
# Thesaurus on Islam: Development and Current State

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

# Text Analytics

- A lot of text data
  - News flows
  - Social network messages
- Information-retrieval and information-analytical systems
  - Analysis of information in a specific domain requires special resources
- This presentation
  - Thesaurus on Islam for automatic document processing

# Outline

- Lexical and terminological resources in natural language processing and information retrieval
- Family of RuThes-like thesauri
- Thesaurus on Islam
  - Scope
  - Terms and relations
  - Applications

# Traditions of Knowledge Representation in Broad Domains

- Information-Retrieval thesauri
  - Controlled vocabulary with formalized relations for improving of information retrieval
- WordNet-like thesauri
  - Hierarchical net of lexical concepts - synsets
  - Describes the lexical system of the general lexicon
  - There were several projects of development domain-specific WordNet-like thesauri
- Formal ontologies
  - Formal description of the domain in form on concepts and relations between them
    - Concepts, instances
    - Attributes, relations
    - Axioms

# Accounting Traditions of Knowledge Representation

- Information-Retrieval thesauri
  - + representation of terms, concepts, multiword expressions, small set of relations
  - - oriented to manual work, absence of ambiguous words, weak formalization
- WordNet-like Thesauri
  - + detailed description of synonyms, senses, multilevel hierarchy
  - -problems with description of multiword expressions, deficiency of relations ,
  - - model is not for terminology representation
- Formal Ontologies
  - + concepts, formal principles of relation description
  - - formalized description is difficult to be matched with language units
  - - it is difficult to create large resources for broad domains

# RuThes Thesaurus

- Linguistic Ontology - most concepts are based on senses of real language expressions
  - Developed more than 20 years
  - Corporate-owned, now partially published
  - http://www.labinform.ru/pub/ruthes/index.htm
  - Used in various IR and NLP projects (i.e. in news processing system of Central Bank of Russia)

- Unified representation – net of concepts
  - For different parts of speech
  - For lexical units and domain terms
  - Words and multiword expressions

- Current size
  - 55 thousand concepts, 4.1 relations per concept
  - 167 thousand Russian words and multiword expressions.

# RuThes units

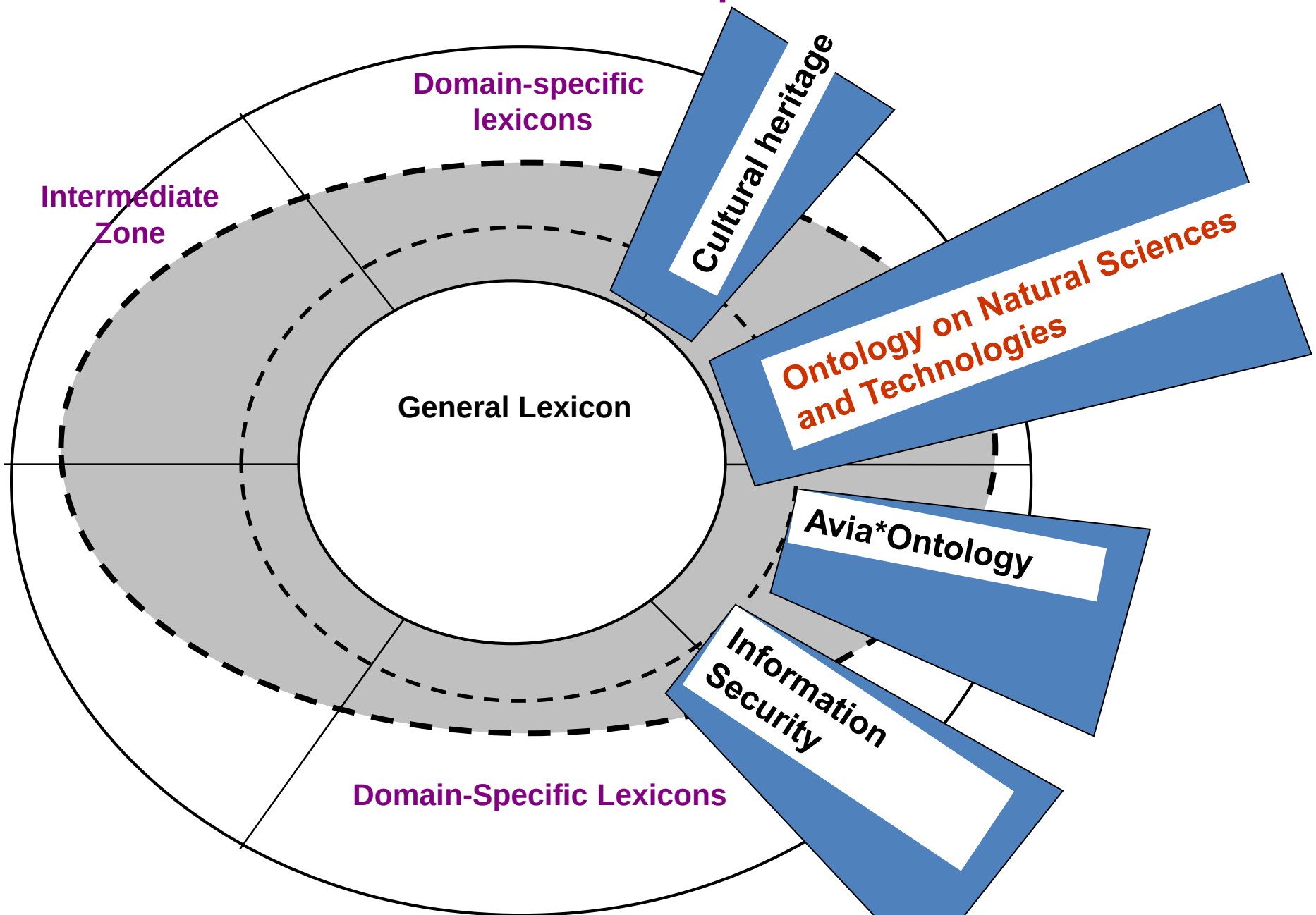
- Main principles
  - Distinguishable concepts – distinctions with neigbour concepts on denotational level
  - Concept should have an unambiguous and concise name
  - Text entries should be equivalent in respect to concept relations

- A concept unites the following language expressions (ontological synonyms):
  - words that belong to different parts of speech (*stabilization, stabilize, stabilized*)
  - linguistic expressions relating to different linguistic styles, genres
  - single words, idioms, free multiword expressions, which senses correspond to the concept

# Conceptual relations

- **Small set of relations**
  - Class – subclass
    - Transitivity, inheritance
  - Part-whole
    - Transitivity of part-whole relations
    - Described parts are strictly related to their wholes
  - External ontological dependence (Gangemi et al., 2001; Guarino, 2009)
    - Forest depends on Tree
    - Forest Fire depends in Forest
- **Main principle for establishing relations – reliable relations**
  - Concepts of lower levels of the hierarchy should be rigidly related to upper concepts

# RuThes and Domain-Specific Resources



# Thesaurus on Islam: Main Stages of Development

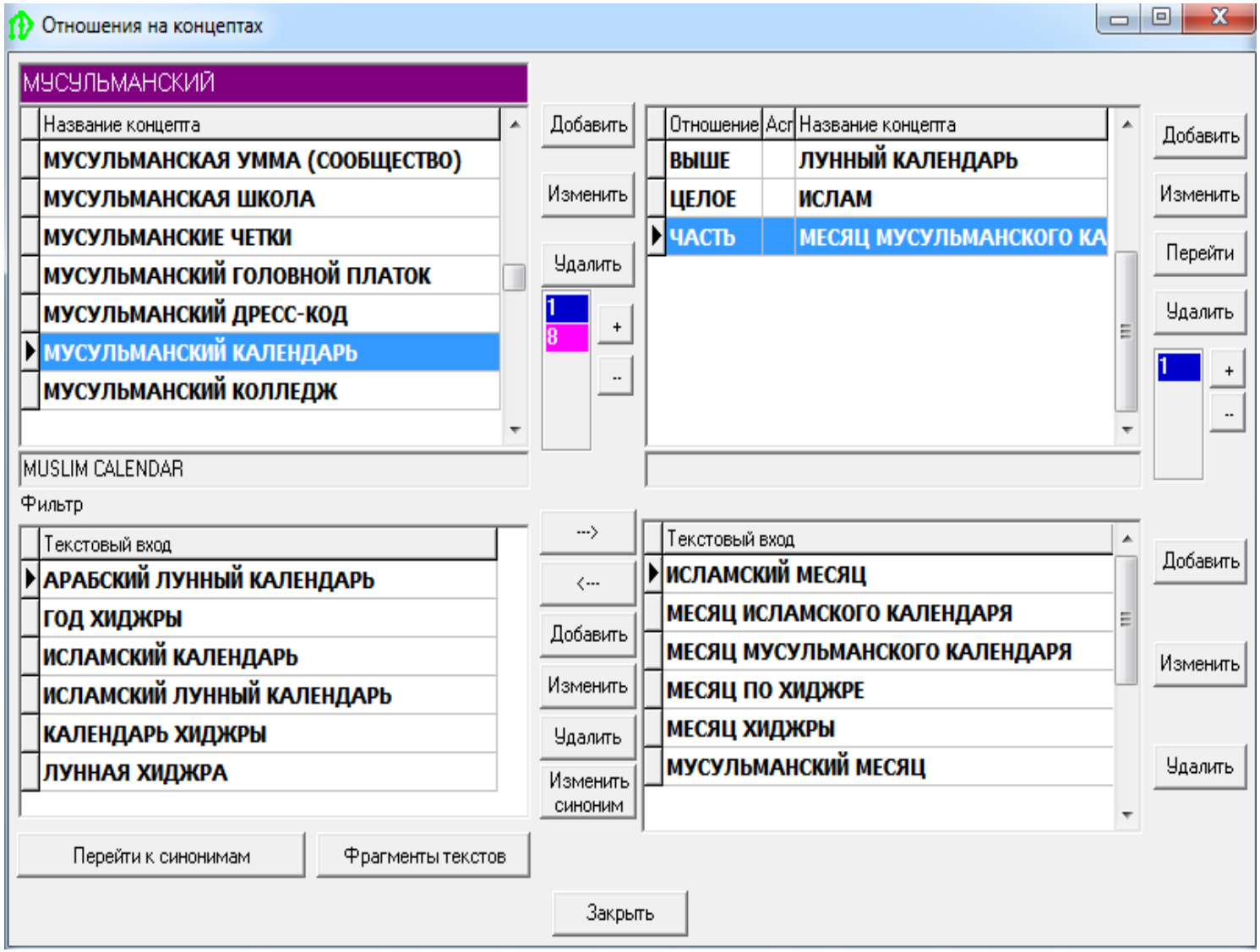
- Resource for automatic document processing of Internet-pages, news articles, or social network messages

- The main stages of the Islam Thesaurus development:
  - to find an important Islam or Muslim-related concept analyzing the results of automatic term extraction, Islam dictionaries or Islam-related texts;
  - to introduce the concept into the thesaurus providing it with a unique understandable name,
  - to provide the concept with text entries that can express it in texts (synonyms);
  - to describe relations of the introduced concept to existing thesaurus concepts.

# Scope of Islam Thesaurus

- basic concepts of Islam*;* various branches of Islam, Muslim law
- religious literature, suras of Qur'an
- forbidden actions in Islam
- Islam buildings, cult objects, worship actions, religious attributes
- Muslim calendar and holidays
- Islamic finance
- famous spiritual and military leaders of Islam
- Islam rituals, types of Muslim clothing
- Muslim education
- Islamic organizations and communities
- general concepts existing in different religions
- conflict with other religions, concepts and specific entities involved in current conflicts, various attacks, attitudes towards Islam
- the political structure of Islamic countries;
- the geography of regions with a predominantly Islamic population etc.

# Collecting various text variants for a concept

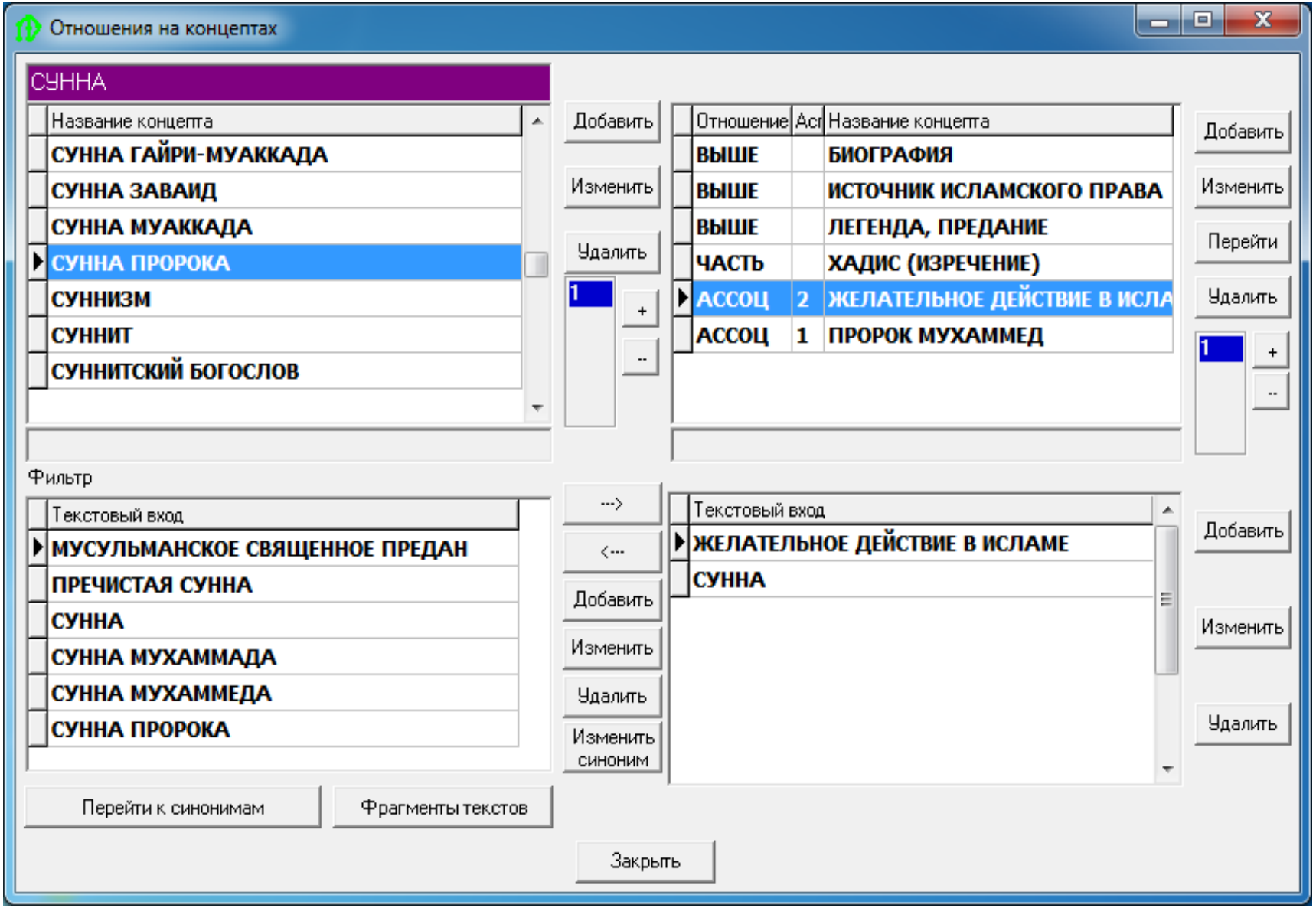


# Description of ambiguity

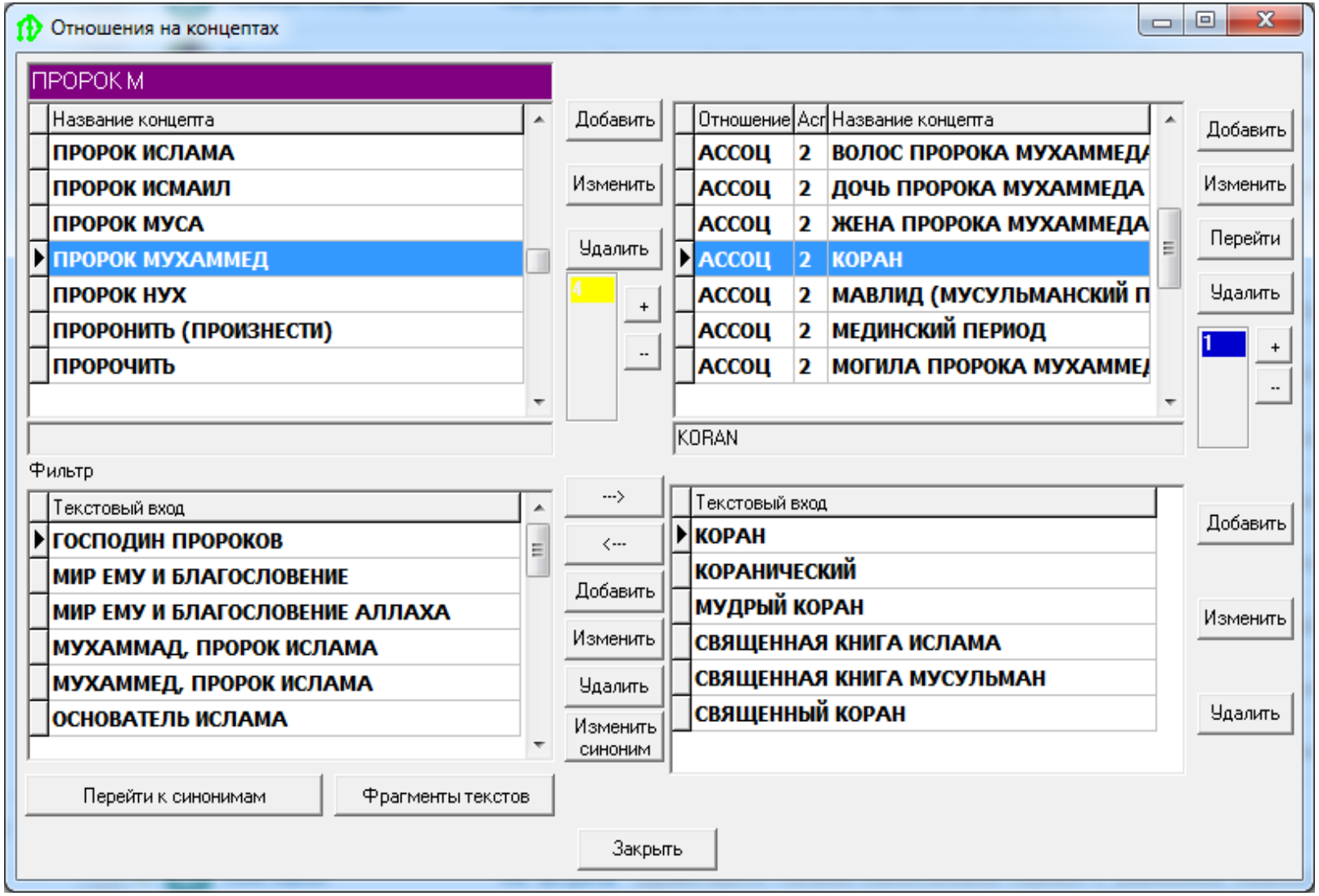
- *Hegira* (*Hijrah*)
  - 1) migration of Muhammad in the year 622,
  - 2) Islamic calendar, which was set from the date of the migration,
  - 3) mass migration of Muslims to Islamic countries from non-Islamic countries

- Introduced concepts
  - *MUHAMMAD'S HEGIRA: hegira, Muhammad's Hegira, Hijrah,*
  - *ISLAMIC CALENDAR*: *Muslim calendar, Islamic calendar Hijri calendar, Muslim moon calendar, Islamic moon calendar, year of hijri, hijri, Arab moon calendar,*
  - *MIGRATION TO MUSLIM LANDS*: *Hijrah, Hijrah mass Muslim migration*.

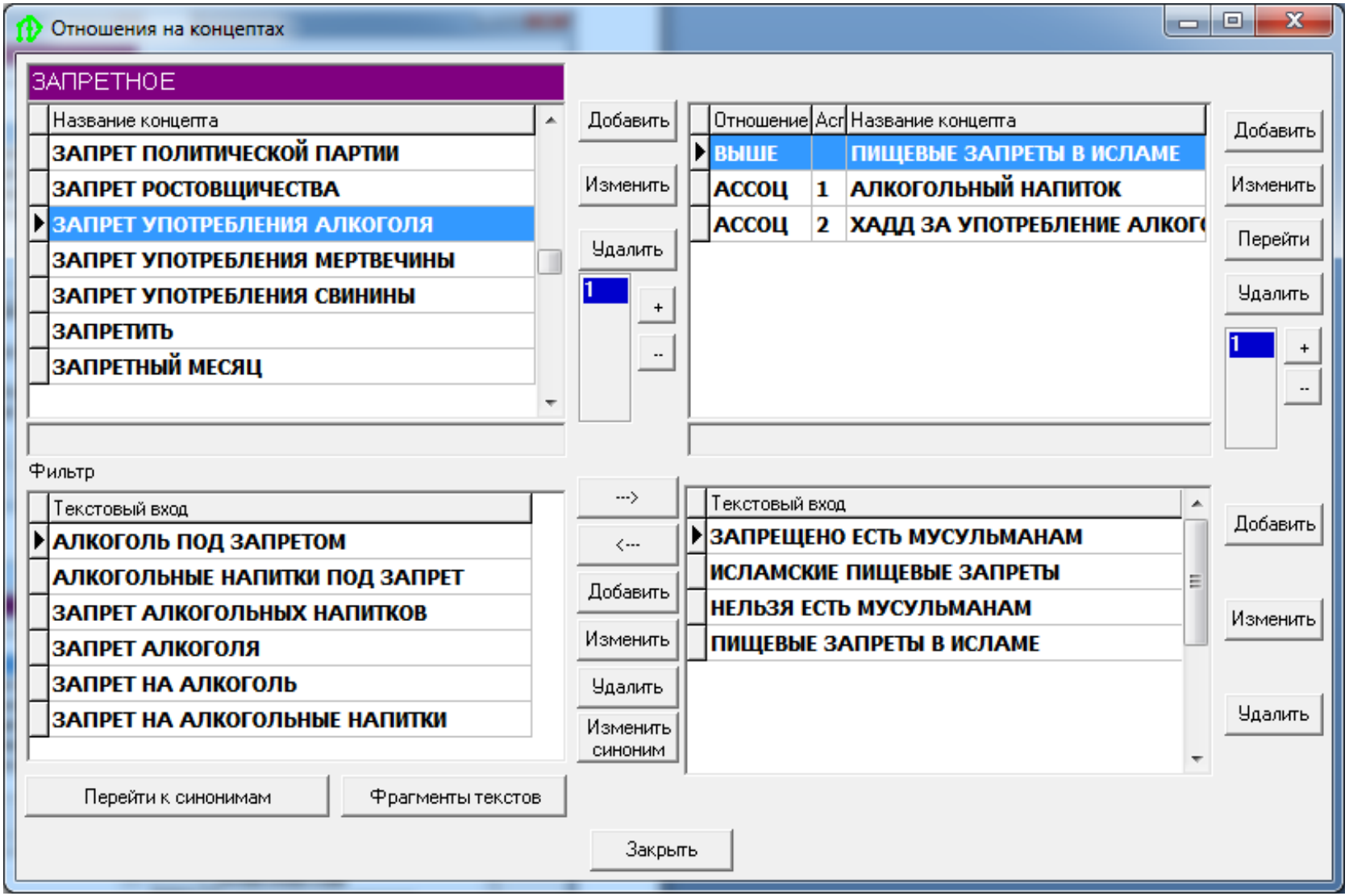
# Two senses of Sunna



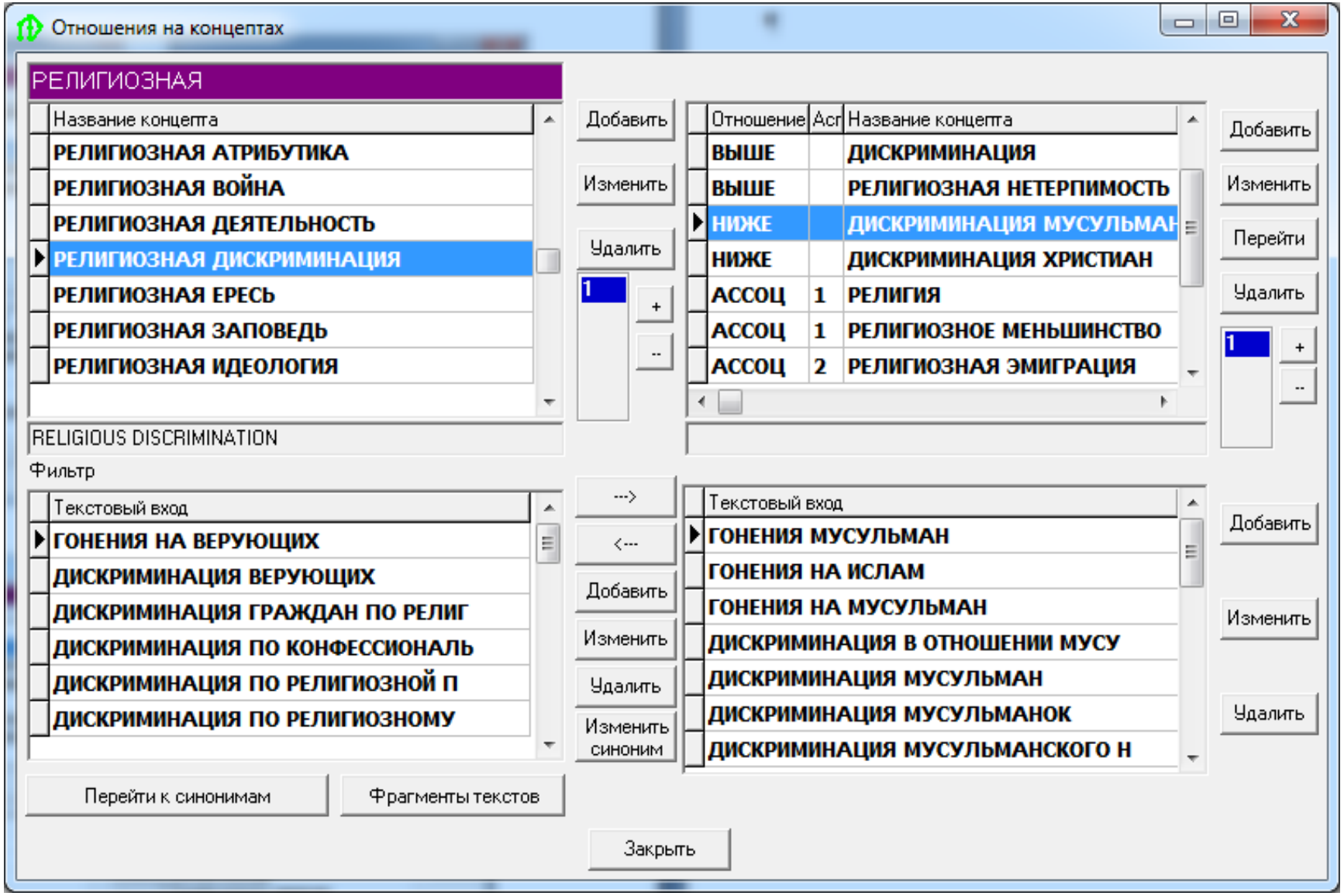
# Use of Dependence Relations



# Description of Prohibited in Islam



# Description of Relations with Other Peoples



# Text Processing: Thesaurus Matching

ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ С | Справка | Новая обработка

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Опознанные | Снять все | T | T_A | T_M | | | |
| | ФИО | Организации | | | | | |
| Сантимент | NECRF | Подсвечивать фон | | | | | |

СПИСОК НАЙДЕННЫХ РУБРИК

СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА

ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

АННОТАЦИЯ

▼ ОБРАБОТАННЫЙ ТЕКСТ

**Шиитские террористы наступают** на **Киркук**
16 октября **в 17:00 Икрамутдин Хан 2**

**Шиитские террористические формирования "Хашд аш-Шааби" в составе т.н. Иракской армии уже примерно сутки ведут наступление на Киркук, к которому они стягивали силы в предшествовавшие несколько дней.**

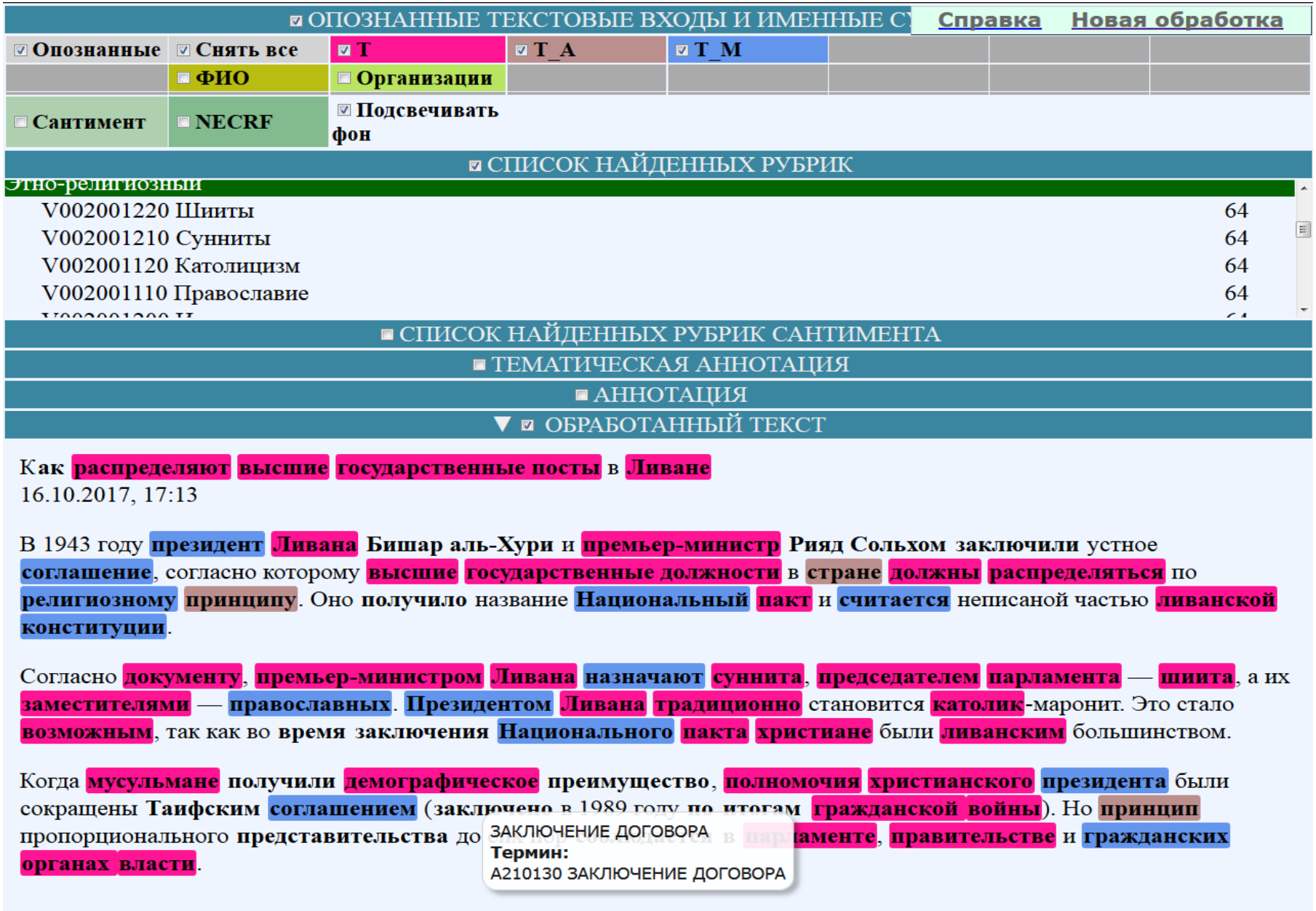
**Вооруженные силы (иракского) Курдистана - Пешмерга - оказывают им ожесточенное сопротивление.**

**Определенные успехи, которых шиитам удалось достигнуть на западе Киркука, связаны с тем, что свои позиции там оставили части Пешмерги, подконтрольные оппозиционной партии ПСК, судя по всему, вошедшей в сговор с Багдадом. В связи с этим в настоящий момент правительство Курдистана осуществляет мобилизацию и стягивают к городу силы, чтобы удержать его.**

**Руководство Курдистана заявило, что не стремится к ведению боевых действий, но и уступать город шиитским бандам не намерено.**

**Лидеры ряда стран мира призвали Эрбиль (столица Курдистана) и Багдад прекратить боевые действия и начать переговоры.**

# Automatic Text Categorization



# Content Analysis of Islam Site

- Golosislama.ru
  - Islam-related news and Islam basics
- More than 26 thousand pages were extracted

- Used thesaurus: RuThes + specialized Islam thesaurus, more than 5000 Islam related terms

- Topic models
  - 100 and 200 topics
  - Manual labeling of topics: understability of topics

| Unigram topic | Phrase-enriched topic | Thesaurus-enriched topic |
|---|---|---|
| **Syria topic (Run 1)**<br>Relation coherence **0.11** | **Syria topic (Run 5)**<br>Relation coherence **0.13** | **Syria topic (Run 12)**<br>Relation coherence **0.36** |
| сирия<br>(Syria) | сирия<br>(Syria) | сирия<br>(Syria) |
| сирийский<br>(Syrian) | башар асад<br>(Bashar al-Assad) | сирийский<br>(Syrian) |
| асад<br>(Assad) | сирийская оппозиция<br>(Syrian opposition) | асад<br>(Assad) |
| оон<br>(UN) | сирийский<br>(Syrian) | дамаск<br>(Damask) |
| оппозиция<br>(opposition) | режим асада<br>(al-Assad regime) | башар асад<br>(Bashar al-Assad) |
| башар<br>(Bashar) | асад<br>(Assad) | сирийская оппозиция<br>(Syrian opposition) |
| страна<br>(country) | сирийский режим<br>(Syrian regime) | оппозиция<br>(opposition) |
| дамаск<br>(Damask) | режим башара асада<br>(Bashar al-Assad regime) | режим асада<br>(al-Assad regime) |
| президент<br>(President) | сирийская власть<br>(Syrian authorities ) | режим<br>(regime) |

# Conclusions

- The Thesaurus on Islam for automatic document processing is presented
  - comprises the concepts of Islam and related concepts of social life.
  - contains more than 5 thousand terms.

- The thesaurus is created by the model of the thesaurus RuThes thesaurus
  - It is also assumed that the Thesaurus on Islam will be compatible with the published version of the RuThes thesaurus,
  - Various applications processing a wide range of texts, including news reports, specialized sites, posts in social networks.