# ОРГАНИЗАЦИЯ ДВУЯЗЫЧНОГО ПОИСКА В УНИВЕРСИТЕТСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ РОССИЯ

Б.В. Добров, Н.В. Лукашевич
Научно-исследовательский вычислительный центр
МГУ им. М. В. Ломоносова;
АНО Центр информационных исследований НИВЦ МГУ, Воробьевы горы, Москва, 119899
{dobroff, louk}@mail.cir.ru

# ORGANIZATION OF BILINGUAL SEARCH IN UNIVERSITY INFORMATION SYSTEM RUSSIA

Boris V. Dobrov, Natalia V. Loukachevitch
Research Computing Center of Moscow State University;
NCO Center for Information Research
339, Research Computing Center of Moscow State University;
Vorobyevy Gory, Moscow, 119899, Russia
{dobroff, louk}@mail.cir.ru

The paper presents the main principles and stages of development of bilingual Thesaurus on sociopolitical life, which specially created as a tool for automatic text processing of large text collections. Now the Russian part of the Thesaurus includes 64 thousand terms, the English part – 55 thousand terms. The Thesaurus allows conceptual indexing of documents, independent of an initial language of a document. Estimation of document relevance in bilingual environment can be made using so called structural thematic summary.

## 1. Введение

Важной задачей современного информационного поиска является разработка многоязычных информационных систем, в которых запрос составляется на одном языке, а найденные документы могут быть написаны как на языке запроса, так и на других языках информационной системы. В последнее время проблема многоязычного поиска стала одной из приоритетных в информационном поиске, например, в рамках конференции TREC выделено соответствующее направление [1]. С этой проблематикой работают также исследователи, объединенные европейской программой CLEF (Cross-Language Evaluation Forum) [2].

Задачи поиска документов в многоязычных текстовых коллекциях решаются уже в течение нескольких десятилетий. Для решения таких задач создавались многоязычные информационно-поисковые тезаурусы [3], в

которых для каждого дескриптора тезауруса были сформулированы его варианты на нескольких языках. Однако традиционные многоязычные тезаурусы создавались как вспомогательные средства для ручного индексирования/поиска специалистами-индексаторами.

В настоящее время возникновение громадных электронных коллекций требует разработки средств автоматической обработки и поиска документов в многоязычных коллекциях. Именно разработка такого многоязычного ресурса для автоматического индексирования документов на европейских языках ставилась в проекте EuroWordNet [4]. В рамках этого проекта для ряда языков были созданы лингвистические ресурсы общего назначения, которые нужно еще дополнять при работе в конкретных предметных областях, что является отдельной проблемой.

Мы опишем основные принципы организации информационного поиска русских и английских документов в Университетской информационной системе РОССИЯ (УИС РОССИЯ, http://www.cir.ru) [5]. Обработка документов основывается на русско-английском Тезаурусе по общественно-политической жизни [6], который в русской части имеет 64 тысячи терминов, в английской части 55 тысяч терминов, представленных как иерархическая сеть 27 тысяч понятий. Тезаурус специально разработан в качестве инструмента для автоматической обработки больших текстовых массивов. На основе Тезауруса производится автоматическое концептуальное индексирование русских и английских документов, при этом производится автоматическое разрешение многозначности терминов. Построенный концептуальный индекс позволяет выполнять поиск одновременно русских и английских документов по запросу на русском или английском языке.

## 2. Традиционные тезаурусы и автоматическое индексирование: тезаурус EuroVoc

Возможности использования традиционных информационно-поисковых тезаурусов для автоматической обработки документов рассмотрим на примере тезауруса Европейского Сообщества EuroVoc.

Тезаурус EUROVOC широко применяется экспертами органов Европейского Союза для содержательной обработки документов. В настоящее время является важным инструментом для обмена информацией в многоязычной среде ЕС.

Русскоязычная версия тезауруса EUROVOC подготовлена в результате многолетнего труда сотрудников Парламентской библиотеки в сотрудничестве со специалистами различных организаций.

Как всякий информационно-поисковый тезаурус, созданный изначально для ручного индексирования, EUROVOC представляет собой искусственный язык, созданный на базе естественного языка предметной области.

Специфика тезауруса, предназначенного для ручного индексирования, влечет за собой определенные проблемы при использовании его для автоматической обработки.

### 2.1. Этап индексирования

#### 2.1.1. Вариативность синонимов

Представление дескрипторов тезауруса в тексте значительно более разнообразно, чем это указано в русской версии тезауруса EUROVOC.

Например, дескриптор *ОХРАНА ОКРУЖАЮЩЕЙ СРЕДЫ* помимо указанных в тезаурусе вариантов может быть выражен также следующими словами и терминами, неописанными в тезаурусе, но встречающимися в текстах российских правовых актов: *защита природы, природозащитный, природоохранный, природоохранительный (меры, деятельность, процесс*); дескриптор *ОХРАНА ЛЕСОВ - защита лесов, защита лесного фонда, лесозащитный (деятельность, мероприятия), лесоохрана, лесоохранный;* дескриптор *СУДЕБНЫЕ РАСХОДЫ – судебные издержки*, дескриптор *РАСХОДЫ НА ОБОРОНУ – оборонные расходы, военные расходы, военный бюджет, оборонный бюджет* и еще сотни примеров.

#### 2.1.2. Неоднозначность терминов

Не указана неоднозначность некоторых терминов, описанных в русской версии только в одном из значений, что не нужно для человека-индексатора, но необходимо для автоматической обработки.

Примеры неоднозначных терминов тезауруса, включенных в русскую версию EUROVOC в одном значении, что может привести к неправильному индексированию: *кожа* (как кожевенная продукция и кожа человека), *печать* (как СМИ, как штамп, как процесс печатания), *питание* (еда и электрическое питание), *корма* (питание животных и часть корабля), *образование* (как обучение и как создание чего либо). Средства описания и работы с многозначностью необходимы для любого ресурса, использующегося для автоматической обработки текстов.

#### 2.1.3. Специфические термины

Тезаурус в своем изложении иерархии понятий останавливается на достаточно высоком уровне иерархии и не включает более конкретные термины. Между тем, например, среди нормативных документов широко представлены такие документы, в которых обсуждается *минтай*, но нет слова *рыба*, обсуждаются *солдаты,* но нет слова *военнослужащий*, обсуждается *пшеница*, но нет слова *зерно* и многие другие подобные примеры. Такие тексты не могут быть проиндексированы правильно из-за нехватки информации в тезаурусе.

### 2.2. Этап поиска

Автоматическая обработка предполагает и автоматизацию поиска, то есть поиск с автоматическим расширением запроса. В связи с этим большую значимость приобретает качество описания отношений между дескрипторами в тезаурусе. Технология описания отношений в тезаурусе EUROVOC не дает возможности использования их для автоматического расширения запроса.

#### 2.2.1. Отношения ВЫШЕ-НИЖЕ

Отношения ВЫШЕ-НИЖЕ, обычно, могут быть использованы для расширения запроса. Но не всегда.

Например,

*ПРИБОРОСТРОЕНИЕ*
НИЖЕ *КОНТРОЛЬНО-ИЗМЕРИТЕЛЬНЫЕ ПРИБОРЫ*
НИЖЕ *НАУЧНЫЕ ПРИБОРЫ*
и т.п.

Значит, выбрав в запрос *ПРИБОРОСТРОЕНИЕ*, получаем тексты о научных приборах, например, о правилах списания, эксплуатации, передачи и т.п.

#### 2.2.2. Ассоциативные отношения

На широко представленные в тезаурусе отношения ассоциации невозможно уверенно опереться при автоматическом расширении запроса:

*ОХРАНА ДЕТСТВА*
АСЦ *ПРОСТИТУЦИЯ*

Ищем тексты о детях, получаем тексты о проституции, из которых лишь некоторые о детях. В обратную сторону тоже не лучше: ищем тексты о проституции, получаем тексты о детях.

*МОНОГРАФИИ*
АСЦ *ТИПОГРАФИИ*

Ищем тексты о монографиях, получаем тексты о типографиях и наоборот.

Отметим, что эти проблемы возникают уже в одноязычной среде, а в многоязычной значительно усугубляются.

Таким образом, для того, чтобы автоматически обрабатывать тексты для различных приложений информационного поиска необходимо разрабатывать специальные лингвистические ресурсы, каким и является Тезаурус по общественно-политической жизни УИС РОССИЯ [7].

## 3. Англоязычные коллекции в УИС РОССИЯ

В настоящее время в УИС РОССИЯ имеются следующие текстовые коллекции англоязычных документов:

- более 160 тысяч реферативных описаний научных статей, собранных в рамках проекта RePEc (Research Papers in Economics - рабочих материалов по экономике), получаемых с сайта проекта СОЦИОНЕТ;
- коллекция документов Совета Министров Совета Европы (в настоящее время загружена текстовая коллекция, в дальнейшем планируется включить все доступные документы Совета Европы);
- материалы переписи населения СССР 1939г., предоставленные университетом Торонто.

В течение 2002-2003 гг. планируется значительно расширить состав текстовых коллекций в УИС РОССИЯ, прежде всего за счет правовых документов Совета Европы, текстов научных статей по экономике из коллекции RePEc.

Для решения различных задач двуязычного поиска по различным текстовым коллекциям, объединяемых в рамках единых предметных областей, потребовалось создать соответствующие технологии. Основой этих технологий является двуязычный Тезаурус по общественно-политической тематике, разработка которого потребовала значительных усилий. Авторы разделяют точку, подтвержденную в работе [2] экспериментально, что качество результатов в многоязычном поиске чрезвычайно сильно зависит от качества используемых лингвистических ресурсов.

## 4. Особенности построения многоязычного лингвистического ресурса для автоматической обработки больших текстовых коллекций

Для предметной области общественно-политической проблематики имеются следующие двуязычные источники: англо-русские и русско-английские словари по экономике, бизнесу, праву и другие, а также был переведен на русский язык многоязычный европейский тезаурус EuroVoc. Основными особенностями двуязычного тезауруса, предназначенного для автоматического концептуального индексирования являются:

- описание как можно более точных соответствий между терминами путем размещения их в синонимические ряды одного и того же понятия, в том числе оставление термина без перевода, если этого перевода нет;
- описание как можно большего количества синонимических вариантов выражения понятия в тексте для обоих языков, что является базой для распознавания понятия в тексте;
- описание многозначности терминов обоих языков;

- описание как можно большего количества многословных синонимических вариантов, что облегчает процедуру разрешения неоднозначности терминов в тексте.

Для развития двуязычного Тезауруса по общественно-политической жизни для автоматического индексирования были проделаны следующие этапы работ.

На первой стадии русскоязычные термины тезауруса были переведены на английский язык. Было получено 33 тысячи англоязычных терминов. Однако при этом вне Тезауруса остались термины, свойственные общественно политической жизни англоязычных стран и не присутствующие в России, кроме того, синонимическая вариативность англоязычной части представлена не достаточно широко.

Поэтому на следующем этапе мы стали вычитывать наиболее известные англоязычные толковые и специализированные словари, информационно-поисковые тезаурусы на предмет выявления новых понятий и новых синонимов уже описанных понятий. Термины извлекались не только из заголовков словарных статей, но часто из примеров и толкований. Эта часть практически завершена и привела к описанию 55 тысяч англоязычных терминов в Тезаурусе.

В настоящее время осуществляется вычитка и корректирование англоязычных синонимичных рядов, при этом осуществляется контроль по Интернет неочевидных случаев перевода. Кроме этого, просматривающий может предложить дополнительные варианты синонимичных многословных терминов и проверить по Интернет их реальное существование.

В ближайшее время начнется разбор коллекций результатов автоматического тематического анализа англоязычных текстов с целью корректировки сделанных в Тезаурусе описаний англоязычных терминов.

## 5. Поиск двуязычных документов в УИС РОССИЯ

### 5.1. Тематический анализ русскоязычных и англоязычных документов

Для документов на русском или английском языке производится их тематический анализ. В результате тематического анализа для документа создается концептуальный индекс, независящий от исходного языка документа и конкретных синонимов, употребленных в тексте. Каждое понятие в концептуальном индексе имеет вес, построенный как на основе частотных характеристик употребления понятий в тексте, так и его тезаурусных и текстовых связей с другими понятиями [7].

### 5.2. Возможности двуязычного поиска в УИС РОССИЯ

Помимо стандартных поисковых атрибутов, включающих средства контекстного поиска, использующие результаты морфологической обработки русских, английских и смешанных русско-английских текстов, в УИС РОССИЯ разработаны как русскоязычный, так и англоязычный интерфейс тематического поиска.

Работая в рамках русскоязычного интерфейса, пользователь может выбрать в запрос те или иные понятия тезауруса. Пусть, например, выбрано понятие "*ГОСУДАРСТВЕННЫЙ СЛУЖАЩИЙ*" (см. Рис. 1).

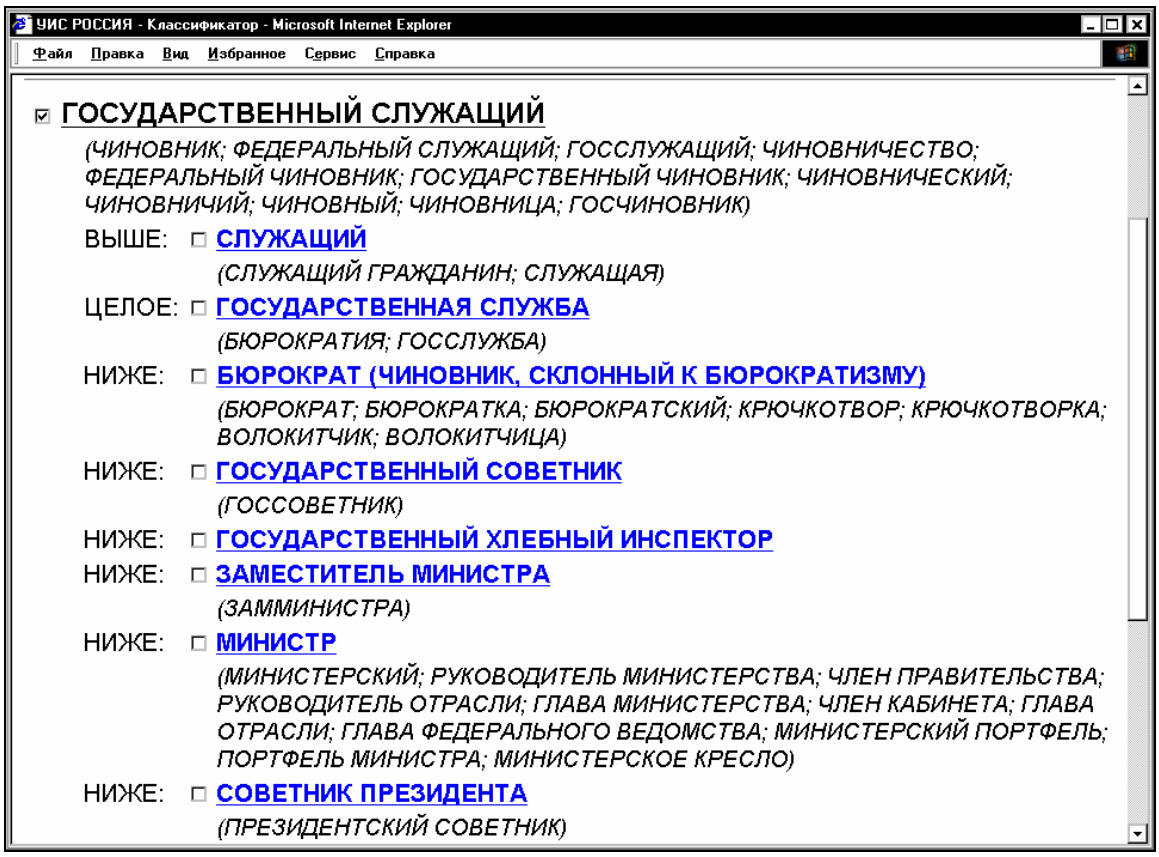


**Рис. 1. Выбор элемента тезауруса в запрос**

Возможен поиск с опцией "РАСШИРЕНИЕ ПО ДЕРЕВУ", когда релевантными считаются документы, содержащие не только синонимы выбранного понятия, но синонимы подчиненных понятий. В многоязычной коллекции результатом поиска могут быть ссылки на документы на разных языках (см. Рис. 2).

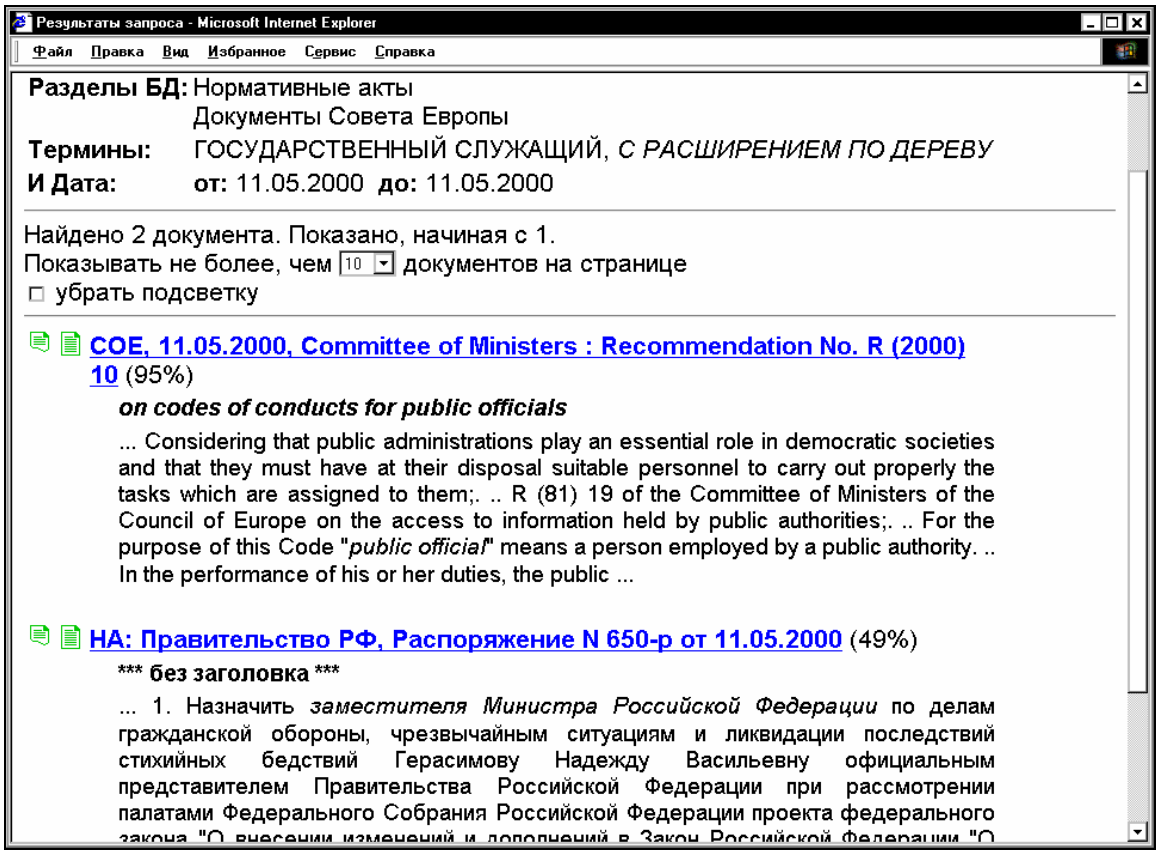


**Рис. 2. Результаты поиска в многоязычной коллекции**

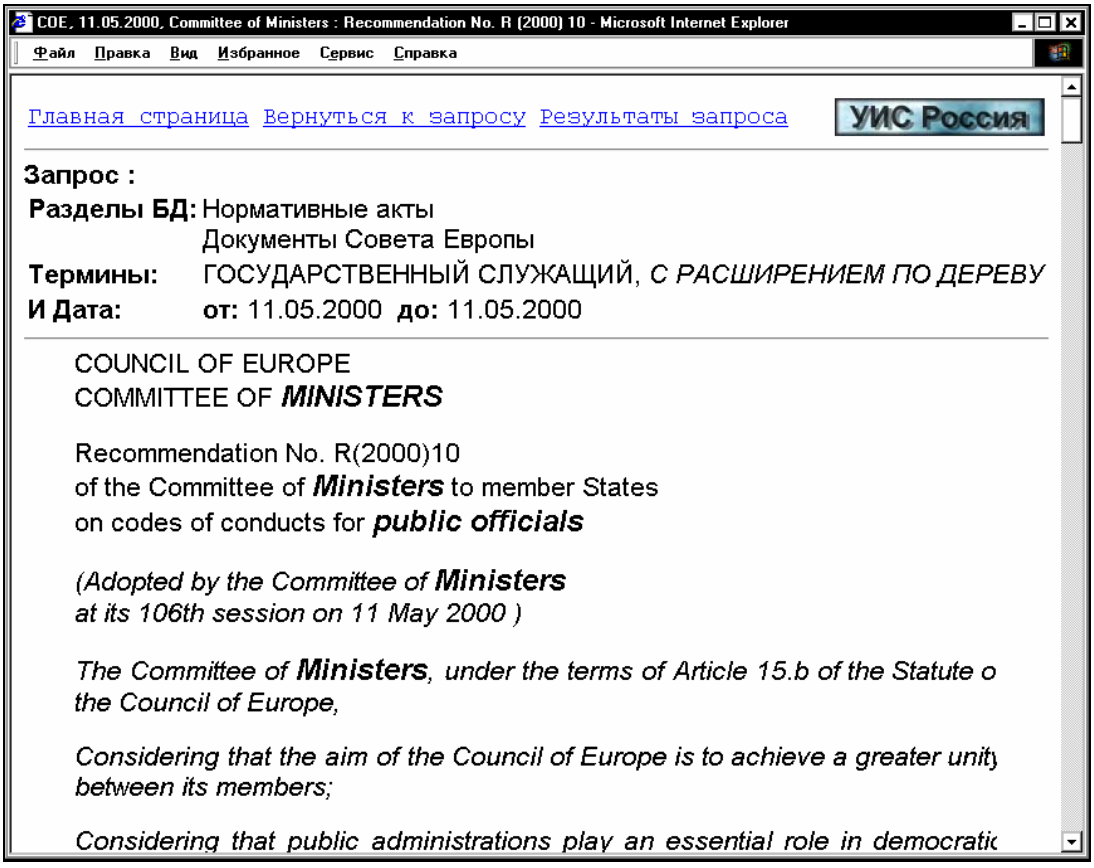


**Рис. 3. Обоснование результатов поиска**

При тематическом поиске важное значение приобретает обоснование релевантности найденного документа, особенно для документов на другом языке. В УИС РОССИЯ релевантные запросу вхождения терминов подсвечиваются в тексте документа (см. Рис.3).

## 6. Представление содержания текста в многоязычной среде

Одной из серьезных проблем в сфере многоязычного поиска является проблема определения релевантности полученных при поиске документов исходному запросу, поскольку пользователь может не понимать или плохо понимать язык документа.

На основе двуязычного тезауруса для автоматического индексирования можно строить так называемую структурную аннотацию текста.

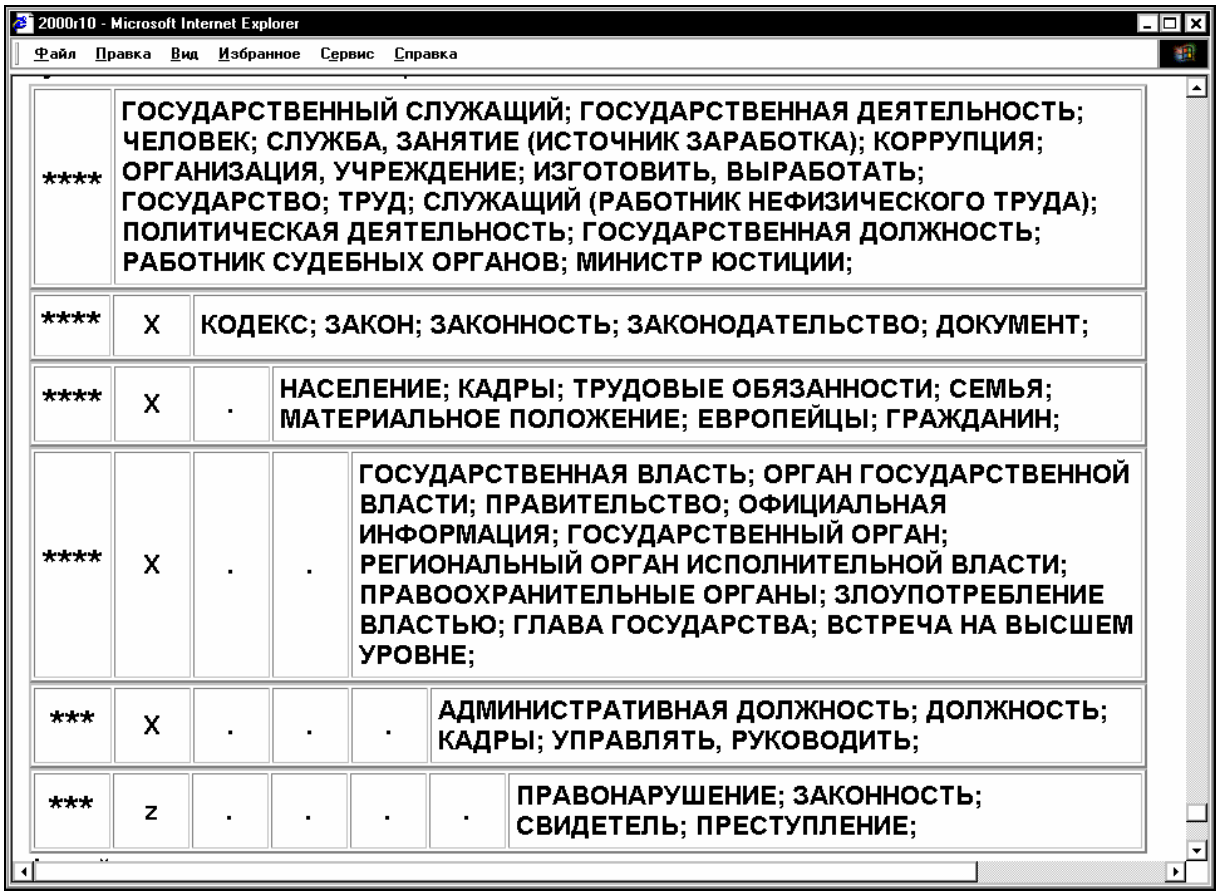


**Рис. 4. Структурная тематическая аннотация**

Структурная аннотация описывает главную тему и подтемы документа, представляя их совокупностями близких по смыслу терминов из этого документа – тематическими узлами. Она позволяет определить основное содержание текста с первого взгляда и может быть представлена на любом из языков тезауруса независимо от языка документа.

В качестве примера рассмотрим (Рис.4) структурную тематическую

аннотацию приведенной выше Рекомендации R(2000)10 Совета Европы.

Структурная тематическая аннотация включает в себя следующие части:

- термины основных тематических узлов, упорядоченных в порядке убывания частотности и расположенных горизонтально;
- отметки об относительно суммированной частотности основных тематических узлов, обозначаемые различным количеством символов “*”;
- отметки о силе взаимоотношений между различными тематическими узлами
  - ”X“-- очень сильное отношение;
  - ”z”-- сильное отношение;
  - ”.” -- отношение.

**Заключение**

Представлены принципы и основные этапы разработки двуязычного тезауруса в общественно-политической области, созданного как инструмент для автоматической обработки больших массивов текстов. Описаны возможности поиска документов в двуязычной среде в Университетской информационной системе РОССИЯ (http://www.cir.ru).

**Благодарности**

Эта работа частично выполняется при поддержке гранта № 01-07-430 Российского фонда фундаментальных исследований.

**Литература**

1. Gay F.C., Oard D.W., The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries // NIST Special Publication 500-250: The Tenth Text Retrieval Conference (TREC-2001). p.16-25. (http://trec.nist.gov/pubs/trec10/papers/clirtrack.pdf)
2. Gonzalo J., Language Resources in Cross-Language Information Retrieval: a CLEF perspective // Cross-Language Information Retrieval and Evaluation: Proceedings of the First Cross-Language Evaluation Forum, LNCS, Springer-Verlag. (http://sensei.lsi.uned.es/NLP/papers/clef00.pdf)
3. Информационно-поисковый тезаурус. Русская версия тезауруса EUROVOC. В 3х томах. - Издание Государственной Думы ФС РФ, 2001.
4. Climent S., Rodriguez H., Gonzalo J., Definition of the link and subsets for nouns of the EuroWordNet project - EuroWordNet (LE2-4003), Technical Report D005. (http://sensei.ieec.uned.es/~julio/D005.ps)

5. Журавлев С.В., Юдина Т.Н., Информационная система РОССИЯ // НТИ, Сер.2. 1995. № 3. С. 18-20.
6. Лукашевич Н.В., Салий А.Д., Представление знаний в системе автоматической обработки текстов // НТИ, Сер.2. 1997. № 3. С. 1-6. (http://www.viniti.ru/cgi-bin/nti/nti.pl?action=show&year=2_1997&issue=3&page=27)
7. Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в Университетской Информационной Системе РОССИЯ // Сборник трудов Третьей Всероссийской конференции по Электронным Библиотекам: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" (RCDL'2001) - Петрозаводск, 2001. - С.78-82. (http://rcdl2001.krc.karelia.ru/papers/papers/ dobrov_lukashevich/dobrov_paper.rtf)