ОСОБЕННОСТИ ПОДГОТОВКИ ТЕРМИНОВ ДЛЯ РУССКО-АНГЛИЙСКОГО ТЕЗАУРУСА ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

Е. Г. Соколова (minegot@rambler.ru)

С. Ю. Семенова (sonya_sem@mail.ru)

Российский государственный гуманитарный университет, Москва, Россия

И. С. Кононенко (irina_k@cn.ru)

Ю. А. Загорулько (zagor@iis.nsk.su)

Институт систем информатики имени А. П. Ершова СО РАН, Новосибирск, Россия

О. Ф. Кривнова (okrivnova@mail.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В. П. Захаров (vz1311@yandex.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Описывается начальная стадия разработки русско-английского терминологического тезауруса по компьютерной лингвистике. В задачи этого этапа входит обоснованный выбор потенциальных источников терминов с учетом двуязычной специфики разрабатываемого ресурса, выделение и подбор терминов для базового словника, изучение особенностей представления терминов, их толкований и отношений между ними.

Ключевые слова: тезаурус, термины, компьютерная лингвистика, двуязычие.

SELECTION AND PREPARATION OF TERMS FOR THE RUSSIAN-ENGLISH THESAURUS OF COMPUTATIONAL LINGUISTICS

E. G. Sokolova (minegot@rambler.ru)

S. Iu. Semenova (sonya_sem@mail.ru)

RSUH, Moscow, Russian Federation

I. S. Kononenko (irina_k@cn.ru)

Iu. A. Zagorul'ko (zagor@iis.nsk.su)

A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russian Federation

O. F. Krivnova (okrivnova@mail.ru)

MSU, Moscow, Russian Federation

V. P. Zakharov (vz1311@yandex.ru)

Saint-Petersburg State University, St. Petersburg, Russian Federation

The initial phase of the development of Russian-English thesaurus on terminology in the fi eld of computational linguistics is described. One of the fi rst tasks is the choice of candidate sources of terms allowing for the bilingual nature of the electronic resource. Other problems to be solved are those of terminology extraction and selection of basic term list as well as the study of peculiarities of representation of terms and relations between them. The diversity of the fi eld of computational linguistics, its interdisciplinary nature and the lack of Russian terminological sources and term defi nitions due to certain lagging of the fi eld in Russia as compared to the English-speaking countries — all these factors explain the kind of decisions made at this stage. One of them concerns the use of the Russian-language corpus of papers presented at the International Conference "Dialogue" (2000–2010). This corpus proved to be a helpful source of terms in real use. Besides, dictionaries as well as indices and glossaries of textbooks and manuals have been examined in order to derive defi nitions. As an additional source of terms for the Russian part of the thesaurus the English-language terminological sources have been utilized and their terms and defi nitions translated into Russian. This is especially important for the terms in some empirical and technologically advanced subfi elds, such as speech technologies.

Key words: thesaurus, terms, computational linguistic, bilingualism.

1. Введение

В статье описывается начальная стадия разработки электронного терминологического тезауруса по компьютерной лингвистике (КЛ). Главной целью этого этапа является выбор терминов и их толкований для дальнейшего включения в тезаурус. Основные трудности в подборе и содержательном определении терминов для русско-английского тезауруса по КЛ связаны с особенностями самой науки и состоянием ее развития в России. Важно отметить, в частности:

1. еще не преодоленное до конца отставание русскоязычной КЛ (РКЛ) от англоязычной КЛ;
2. неоднородность предметной области (ПО) КЛ;
3. неравномерность развития отдельных направлений КЛ;
4. междисциплинарный характер КЛ.

Определенные трудности обусловлены также нашей установкой на отражение терминологии РКЛ в составе оригинальных русскоязычных работ, а не обзоров и таких учебников, которые во многом являются пересказом англоязычных источников[1]. Следствием пп. 1, 3 является отсутствие русскоязычных учебных и лексикографических источников, достаточно полно отражающих структуру современной КЛ в отличие от англоязычных источников, где она представлена детально и отчетливо. До сих пор термины РКЛ входили лишь в состав словарей и глоссариев по лингвистике и смежным направлениям. Следствием пп. 2, 3 является наличие источников только по отдельным разделам смежных направлений и КЛ, например, по искусственному интеллекту (ИИ), информационному поиску (ИП) и почти полное отсутствие русскоязычных терминов по другим разделам КЛ, например, по оценке систем (system evaluation). Следствием пп. 2, 4 являются ситуации, когда один и тот же термин в смежных науках имеет различные толкования, например, «синтаксический анализ» в ИИ и в КЛ. Место КЛ среди смежных наук и состав КЛ обсуждаются в р. 2.

Учитывая вышеперечисленные особенности КЛ, мы стремились в начальной версии тезауруса найти источники «живых» терминов РКЛ и их толкований и именно их зафиксировать в словарных статьях терминов. Список основных лексикографических источников терминов — тезаурусов и толковых словарей смежных наук, — а также принятая в нашем проекте структура описания термина, представлены в статье [1] данного сборника.

В соответствии с направленностью тезауруса на РКЛ, в условиях недостаточности основных лексикографических источников рассматривались также дополнительные источники в виде предметных указателей и глоссариев русскоязычных учебников и монографий и самих их текстов, а также

[1] Это не противоречит тому факту, что локально мы опираемся именно на англоязычные источники, стараясь дополнить картину КЛ в отдельных ее частях.

коллекция научных текстов [Прил., рус. 1][2], из которой были получены статистические показатели встречаемости терминов. Параметры и методика исследования массива научных статей конференции «Диалог» описываются в р. 3.

Мы используем в качестве источника терминов массив текстов «Диалога», поскольку он содержит «живые» термины, которые реально употребляются в РКЛ, а также предметные указатели, хотя они имеют те же недостатки, что и сами источники — локально являются пересказом англоязычных источников, что может приводить к выделению термина, реально в РКЛ не используемого. См. р. 4 и 5, в которых обсуждаются дополнительные лексикографические источники.

Для английской части тезауруса рассматривались только лексикографические источники и учебники, коллекции текстов не исследовались. В р. 6 описывается тематика терминов начальной версии тезауруса.

2. РКЛ как предметная область

В РКЛ для обозначения всего направления в основном используется термин *Прикладная лингвистика*. В последние два десятилетия все более употребительным становится термин *Компьютерная лингвистика*[3]. Эти термины и термины, представляющие близкие направления и разделы КЛ, имеют следующие частотные характеристики в [Прил., рус. 1]:

автоматическая обработка текста — 155
автоматическая обработка (естеств.) языка –7+8
искусственный интеллект — 265 + ИИ (108)
когнитивная лингвистика — 58
компьютерная лингвистика — 900[4]
корпусная лингвистика — 159
лингвистическая технология — 11
прикладная лингвистика (языкознание — 9) — 120
речевая технология — 74
синтез речи — чуть более 300.

В связи с отмеченными особенностями 2, 3, 4 научной области КЛ трудно дать определение. В русскоязычных источниках даются экстенсиональные

---

[2] Ссылка на Приложение в конце статьи, в котором перечислены наиболее употребляемые составителями статей источники как терминов, так и толкований. Таким образом, библиографические ссылки в статье отсылают к одному из двух взаимодополняющих списков: Литература и Приложение.

[3] Включенный, в частности, А. С. Нариньяни в название конференции «Диалог» с момента ее возрождения после перестройки в 1995 г.

[4] Большинство этих вхождений относится к библиографическим ссылкам на статьи в сборниках Диалога.

определения, например, в [2]: «Раздел лингвистики, задачей которого является исследование проблем, связанных с машинной обработкой текста: организацией естественно-языкового интерфейса, машинным переводом и реферированием, статистическим анализом словарей и текстов на ЭВМ, автоматическим распознаванием речи». При этом КЛ рассматривается как ветвь разных наук, в приведенном определении — лингвистики, в других источниках — ИИ [Прил., рус. 6]) и прикладной лингвистики [Прил., рус. 7]. РКЛ также пересекается с социо-, квантитативной и другими «лингвистиками». В англоязычных источниках КЛ подчиняется когнитивной науке (cognitive science), в частности, в меморандуме Х. Ускорайта [3]. Он справедливо разделяет КЛ на два направления: «теоретическая КЛ» (*Theoretical CL*) и «прикладная КЛ» (*Applied CL)*. Теоретическая КЛ базируется на теоретической лингвистике и пересекается с психолингвистикой и когнитивной психологией. Относительно прикладной КЛ указывается, что эта область также иногда обозначается терминами «language engineering» и «(human) language technology» и направлена на достижение практических результатов в моделировании ЕЯ.

Представление о том, что входит в ПО КЛ, также изменялось в истории этого направления в течение чуть более 60 лет. Современный состав ПО КЛ можно обрисовать по англоязычным источникам, в частности — [Прил., англ. 7, 8, 10]. Согласно обзору [Прил., англ. 10], в состав КЛ входит анализ/понимание vs. генерация текстов, распознавание и синтез речи, а также диалог и дискурс, мультимодальность, математические методы, лингвистические ресурсы и оценка систем[5]. Учебники [Прил., англ. 7 и 8] выделяют эти же направления, но добавляют к ним прикладную тематику: машинный перевод (МП), ИП, автоматическое реферирование, вопросно-ответные системы, извлечение знаний, автоматическое индексирование, взаимодействие с компьютером на ЕЯ, интеллектуальный поиск в текстах и некоторые другие.

С распространением эмпирических компьютерных технологий в конце 80–90 гг. зарождается потребность разделения «широкой» КЛ и «узкой», собственно технологической, области скорее ИИ, чем КЛ, для которой письменные тексты и звучащая речь являются только одним из видов данных. Особенно это характерно для направлений, связанных с обработкой речи. Специфика этого направления обсуждается в р. 5.

Отличие КЛ от традиционной лингвистики, которая является основой, базой для КЛ, состоит в том, что предметом КЛ (и, в конечном счете, ее объектом) является информация, а не языковая форма. Это верно как для теоретических направлений (анализ/понимание текста — генерация текста),

---

[5] Здесь опять возникает неопределенность. По [3] этот список относится скорее к «теоретической КЛ», которая «…deals with formal theories about the linguistic knowledge that a human needs for generating and understanding language… and implement(s) them as computer programmes.». При этом обзор называется «The state of the art in Human Language Technology» — термин, который сам Х. Ускорайт вместе с термином «language engineering» относит к прикладной КЛ.

так и для прикладных разделов КЛ: МП, ИП, речевые технологии (РТ) и др. При этом КЛ опирается преимущественно на «новейшие» разделы теоретической лингвистики — семантику, теорию дискурса и прагматику, быстрое развитие которых в последние десятилетия было вызвано в том числе и самой КЛ.

В начальной версии тезауруса мы старались собрать наиболее частотные термины, которые реально встречаются в РКЛ. В связи с этим в качестве основного источника терминов РКЛ взята коллекция русскоязычных текстов, представленная в р. 3.

3. Коллекция текстов как источник терминов

Складывающаяся терминологическая система отражается в учебниках, справочниках, электронных ресурсах и материалах конференций. Учитывая недостаток современной справочной русскоязычной литературы по КЛ, было принято решение создать базовый словник русскоязычных терминов по КЛ, используя материалы ежегодной международной конференции «Диалог». Собранная коллекция документов содержит тексты докладов, представленных на конференции «Диалог» в 2000–2010 гг., и имеет следующие характеристики: число документов — 1 193, объем — 4 610 694 словоупотреблений, суммарный размер — 27,5 Мб.

На этапе создания словника применялась словарная технология КЛАН [7], которая позволяет на базе коллекции текстовых документов создать список использованных в ней слов и словосочетаний — кандидатов в термины ПО, — причем каждый термин снабжен следующими статистическими показателями: частота встречаемости в коллекции и частота по документам. В процессе автоматической обработки этой коллекции было реализовано первоначальное наполнение словника с использованием технологии обучения по массиву текстов на базе лингвистических моделей: универсальный морфологический анализ, сборка именных словосочетаний на основе 20 синтаксических шаблонов, предсказание незнакомых слов. В результате был получен исходный словник объемом 79 678 слов и 512 783 словокомплекса (СК).

На этапе фильтрации полученный список терминов-кандидатов был полуавтоматически отсортирован для выявления наиболее важных (статистически значимых) в данной ПО слов и СК, имеющих терминологический характер:

- удалены термины с частотой встречаемости 1–3;
- удалены термины, встретившиеся только в одном тексте;
- удалены или перенесены в стоп-словарь служебные слова;
- отфильтрована нетерминологичная лексика специальных лексико-семантических разрядов (топонимы, имена персон и организаций);
- удалены ошибочные предсказания, в том числе: ошибки в предсказании части речи или морфологического класса, ошибки при определении лемм, некорректности, основанные на ошибках в тексте.

При удалении неверных гипотез были автоматически удалены построенные на их основе словосочетания.

В результате этапа фильтрации объем словника сократился до 23 760 слов и 31 709 СК. Словарь отфильтрованных терминов-кандидатов был передан экспертам в данной ПО для проведения экспертной оценки. Работа экспертов поддерживается конкордансом, который позволяет получить все примеры употребления термина словаря вместе с его контекстами.

При автоматизированной подготовке базового словника были использованы возможности сортировки знаменательных слов по части речи, морфологическому типу и убыванию встречаемости, а также автоматической проверки вхождения слов в состав СК:

- терминообразующие лексические единицы разряда фамилий — перенесены в стоп-словарь (*Зализняк — словарь Зализняка*; *Мельчук — модель Мельчука*);
- нарицательные одушевленные существительные — выборочно удалены (*адъюнкт*, *коллекционер*);
- глаголы, наречия — отсортированы по части речи и в порядке убывания встречаемости, выборочно удалены общеупотребительные.

В результате такой обработки объем словника сократился до 13 865 слов и 27 458 СК.

Далее эксперты провели отсев терминов, не относящихся к КЛ. Итоговый объем словаря составил 6013 слов и 8489 СК. Была проведена разметка словника с помощью системы семантических признаков, которая соответствует делению КЛ на три направления: АОТ (автоматическая обработка текста), РТ (речевые технологии) и КорпЛ (корпусная лингвистика). Наиболее представительным оказался подсловник АОТ (1524 слова), из них в топ-список (встречаемость выше 20) отнесено 941 слово; топ-список многословных единиц АОТ составляет 1001 СК. Как и следовало ожидать, направление РТ представлено на «Диалоге» слабо: в топ-список вошли 105 терминов по РТ и прикладной фонетике. Эксперты могут работать с различными выборками из соответствующего итогового подсловника, сформированными на базе таких признаков как частота, часть речи, структура СК и т. п. Так, для базовой версии тезауруса было принято решение ограничиться топ-списком, составленным из наиболее важных терминов-существительных и именных групп.

На предварительном этапе эксперты существенно опирались не только на знания о предмете и направлениях КЛ, но и на общелингвистические представления о терминологичности и путях формирования терминологических словников. Так, наш подход, основанный, в том числе, на предварительном структурировании ПО, согласуется с общей методикой формирования словников на базе классификационных схем предметных областей, см., например, [8].

Что касается английской части словника, то для данной версии тезауруса, имеющей русско-английскую направленность, выбирались переводные эквиваленты из доступных англоязычных источников по КЛ.

4. Особенности русскоязычных терминоисточников по РКЛ

Термины КЛ из основных источников — толковых словарей и тезаурусов, относящихся к лингвистике и смежным с КЛ областям, — требуют проверки. Например, тезаурус ИНИОН [8] и ЛЭС [Прил, рус. 11] считают основным термином *автоматический перевод,* присвоив ему статус дескриптора, а *машинный перевод* — аскриптором к нему. Встречаемость в [Прил, рус. 1]: *машинный перевод* 318; *автоматический перевод* 58[6]. Более высокая частотность первого не объясняется ссылками на литературу, в частности, на название сборника «*Машинный перевод и прикладная лингвистика*» приходится лишь 28 вхождений термина *машинный перевод*. Интернет-энциклопедии [Прил, рус. 2,3] и учебники придерживаются этой же традиции, которую и мы не стали нарушать. На сайте Европейской ассоциации машинного перевода [10] также отмечается, что термин *machine translation* звучит архаично, но тем не менее сохраняется как основной общий термин для всей области.

Дополнительные источники терминов — предметные указатели и глоссарии к научным текстам, а также сами тексты — более субъективны. Указатели отражают текст конкретной книги, статьи, а не ПО. В описании термина в тезаурусе мы отмечаем, где он встретился, в тексте издания или в предметном указателе (глоссарии), считая, что включение термина в предметный указатель повышает его терминологический статус, хотя бывает, что достаточно значимый термин встречается только в тексте. Так, термин *structural transfer* упоминается в тексте учебника [Прил., англ. 8] (и определяется как *transformation of source language structures into equivalent target language forms*), но отсутствует в предметном указателе. Кроме того, если термин заимствован из английского, то он может органично выглядеть в тексте, например, «*По Р. Шенку скрипт — это некоторая общепринятая, общеизвестная последовательность причинных связей*», но становиться неадекватным, когда автор выносит его в предметный указатель. Термин «скрипт» входит в предметные указатели учебников [7] и [Прил, рус. 7], в обоих случаях являясь калькой английского термина. В [Прил, рус. 1] «скрипт» употребляется исключительно как термин информатики для обозначения определенного типа программ, например, в следующем контексте: «база данных жестов, включающая в себя файлы скриптов, управляющих виртуальным демонстратором». Таким образом, «скрипт» является дескриптором в информатике, а частотность его употребления в РКЛ со значением «сценарий» по [Прил, рус. 1] равна 0.

---

[6] Поиск в Интернете дает обратное соотношение: *машинный перевод* 640 000, *автоматический перевод* 1 960 000, которое объясняется тем, что если речь идет о МП с языка на язык (а не о переводе на другой тариф и т. п.), основную часть ответов составляет реклама он-лайн переводчиков, т. е. имеется в виду разновидность *полностью автоматического перевода* (*онлайн-перевод*).

5. Англоязычная литература как источник терминов

Учитывая скачок, совершенный в области РТ в течение последних нескольких лет, когда эта область окончательно сложилась как высокотехнологичное направление, имеющее огромный практический и коммерческий выход, а также тот факт, что это направление слабо представлено в [Прил, рус. 2], авторы избрали методику сбора терминов, обратную методике, принятой в разделах АОТ и КорпЛ, используя в качестве основных англоязычные источники. В некоторой степени этот подход применяется и в других разделах, таких как «направления КЛ».

В качестве основы для сбора терминологического материала по РТ были взяты предметные указатели нескольких современных и наиболее авторитетных англоязычных книжных источников обзорно-учебного профиля. Кроме того, активно использовались глоссарии, входящие в состав известных звуковых анализаторов Adobe Audition 1.5. 2004 и Speech Analyzer 1.5–2002 [Прил, англ. 2].

На данной терминологической базе был составлен англо-русский словник параллельных терминов по РТ, включающий более 700 парных терминов (англо-русских эквивалентов). На следующем этапе из собранного англо-русского словника была выделена базовая часть, которая далее была включена в состав первой редакции проектируемого тезауруса по направлению РТ. Выделение базовой части словника осуществлялось экспертами по данному направлению с учетом списка частотных терминов по РТ и прикладной фонетике, полученного в результате компьютерной обработки и анализа электронных материалов конференции «Диалог».

Направление РТ характеризует большой массив собственной терминологии, например, в подразделах «автоматический синтез речи», «автоматическое распознавание речи» и др. Но имеются и точки пересечения с АОТ (см. пример ниже). Имеются и общие проблемы, к числу которых относятся пробелы в русскоязычной терминологии, ведущие к необходимости перевода терминов, а также отсутствие сложившейся традиции в понимании и употреблении уже имеющихся терминов.

Рассмотрим в качестве примера термин *spoken language machine translation*. Задача автоматического перевода устной речи возникла на стыке МП и РТ. *Spoken Language Processing* обычно переводится как *Автоматическая обработка устного языка*, одной из задач которой является автоматический устный перевод (АУП) с его разновидностями, соответствующими АУП типа «Речь(L1) --> Текст(L2)» и АУП типа «Речь(L1) --> Речь(L2)». Вторая разновидность представлена английским термином *speech-to-speech translation*. В русскоязычной литературе такой традиции нет, как нет (или практически нет) и такого типа приложений. Поиск в Интернете дал в качестве эквивалента для *spoken language machine translation* вариант *автоматический перевод устной речи*, который встретился дважды: в рецензии на англоязычную книгу по МП и на сайте «Лингвистика в России» со ссылкой на Группу речевой информатики Санкт-Петербургского института информатики и автоматизации РАН (впрочем, в русскоязычной части сайта этого института термин найти не удалось).

Определенную трудность вызывает представление базовых терминов *Лингвистические технологии* и *Речевые технологии* и их дифференциация,

соответственно, с понятиями *АОТ* и *Автоматическая обработка звучащей/устной речи (АОЗР)*. В англоязычной литературе разграничение между *Speech Technology = РТ* и *Speech Processing = АОЗР* проводится нечетко. Последний термин покрывает все прикладные задачи, связанные с автоматической обработкой устной речи, и здесь можно выделить четыре основных направления: цифровая (параметрическая) обработка речевого сигнала (ЦОРС), автоматический синтез речи, автоматическое распознавание речи, создание речевых корпусов. Одна из возможных точек зрения: термин РТ равнозначен термину Автоматическая обработка звучащей/устной речи, так как ЦОРС — тоже речевая технология, направленная на создание автоматических звуковых анализаторов/редакторов речи. Однако некоторые авторы к РТ относят только синтез, распознавание и корпусные технологии, т. е. понимают РТ уже, чем АОЗР.

6. Тематические классы терминов начальной версии русско-английского тезауруса

В итоге для начальной версии базового словника тезауруса было выделено порядка 3 тысяч терминов, к настоящему моменту описаны и внесены в электронный ресурс около 1100 терминов: дескрипторов — около 700, аскрипторов — более 400, связей между терминами — около 2500, источников терминов и их определений — 126. Множество терминов распадается на пять основных терминологических областей:

1. «Направления КЛ» — термины, называющие отдельные направления КЛ. Мотив выбора этой группы — получение по возможности полной картины о возможном предметно-тематическом составе тезауруса. Термины этой группы включены экспертом независимо от частоты их встречаемости в [Прил рус. 1];
2. «РТ» — относительно самостоятельное и минимально пересекающееся с остальными направление КЛ;
3. «КорпЛ» — базовое направление для статистических методов ИИ и различных эмпирических подходов, которые проникают во все направления современной КЛ;
4. «ИП» — одно из основных прикладных направлений КЛ;
5. «МП» — важнейшее приложение КЛ, традиционно интегрирующее всю проблематику АОТ, а в последнее время тесно взаимодействующее с РТ в рамках задачи автоматического устного перевода;
6. Группа терминов «метаязык». К этой области относятся термины фонетического, морфологического, лексического, синтаксического и семантического уровней языка и представлений этих уровней. Здесь систематично рассматривались терминологические обозначения семантических отношений — основополагающая для КЛ группа терминов, используемых как в ресурсах, например, лексико-семантических базах, так и в моделях языка и описаниях для систем АОТ, например, *агенс, начальная точка* и т. д.

Заключение

В статье излагаются наблюдения над структурой ПО КЛ и терминологией КЛ, сделанные в процессе создания начальной версии русско-английского тезауруса по КЛ. Отмечены особенности самой ПО (междисциплинарность, неравномерность развития разных направлений КЛ и др.), показано их влияние на терминологию и создание терминологических словарей. Кроме того, сформулирована главная особенность КЛ по сравнению с общей лингвистикой — направленность на передаваемую информацию, а не на формы языка. Именно наличие объекта, отличного от традиционной лингвистики, выделяет КЛ в самостоятельную научную дисциплину. Описаны методика и процесс обработки корпуса текстов для выделения терминов РКЛ и частота встречаемости некоторых терминов. Проанализированы особенности терминов КЛ и их отбора из основных лексикографических и дополнительных источников.

Благодарности

Работа выполнена при поддержке Российского гуманитарного научного фонда (грант № 10-04-12108в).

Авторы также выражают благодарность за плодотворное сотрудничество студентам и аспирантам МГУ им. М. В. Ломоносова, принимавшим участие в исследовании.

References

1. *Artificial* Intellect Explanatory Dictionary [Tolkovyi Slovar' po Iskustvennomu Intellektu].1992, available at: http://www.raai.org/library/tolk/aivoc.html
2. *Computaional* Linguistics Knowledge Web-site, available at: http://uniserv.iis.nsk.su/cl
3. *Krongauz M. A.* 2001. Semantics.
4. *Linguistics.* Information and Search Thesaurus of INION RAS. [Iazykoznanie. Informatsionno-Poiskovyi Tezaurus INION RAN]. 2007.
5. *Pererva V. M.* 1976. On Principles and Problems of Terms Selection and Terms Dictionary Formation [O Printsipakh I Problemakh Otbora Terminov I Sostavleniia Slovnika Terminologicheskikh Slovarei]. Problematika Opredelenii Terminov v Slovariakh Raznykh Tipov : 190–204.
6. *Sidorova E. A. Сидорова Е. А.* Multipurpose Dictionary Subsystem of Object Vocabulary Extraction [MnogotselevaIa Slovarnaia Podsistema Izvlecheniia Predmetnoi Leksiki]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14) : 475–481.
7. *Sokolova E. G., Zagorul'ko Iu. A., Kononenko I. S.* 2009. The Experience of Knowledge and Web Resources Classification for the Computaional Linguistics Knowledge Web-site [Opyt Sistematizatsii Znanii I Internet-resursov dla Portala Znanii po Komp'iuternoi Lingvistike]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 465–470.
8. *Uszkoreit H.* What is computational linguistics?, available at: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html
9. *Website* EAMT (The European Association for Machine Translation), available at: http://www.eamt.org/
10. *Zagorul'ko Iu. A., Borovikova O. I., Kononenko I. S., Sokolova E. G.* 2011. Designing of Russian-English Computational Linguistics Thesaurus [Razrabotka Russko-Angliiskogo Tezaurusa po Komp'iuternoi Lingvistike]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011").

Таблицы содержат источники, наиболее часто используемые в начальной версии тезауруса, в основном, по указанным в статье тематикам. Всего на начальном этапе создания тезауруса зарегистрировано 127 источников.

Русскоязычные источники

	Тип источника	Название источника, библиографическая ссылка, URL	Дается определение	Встреч. дескриптор	Встреч. аскриптор
1	коллекция текстов	Коллекция текстов Диалог 2000–2010	1	116	78
2	интернет-ресурс	Интернет- энциклопедия «Википедия» http://ru.wikipedia.org	61	8	8
3	интернет-ресурс	Интернет- энциклопедия «Кругосвет» http://www.krugosvet.ru	11		2
4	книга	Трахтеров А. Л. Английская фонетическая терминология. М., Изд-во литературы на иностранных языках, 1962.	22	14	8
5	коллекция текстов	Корпус текстов по корпусной лингвистике		18	
6	книга	Искусственный интеллект. Справочник в 3-х томах. — М.: Радио и связь, 1990.	7	4	2
7	учебник	Баранов А. Н. Введение в прикладную лингвистику. Учебное пособие. — М.: Эдиториал УРСС, 2001. — 360 с.	10	2	4
8	учебник	Кобозева И. М. Лингвистическая семантика: Учебник. Изд. 4-ое — М.: Книжный дом «ЛИБРОКОМ», 2009. — 352 с. (Новый лингвистический учебник).	10	3	1
9	учебник	Кодзасов С. В., Кривнова О. Ф. Общая фонетика. М, РГГУ, 2001.	20	27	13
10	учебник	Тестелец Я. Г. Введение в общий синтаксис М.:РГГУ, 2001.	18		4
11	энциклопедический словарь	Лингвистический энциклопедический словарь. / Под ред. В. Н. Ярцевой. М.: Советская энциклопедия, 1990. — 685 с. [3 изд. 2002.]	6	1	1
12	энциклопедия	Энциклопедия «Русский язык». Гл. ред. Ю. Н. Караулов. Научное издательство «Российская энциклопедия», М., 1997.	9	7	3
13	интернет-ресурс	Сайт кафедры перевода и переводоведения ТюмГУ http://tc.utmn.ru	7	4	4

Англоязычные источники

1	интернет-ресурс	Интернет- энциклопедия «Wikipedia» http://en.wikipedia.org	24	4	18
2	документация	Документация к комп.программе Speech Analyzer 1.5–2002 http://www.sil.org/computing/speechtools/	39	18	2
3	интернет-ресурс	Интернет- энциклопедия «Glottopedia» http://www.glottopedia.de/index.php/Main_Page	13	5	
4	книга	J. Harrington, S. Cassidy. Techniques in Speech Acoustics. Text, Language, Technology, vol.8. Kluwer Academic Publishers. Dordrecht/Boston/London, 1999.	14	7	2
5	книга	J. Holmes and W. Holmes. Speech Synthesis and Recognition. 2nd edition/ Taylor&Francis, London/New York, 2001.	9	8	1
6	книга	K. Johnson. Acoustic and Auditory Phonetics. Blackwell Publishers, Cambridge, 1997.	38	19	4
7	учебник	Jurafsky Danial, Martin James H. Speech and language Processing: An Introduction to Natural language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, 2000	13	12	5
8	обзор	The Oxford handbook of computational linguistics (Ruslan Mitkov ed.) N.Y.: Oxford university press, 2003	62	43	21
9	статья	Igor Mel'čuk. Actants in semantics and syntax I: actants in semantics //Linguistics, 2004, 42:1, p.1-66	3	1	7
10	обзор	Survey of the State of the Art in Human Language Technology (Ronald Cole, editor in chief) 1996 http://cslu.cse.ogi.edu/HLTsurvey/	3	1	6
11	интернет-ресурс	Glossary of linguistic terms http://www.sil.org/linguistics/GlossaryOflinguisticTerms/	5	4	3