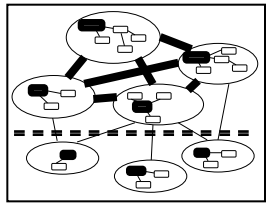


RuThes-Lite, a Publicly available version of Thesaurus of Russian Language RuThes

Loukachevitch Natalia, Dobrov Boris,
Chetviorkin Ilia

Lomonosov Moscow State University

Russian WordNets

- Automatically-generated
  - Balkova et al., 2008
    - State of the project is unknown
  - http://wordnet.ru/ (Gelfenbeyn et al., 2003)
    - direct translation without any manual revision

- Developed from scratch
  - RussNet (Azarowa, 2008)
    - State of the project is unknown
  - YARN – Yet Another RussNet (2012)

RuThes Linguistic Ontology

- Main principles and general structure
- Units and relations of RuThes
- Publication of RuThes-lite
  - Site version
  - XML-version

RuThes Linguistic Ontology

- Linguistic Ontology - most concepts are based on senses of real language expressions
  - Developed more than 15 years
  - Corporate-owned, now partially published
- Unified representation – net of concepts
  - For different parts of speech
  - For lexical units and domain terms
  - Words and multiword expressions

- Current size
  - 54 thousand concepts, 4.1 relations per concept
  - 158 thousand Russian words and multiword expressions.
  - English part: 135 thousand entries

RuThes: General Structure

- **General Lexicon**
  - concepts that can be met in various specific domains. (similar to the Factotum domain in the Wordnet domain set),
- **Sociopolitical Thesaurus**
  - thematically oriented lexemes and multiword expressions as well as domain-specific terms of the broad sociopolitical domain.
- Sociopolitical domain
  - Broad domain of contemporary social relations, entities (economy, law, finance, politics, international affairs etc.)
  - Subdomains are interwoven between each other
  - Lexical units are very closely related to domain terms
    - *Credit line, consumer credit, microcredit, account opening*

▪ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

▪ АННОТАЦИЯ

▼ ▪ ОБРАБОТАННЫЙ ТЕКСТ

Япония оказалась между **молотом** и **наковальней Япония уже** в который ра **меняет коней** на **переправе**. **Новому премьер-министру достался чудовищно разросшийся ворох экономических** и **внешнеполитических проблем** и **перспектива вооруженного конфликта мирового масштаба**. В **результате недавних парламентских выборов** к **власти вернулась Либерально-демократическая партия**. Ее **лидеру** Синдзо Абэ теперь **надлежит занять кресло премьер-министра**. Надо **заметить**, он **уже** сиживал в нем **пять лет назад**. **Продержался недолго** – не **справился** с **экономикой**, но **успел сделать кое-какие шаги** в **сторону улучшения отношений** с **Китаем**, **основательно испорченные** его **предшественником** Дзюнъитиро Коидзуми. Теперь, однако, на **повестке дня** в японо-**китайских отношениях остро стоит вопрос** о **принадлежности островов** Сэнкаку (Дяоюйдао). Сами по себе они не **представляют** никакой **ценности**, но **обладание** ими **дает право владеть обширным континентальным шельфом**, где недавно были **обнаружены большие запасы углеводородов**. Для **Японии**, практически **лишенной энергоресурсов**, это ценнейшая **находка**. Поэтому при **предшествующем правительстве государство выкупило островки** у **частных владельцев**, чем **вызвало бурю негодования** в **Китае**. Абэ **уже поспешил заявить**, что **переговоры** с **Пекином** о **принадлежности островов невозможны**.

Units of RuThes

- Main principles
  - Distinguishable concepts – distinctions with neigbour concepts on the 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

Examples of ontological synonyms

- *ДУШЕВНОЕ СТРАДАНИЕ (wound in the soul)*
- *боль, боль в душе, в душе наболело, душа болит, душа саднит, душевная пытка, душевная рана, душевный недуг, наболеть, рана в душе, рана в сердце, рана души, саднить*
- English ontological synonyms can look as:
- emotional hurt, *emotional pain, emotional wound, heartache, pain, pain in the soul, wound, wound in the heart, wound in the soul*
- *WN 3.0:* ***pain***, *painfulness (emotional distress; a fundamental feeling that people try to avoid) "the pain of loneliness"*

Synonyms and variants of "money laundering"

- *Оттывание доходов, легализация доходов, легализация денежных средств, легализация незаконных доходов, легализация средств, легализовать доходы, незаконное оттывание, оттыв, оттывать деньги, оттываться…*
- *criminal laundering, illegal laundering, laundering, laundering activities, laundering of money, laundering operations, money laundering, money laundering activities, money legalization, money washing, profit laundering, profit washing…*

- !! Variants allow better matching concepts in texts
- !! Multiword variants decrease lexical ambiguity

MWE-based Concepts

- MWE-based concept should add new information to the ontology
  - Relations do not follow from component structure
  - Interesting synonyms, high ambiguity of components etc.
    - *LEASE AGREEMENT*
      - important element of leasing procedure
      - subclass of legal agreements
    - *PAPAL ELECTION*
      - relation to *CONCLAVE*
    - *FALLING ASLEEP AT THE WHEEL*
      - interesting synonym" - *falling asleep while driving*
      - relation to *ROAD ACCIDENT*

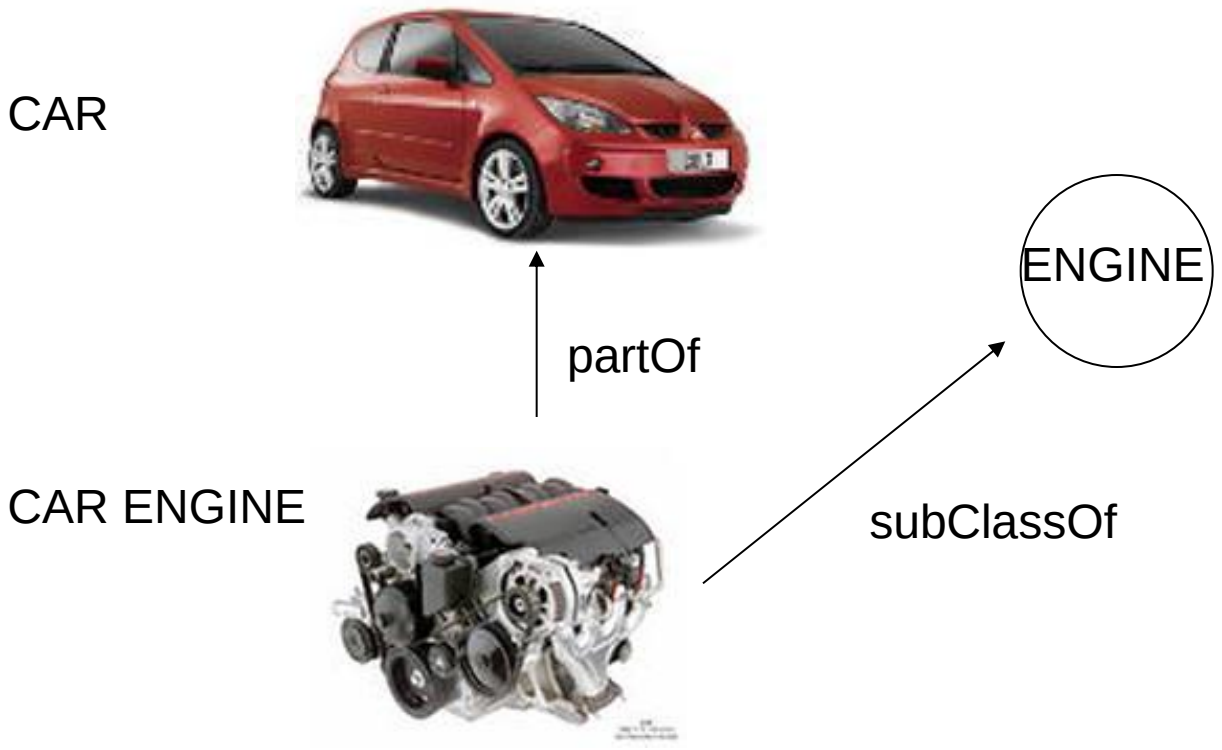
Conceptual relations

- **Small set of relations**
  - Class – subclass
    - Transitivity, inheritance
  - Part-whole
    - Transitivity of part-whole relations
  - External ontological dependence (Gangemi et al., 2001; Guarino, 2009)
    - Existence of *Car plant* depends on existence of *car*
- **Main principle for establishing relations – reliable relations**
  - Concepts of lower levels of the hierarchy should be rigidly related to upper concepts

Part-Whole Relations in RuThes

- Parts described in RuThes should be attached to their wholes
  - Existential or generic dependence of part from whole (Gangemi et al., 2001 Guizzardi, 2011)
    - Inseparable parts
    - Mandatory wholes
  - Different semantic types
    - Physical entities, elements, processes
    - Roles in processes (investor – investing)
    - Processes in spheres of activities
    - Properties of entities
    - Such a part-whole relation is close to Guarino internal relations (Guarino, 2009)

How to describe a relation between car and engine



Car engine is generically dependent from car (mandatory whole)

RuThes Linguistic Ontology: combination of three traditions

- methods of construction of information-retrieval thesauri (concept-based units, a small set of relation types, rules of multiword expression inclusion)
- development of wordnets for various languages (language-motivated units, description of ambiguous expressions)
- ontology research (concepts as main units, strictness of relations, necessity of many-step inference).

RuThes-Based Projects

- Informational-retrieval applications
  - Conceptual indexing
  - Semantic search and query expansion
  - Visualization of search results
  - Document clustering
  - Single document and multidocument summarization
  - Sentiment analysis
  - Developmnet domain-specific ontologies

- Project with
  - State Bodies
    - Central Bank of the Russian Federation (2006 – ..)
    - Central Election Committee of the RF (1999 – 2011) ...
  - Commercial organizations
    - Rambler Media company (2007– 2012)
    - Garant Legal Information Company (2002 – …)
    - Yandex (2014) …

Publication of RuThes

- At present, RuThes thesaurus is partially involved in several commercial projects and therefore it cannot be published as a whole.
- The current publicly available version of RuThes (RuThes-lite) contains around 100 thousand words and expressions and is available from http://www.labinform.ru/ruthes/index.htm.
  - Site-version
  - XML-version
- We plan to distribute RuThes-lite as free for noncommercial use (Attribution-NonCommercial-ShareAlike 3.0 Unported license).

Steps for generation of RuThes-lite

- News collection – 2mln. Documents
- Matching with RuThes – frequency list of thesaurus text entries
- Cleaning from word fragments, names, specific terms
- 30 thousand the most frequent words and expressions were taken
- The following concepts were included to RuThes-lite
  - Concepts corresponding to the chosen text entries
  - All upper-level concepts

Список текстовых входов

С БЛЕСКОМ	С ВЕДОМА
С ГЛАЗУ НА ГЛАЗ	С ГЛЯНЦЕМ
С ДУШИ ВОРОТИТ	С КОНВЕЙЕРА СОЙТИ
С КОНВЕЙЕРА СХОДИТЬ	С ЛЕГКОСТЬЮ
С ЛИХВОЙ	С МИРОВЫМ ИМЕНЕМ
С НАТУРЫ	С НОГ НА ГОЛОВУ ПЕРЕВЕРНУТЬ
С НОГ НА ГОЛОВУ ПОСТАВИТЬ	С ПЕРЕЛИВАМИ
С ПОЛУСЛОВА	С ПРАВИЛЬНЫМИ ПРОПОРЦИЯМИ
С ПРОХЛАДЦЕЙ	С ПЬЯНЫХ ГЛАЗ
С РАСПРОСТЕРТЫМИ ОБЪЯТИЯМИ	С РИСУНКОМ
С РОГАМИ	С СИЛОЙ НАВАЛИТЬСЯ
С ТАКТОМ	С ТВЕРДЫМИ ПРИНЦИПАМИ
С ТРЕСКОМ ПРОВАЛИТЬСЯ	С ТРИ КОРОБА
С ТРУДОМ	С ТРУДОМ ВЫБРАТЬСЯ
С ТРУДОМ ВЫЛЕЗТИ	С ТРУДОМ ПРОНИКНУТЬ
С УДОВОЛЬСТВИЕМ ЖДАТЬ	С УЗОРОМ
С ХОРОШЕЙ ТЕХНИКОЙ	С ХОРОШЕЙ ФИГУРОЙ
С ХОРОШИМИ ПЕРСПЕКТИВАМИ	С-ПЕТЕРБУРГ
С/Х	С/Х ЖИВОТНОЕ
С/Х ЗЕМЛИ	С/Х ИНВЕНТАРЬ
С/Х МАШИНА	С/Х МАШИНОСТРОЕНИЕ
С/Х ОБОРУДОВАНИЕ	С/Х ОРГАНИЗАЦИЯ
С/Х ОРУДИЯ	С/Х ПОКАЗАТЕЛЬ

Текстовый вход: *САД*

ДЕТСКИЙ САД

(*ДЕТСАД*, *ДЕТСАДИК*, *ДЕТСАДОВСКИЙ*, *ДЕТСКИЙ САД*, *САД*, *САДИК*, *САДОВСКИЙ*, *САД-ЯСЛИ*, *ЯСЛИ-САД*)

ВЫШЕ ДОШКОЛЬНОЕ УЧРЕЖДЕНИЕ
ЧАСТЬ ЯСЛИ

САД (УЧАСТОК ЗЕМЛИ)

(*САД*, *САДИК*, *САДОВЫЙ*)

ВЫШЕ ЗЕМЕЛЬНЫЙ УЧАСТОК
АССОЦ$_1$ САДОВАЯ КУЛЬТУРА
АССОЦ$_2$ БЕСЕДКА
АССОЦ$_2$ САДОВНИК
АССОЦ$_2$ САДОВОДСТВО

ДОШКОЛЬНОЕ УЧРЕЖДЕНИЕ

(*ДЕТСКОЕ ДОШКОЛЬНОЕ УЧРЕЖДЕНИЕ*, *ДОШКОЛЬНОЕ ДЕТСКОЕ УЧРЕЖДЕНИЕ*, *ДОШКОЛЬНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ*, *ДОШКОЛЬНОЕ УЧРЕЖДЕНИЕ*, *ДОШКОЛЬНЫЙ КОМБИНАТ*, *УЧРЕЖДЕНИЕ ДОШКОЛЬНОГО ВОСПИТАНИЯ*)

ВЫШЕ	ДЕТСКОЕ УЧРЕЖДЕНИЕ
ВЫШЕ	СОЦИАЛЬНО-КУЛЬТУРНОЕ УЧРЕЖДЕНИЕ
ЦЕЛОЕ	ДОШКОЛЬНОЕ ВОСПИТАНИЕ
НИЖЕ	ДЕТСКИЙ САД
НИЖЕ	ЯСЛИ
АССОЦ$_1$	ДОШКОЛЬНИК

XML-version

- File of concepts
  - Identifier, name, gloss from Wiktionary
- File of relations
- File of text entries
  - Dictionary form
  - Lemmatized form
  - Syntactic type
  - Main word
  - Parts of speech for every word
- File of synonyms: concept – text entry relations

Conclusion

- RuThes-lite linguistic ontology was published
- Plans
  - Revision and preparation of larger versions
  - Generation of a Russian thesaurus in WordNet form for comparison
  - Alignment with Princeton WordNet
  - Experiments and applications