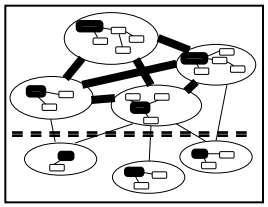


Creating Russian WordNet by Conversion

Natalia Loukachevitch (Moscow)

G. Lashevich (Kazan), A.Gerasimova (Moscow),

V. Ivanov (Kazan), B. Dobrov (Moscow)

WordNet

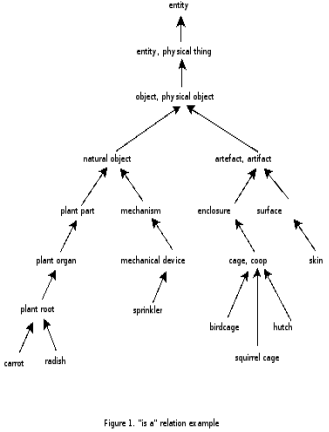


Figure 1. "is a" relation example

- Popular resource for NLP applications
- Hierarchical net of synsets
  - Four nets for each part of speech (nouns, verbs, adjectives, adverbs)
  - Different sets of relations
  - Cross-links
- Projects of creating wordnets in many languages
- This talk:
  - semi-automatic conversion of data from another thesaurus RuThes into WordNet structure
  - Why? People want to have a wordnet for their own language

Outline

- Russian wordnet projects
- WordNet Structure
- RuThes structure, applications
- Creating RuWordNet
- Publication of resources

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)
    - Crowdsourcing, use of Wiktionary
    - https://russianword.net/
    - Mainly contains synsets, often «naive»
      - *Казах, казахстанец*
      - *Инъекция, прививка*

WordNet Noun Relations

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **yard**, pace (a unit of length equal to 3 feet; defined as 91.44 centimeters; originally taken to be the average length of a stride)
  - *part meronym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *part holonym*
  - *derivationally related form*
- S: (n) **yard**, grounds, curtilage (the enclosed land around a house or other building) *"it was a small house with almost no yard"*
- **S: (n) yard (a tract of land enclosed for particular activities (sometimes paved and usually associated with buildings)) *"they opened a repair yard on the edge of town"***
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
- S: (n) thousand, one thousand, 1000, M, K, chiliad, G, grand, thou, **yard** (the cardinal number that is the product of 10 and 100)
- S: (n) cubic yard, **yard** (a unit of volume (as for sand or gravel))
- S: (n) **yard** (a tract of land where logs are accumulated)
- S: (n) **yard**, railway yard, railyard (an area having a network of railway tracks and sidings for storage and maintenance of cars and engines)
- S: (n) **yard** (a long horizontal spar tapered at the end and used to support and spread a square sail or lateen)
- S: (n) **yard** (an enclosure for animals (as chicken or livestock))

WordNet Adjective Relations

Adjective

- S: (adj) **light** (of comparatively little physical weight or density) "
  "magnesium is a light metal--having a specific gravity of 1.74 at
  - *similar to*
  - ***attribute***
    - S: (n) weight (the vertical force exerted by a mass as
  - *antonym*
  - *derivationally related form*
- S: (adj) **light**, light-colored ((used of color) having a relatively sm
  agent) *"light blue"; "light colors such as pastels"; "a light-colored*
- S: (adj) **light** (of the military or industry; using (or being) relativel
  equipment) *"light infantry"; "light cavalry"; "light industry"; "light*
- S: (adj) **light** (not great in degree or quantity or number) *"a light*
  accent"; "casualties were light"; "light snow was falling"; "light m
  smoke from the chimney"
- S: (adj) **light** (psychologically light; especially free from sadness
  heart"
- S: (adj) **light** (characterized by or emitting light) *"a room that is l*
  are open"; "the inside of the house was airy and light"
- S: (adj) unaccented, **light**, weak ((used of vowels or syllables) p

WordNet Verb Relations

Verb

- **S: (v) give (cause to have, in the abstract sense or physical *him a black eye"; "The draft gave me a cold"***
- S: (v) yield, **give**, afford (be the cause or source of) *"He gave m "Our meeting afforded much interesting information"*
- S: (v) **give** (transfer possession of something concrete or abstra *gave her my money"; "can you give me lessons?"; "She gave t and tender loving care"*
  - *direct troponym* / *full troponym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *cause*
  - *phrasal verb*
  - *antonym*
  - *derivationally related form*
  - *sentence frame*
- S: (v) **give** (convey or reveal information) *"Give one's name"*
- S: (v) **give**, pay (convey, as of a compliment, regards, attention, *pay him any mind"; "give the orders"; "Give him my best regard*
- S: (v) hold, throw, have, make, **give** (organize or be responsible *reception"; "have, throw, or make a party"; "give a course"*

RuThes Linguistic Ontology

- Linguistic Ontology - most concepts are based on senses of real language expressions
  - Developed more than 20 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
  - 164 thousand Russian words and multiword expressions.
  - English part: 138 thousand entries

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
  - Development 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) …

Units of RuThes

- Main principles
  - Distinguishable concepts – distinctions with neighbor 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*
- *but:*
- *WN 3.0:* ***pain***, *painfulness (emotional distress; a fundamental feeling that people try to avoid) "the pain of loneliness"*

RuThes 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

Generating RuWordNet

- Source: RuThes-lite 2.0
  - 115 thousands words and expressions
- Division to part of speech nets
  - Semi-automatic parsing of RuThes entries to obtain morpho-syntactic representations
  - Division to three nets
  - POS-synonyms
- Providing WordNet relations

RuWordNet statistics

Part of speech	Number of synsets	Number of unique entries	Number of senses
Noun	29296	68695	77153
Verb	7634	26356	35067
Adjective	12864	15191	18195

Part of speech	Hypernyms	Instance-class	Wholes	Pos-synonymy	Antonyms
Noun	39155	1863	10010	18179	455
Verb	10440	0	117	7451	20
Adjective	17834	66	14139	14139	457

RuWordNet: Noun Relations

- Hyponym-hypernym
- Instance-hypernym (geographical locations)
- Antonyms (properties and states)
- POS-synonymy
- Part-whole relations
  - functional parts (*nostrils* – *nose*),
  - ingredients (*additives* – *substance*),
  - geographic parts (*Sevilia* – *Andalusia*),
  - members (*monk* – *monastery*),
  - dwellers (*Moscow citizen* – *Moscow*),
  - temporal parts (*gambit* – *chess party*)

RuWordNet: Adjective Relations

- hyponym-hypernym relations
  - Hierarchies as in GermaNet and Polish wordnet

- Antonyms
- POS-synonymy links to noun and verb synsets:
  - word *строительный* – POS links
  - to the noun synset {*стройка, постройка, возведение, сооружение*..}
  - to the verb synset {*строить, построить, возводить* ...}.

RuWordNet: Verb Relations

- Hyponyms-hypernyms
- Antonyms
- POS-synonymy
- Part-whole relations
  - Princeton WordNet entailment relation
  - {*видеть во сне, сниться, грезиться, присниться, привидеться во сне, пригрезиться, пригрезиться во сне"*} [*to dream*] - {*спать, поспать, доспать, соснуть, досыпать, почивать, проспать, просыпать*} [*to sleep*]
  - {*оппонировать, оппонировать диссертацию*} - {*защитить диссертацию*}

Accessibility of RuThes and RuWordNet

- RuThes web-site
  - http://www.labinform.ru/pub/ruthes/index.htm
- RuWordNet web-sites
  - http://www.labinform.ru/pub/ruwordnet/index.htm
  - ruwordnet.ru

- Xml-files can be obtained non-commercial use

RuThes Web Representation

Текстовый вход: *ДВОР*

УСАДЬБА

(*ДВОР*, *ПОДВОРЬЕ*, *УСАДЕБНЫЙ*, *УСАДЬБА*)

ВЫШЕ	ЗАГОРОДНЫЙ ДОМ
ВЫШЕ	ИНДИВИДУАЛЬНЫЙ ЖИЛОЙ ДОМ
ЦЕЛОЕ	ПОМЕСТЬЕ
НИЖЕ	ХУТОР
ЧАСТЬ	ПРИУСАДЕБНЫЙ УЧАСТОК

ДВОР (ЗЕМЕЛЬНЫЙ УЧАСТОК)

(*ДВОР*, *ДВОРИК*, *ДВОРОВЫЙ*, *ДВОРОВЫЙ УЧАСТОК*, *ПРИДОМОВЫЙ УЧАСТОК*)

ВЫШЕ	ЗЕМЕЛЬНЫЙ УЧАСТОК
НИЖЕ	ГОРОДСКОЙ ДВОР
НИЖЕ	ПРИУСАДЕБНЫЙ УЧАСТОК

КОРОЛЕВСКИЙ ДВОР

(*ДВОР*, *КОРОЛЕВСКИЙ ДВОР*, *ЦАРСКИЙ ДВОР*)

ВЫШЕ	ОКРУЖЕНИЕ, ОКРУЖАЮЩИЕ ЛЮДИ

RuWordNet Web Representation

Текстовый вход: *ДВОР*

Синсет: *ДВОР* *УСАДЬБА* *ПОДВОРЬЕ*

ГИПЕРОНИМ	*ДОМ ДЛЯ ОДНОЙ СЕМЬИ* *ДОМ УСАДЕБНОГО ТИПА* *ОДНОСЕМЕЙНЫЙ ДОМ* *ИНДИВИДУАЛЬНОЕ ЖИЛЬЕ* *ИНДИВИДУАЛЬНЫЙ ЖИЛОЙ ДОМ* *ИНДИВИДУАЛЬНЫЙ ЖИЛИЩНЫЙ ФОНД* *ИНДИВИДУАЛЬНОЕ ДОМОВЛАДЕНИЕ*
ГИПЕРОНИМ	*ЗАГОРОДНЫЙ ДОМ*
ГИПОНИМ	*ХУТОР* *ХУТОРОК*
ЦЕЛОЕ	*ИМЕНИЕ* *ПОМЕСТЬЕ*
ЧАСТЕРЕЧНЫЙ СИНОНИМ	*УСАДЕБНЫЙ*
ЧАСТЬ	*ПОДВОРЬЕ* *ПРИУСАДЕБНАЯ ЗЕМЛЯ* *ПРИУСАДЕБНЫЙ УЧАСТОК* *ПРИУСАДЕБНЫЙ ЗЕМЕЛЬНЫЙ УЧАСТОК*

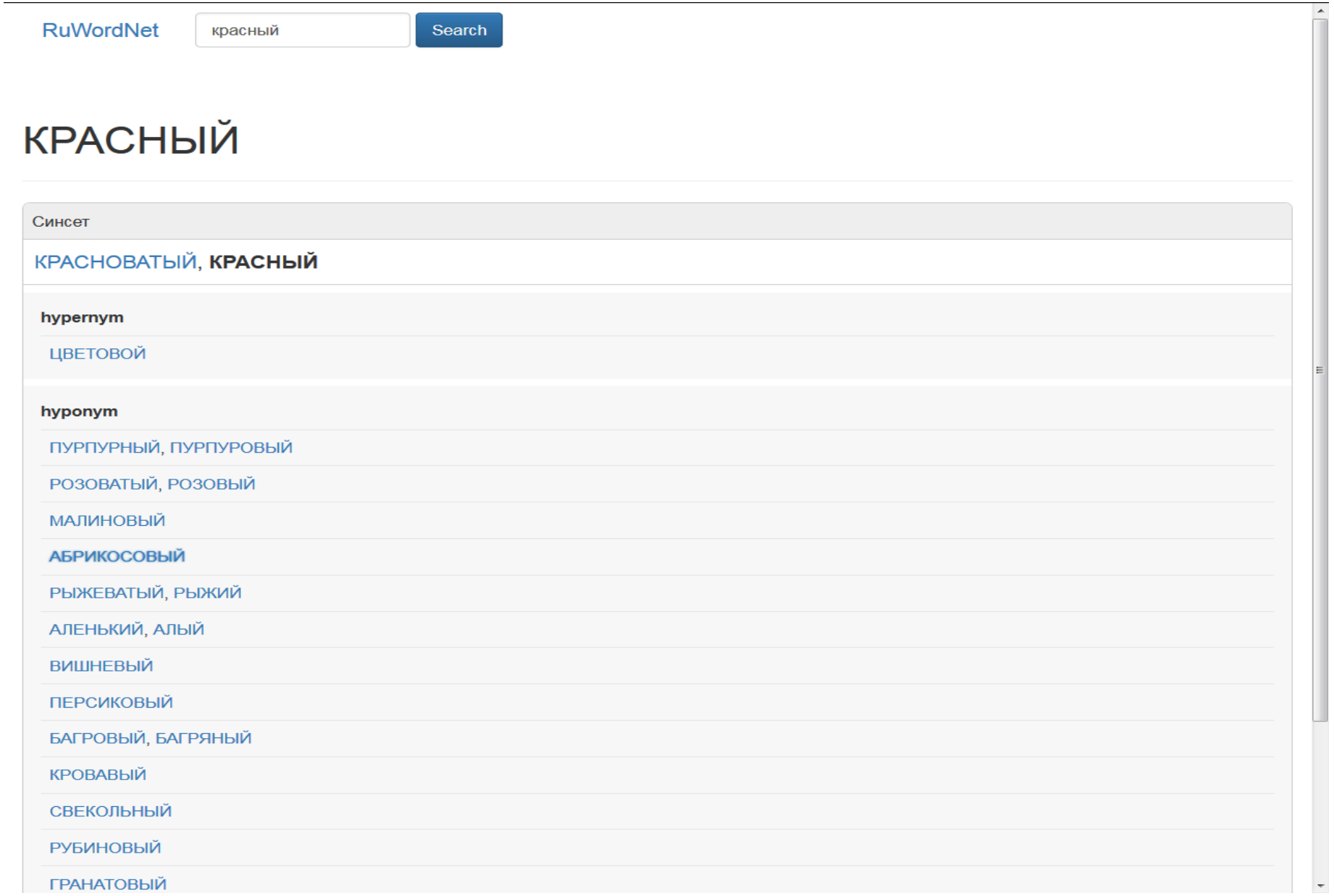
Синсет: *ДВОР* *ЦАРСКИЙ ДВОР* *КОРОЛЕВСКИЙ ДВОР*

ГИПЕРОНИМ	*КРУГ* *СРЕДА* *БЛИЗКОЕ ОКРУЖЕНИЕ* *ОКРУЖЕНИЕ* *БЛИЖАЙШЕЕ ОКРУЖЕНИЕ* *ОКРУЖАЮЩИЕ ЛЮДИ*

Синсет: *ДВОР* *ДВОРИК* *ДВОРОВЫЙ УЧАСТОК* *ПРИДОМОВЫЙ УЧАСТОК*

ГИПЕРОНИМ	*ЗЕМЛЯ* *НАДЕЛ* *НАДЕЛ ЗЕМЛИ* *ДЕЛЯНКА* *ЗЕМЛИЦА* *УЧАСТОК СУШИ* *ДЕЛЯНКА ЗЕМЛИ* *УЧАСТОК ЗЕМЛИ* *ЗЕМЕЛЬНЫЙ НАДЕЛ* *ЗЕМЕЛЬНЫЙ УЧАСТОК*
ГИПОНИМ	*ПОДВОРЬЕ* *ПРИУСАДЕБНАЯ ЗЕМЛЯ* *ПРИУСАДЕБНЫЙ УЧАСТОК* *ПРИУСАДЕБНЫЙ ЗЕМЕЛЬНЫЙ УЧАСТОК*
ГИПОНИМ	*ДВОРОВАЯ ТЕРРИТОРИЯ* *ГОРОДСКОЙ ДВОР* *ПРИДОМОВАЯ ТЕРРИТОРИЯ*
ГИПОНИМ	*ЗАДНИЙ ДВОР*

Ruwordnet.ru provides word search



Conclusion

- We have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet.
- In this procedure we attempted to achieve two main characteristic features of wordnet-like resources:
  - division of data into part-of-speech-oriented structures with cross-references between them
  - providing a set of relations similar to wordnet-like relations
- Both thesauri, RuThes-lite 2.0 and RuWordNet, are currently published as static web-pages.
  - Researchers can obtain both types of thesauri, compare them in applications