилагательных и существительных, а также генеративные конструкции (существительное + существительное в родительном падеже и т.п.).

Второй алгоритм может выделять часто повторяющиеся именные группы в несколько слов, в том числе предложные.

Полученные терминоподобные слова и словосочетания упорядочивались по убыванию суммарной частотности и убыванию количества содержащих их документов.

Так, самыми частотными словосочетаниями, выявленными по текстовой коллекции, были следующие: *федеральный бюджет, Российская Федерация, Счетная палата, федеральный закон,•общая сумма, средства федерального бюджета, областной бюджет, денежные средства, использование средств, заработная плата, Минфин России, бюджетные средства, налоговый орган* и др.

Эксперты, двигаясь по списку сверху вниз (по убыванию частотности), рассматривали очередное терминоподобное слово или словосочетание в качестве кандидата на включение в Словарь, последовательно применяя принципы, сформулированные в п.2.

4. Анализ и редактирование состава и определений Словаря

4.1. Анализ тематического состава версий Словаря

Было выявлено, что имеется ряд высокочастотных терминов, относящихся к вышеперечисленным терминологическим категориям, которые не были включены в существующие версии словаря:

- термины процедуры контроля: *контрольное мероприятие*, *выборочная проверка* и др.*;*
- бюджетные термины: *федеральный бюджет, бюджетные средства, налоговый доход, оборонный заказ, налоговый учет, государственный контракт, территориальный бюджет, бюджетный счет* и др.;
- типы и формы организаций: *муниципальное предприятие, коммерческая организация, дочернее предприятие, государственное учреждение, холдинг* и др.

4.2. Анализ системности определений существующих версий Словаря

Были выявлены случаи нарушения принципа системности. Например**,** для термина *акциз* было представлено следующее определение

> *Акциз – косвенный налог, включаемый в цену товара (продукции)* ,

при этом термин *косвенный налог* не был определен в словаре.

Для исключения ситуаций нарушения принципа системности словаря могут использоваться следующие методы:

- переформулирование имеющегося определения для того, чтобы исключить введение дополнительных терминов;
- включение в состав Словаря термина, необходимого для определения исходного термина;
- исключение из словаря исходного термина.

Так, например, в определении Словаря

> *Случай страховой – событие, при наступлении которого в силу закона или договора страховщик обязан выплатить страховую сумму*

содержатся термины *страховщик* и *страховая сумма*, которые не были определены в данном Словаре. В результате анализа исходный словарный вход *случай страховой* исключен из словаря как несоответствующий предметной области.

4.3. Анализ состава Словаря по частотности употребления терминов

В составе словаря были выявлены следующие термины, которые не употреблялись в открытых публикациях Счетной палаты или частотность их употребления очень низка:

варрант – употребление не обнаружено
верификация - употребление не обнаружено
дефляция - употребление не обнаружено
дефолт - в 1 документе
и др.

Часть терминов, которые мало употреблялись в документах Счетной палаты РФ, были удалены из Словаря. Малочастотный термин мог быть оставлен в словаре, если для этого имелись дополнительные причины, например, термин принадлежит к основным терминам аудиторской деятельности (*небухгалтерские данные, аудиторские доказательства), термин необходим для определения других терминов словаря, термин однозначно отражает одно из значений употребительного многозначного термина и др.*

4.4. Работа с многозначными словарными входами

Текущая версия Словаря содержала 47 многозначных слов и словосочетаний в качестве словарных входов. Для всех этих многозначных выражений была проведен анализ, не существует ли однозначное выражение, совпадающее по значению с одним из значений многозначного слова.

Например, словарная статья для слова *счет* была заменена на 2 словарные статьи следующим образом:

Исходный вариант:

> *Счет – 1. Совокупность записей бухгалтерского учета, отслеживающих движение денежных средств по какому-либо конкретному направлению. (повтор) 2. Товарный документ (фактура), выписываемый продавцом на имя покупателя и удостоверяющий поставку товара или оказание услуг и их стоимость.*

Замена словарной статьи:

> *Счет – см. Счет бухгалтерского учета; Счет-фактура*

Значения некоторых неоднозначных словарных единиц не использовались в текстах предметной области, поэтому в таких случаях такие значения удалялись, а оставшиеся значения приписывались к однозначно сформулированным словарным входам. Нужно отметить, что каждый раз в качестве таких однозначных словарных входов выбирались термины, реально употребляющиеся специалистами в предметной области.

Например, исходный вариант содержал следующую словарную статью:

> *Мораторий – 1. Приостановление исполнения обязательств, устанавливаемое государством на определенный срок или до окончания каких-либо чрезвычайных событий. Распространяется как на все виды обязательств, так и на некоторые их виды или на отдельные категории должников. 2. В международном праве – договоренность государств об отсрочке или воздержании от каких-либо действий как на определенный, так и на неопределенный срок. 3. Приостановление исполнения должником денежных обязательств и уплаты обязательных платежей.*

В окончательной версии, учитывая, что два первых значения не использовались в документах проанализированного массива, получаем:

> *Мораторий на удовлетворение требований кредиторов - приостановление исполнения должником денежных обязательств и уплаты обязательных платежей.*

В некоторых случаях преобразование многозначных словарных статей в совокупность словарных статей с однозначными заголовками оказалось невозможным. Такие словарные статьи были оставлены в первоначальном виде.

Например,

> *Долг государственный внешний – 1. Обязательства, возникающие в иностранной валюте. 2. Государственный долг по непогашенным внешним займам и не выплаченным по ним процентам. Складывается из задолженности международным и государственным банкам, правительствам, частным иностранным банкам. Различают общий (накопленный) и текущий внешний долг.*

Были также выявлены случаи ложной многозначности, когда словарный вход с приписанными несколькими значениями на самом деле не является многозначным.

Для выявления таких ситуаций проводится следующая процедура. Пусть словарный вход *С* имеет два значения, описываемые выражениями *31* и *32*. Мы формулируем два вопроса:

- существует ли такая сущность *А*, которая является *31* и не является *32*,
- существует ли такая сущность *Б*, которая является *32* и не является *31*.

Далее на основе анализа документов и словарей производится поиск таких *А* и *Б*. Если привести примеров существования таких *А* и *Б* не удается, то два значения склеиваются.

Так, например, для термина *таможенный тариф* в анализируемой версии Словаря описаны следующие два значения:

> *Тариф таможенный – 1. Инструмент торговой политики и государственного регулирования внутреннего товарного рынка при его взаимодействии с мировым рынком, а также правила обложения товаров пошлинами при их пересечении через таможенную границу. 2. Свод ставок таможенных пошлин, применяемых к товарам, перемещаемым через таможенную границу данной страны.*

Между тем анализ документов показал, что эти значения не являются взаимоисключающими. Если таможенный тариф является сводом ставок таможенных пошлин, то именно как свод ставок он является инструментом регулирования и задает правила обложения товаров пошлинами. Таким образом, эти определения объединены следующим образом:

> *Таможенный тариф - свод ставок таможенных пошлин, применяемых к товарам, перемещаемым через таможенную границу Российской Федерации и систематизированным в соответствии с*

Товарной номенклатурой внешнеэкономической деятельности. Таможенный тариф рассматривается как инструмент торговой политики и государственного регулирования внутреннего рынка товаров Российской Федерации при его взаимосвязи с мировым рынком.

5. Изменение состава словаря

Исходная версия словаря включала 339 словарных статей.

В качестве отдельных словарных единиц было удалено 60 словарных входов по следующим причинам:

- слишком общие (*баланс, факт*) – 13,
- малоупотребительные – 33,
- не соответствуют ПО – 4 (*страховой случай),*
- заменены на ссылки к словарным статьям однозначных терминов– 10 (*счет, мораторий, сальдо).*

Анализ реального языка документов Счетной палаты показал, что в Словарь необходимо добавить определения достаточно большого числа терминов.

Следует подчеркнуть, что ни один из существующих словарей по экономике, праву и смежным дисциплинам не обеспечивал всей совокупности необходимых определений терминов. Количество источников новых определений для Словаря достигло 50 единиц, включая тексты законов и нормативных актов, опубликованные словари, электронные словари, научные публикации.

Сопоставление большого числа словарных источников выявило серьезные проблемы имеющихся словарей, такие как нехватка определений широко употребляющихся терминов, противоречие определений близких по смыслу терминов, несоответствие определений текущему употреблению термина.

Так, например, все обнаруженные словарные источники давали определение термина *заемные средства* как *денежные средства, либо иная имущественная ценность, которые выдаются банком в виде ссуды предприятию или учреждению во временное пользование,* что значительно уже современного употребления этого термина**.**Всего было добавлено 195 новых словарных статей, в том числе для терминов, которые достаточно редко встречаются в опубликованных словарях: а*мортизация долга, бюджетная заявка, временный кассовый разрыв, движение денежных средств, перефинансирование, разассигнование, расчетные документы, платежные документы, хозяйственный договор и др.*

Заключение

Результаты работы показали, что разработка качественного терминологического словаря в короткие сроки существенно зависит от:

- наличия программ автоматического извлечения терминоподобных словосочетаний,
- наличия представительных текстовых коллекций, загруженных в информационные системы.

Практически мы пользовались системой текстовых коллекций, состоящей из трех составных частей:

- Документы Счетной Палаты РФ,
- Законодательство РФ,
- Интернет.

Некоторые термины для качественного описания потребовали серьезного лингвистического анализа контекстов употребления, сопоставления существующих словарей.

Благодарности

Настоящее исследование частично поддержано за счет гранта Российского фонда фундаментальных исследований № 03-01-00472.

Список литературы:

1) Ахманова О.С., Словарь лингвистических терминов. Предисловие. - М. – 1966.
2) Воронин Ю.М., Мешалкина Р.Е. Стандартизация финансового контроля: Россия и мировой опыт. - Изд. дом «Финансовый контроль», 2003. – С.112-153
3) Добров Б.В., Лукашевич Н.В., Сыромятников С.В., Формирование базы терминологических словосочетаний по текстам предметной области // Пятая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Санкт-Петербург, 28 -31 октября 2003 г. – СПб.: СпбГУ – 2003. – С.201-210.
4) Лукашевич Н.В., Добров Б.В., Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // Труды международного семинара Диалог-2001. - Аксаково-2001.- с.273-279.
5) Суперанская А.В., Подольская Н.В., Васильева Н.В., Общая терминология: Вопросы теории / Отв. Ред. Т.Л.Канделаки. Изд. 2-е, стереотипное. – М.: Едиториал УРСС, 2003. – 248 с.