Research Computing Center
of Moscow State University

NCO Center for Information Research

Concept Formation in Lingustic Ontologies

Natalia V. Loukachevitch

louk@mail.cir.ru

Plan of Presentation

- **Formal ontologies and natural language**
- **Linguistic ontologies (WordNet, MikroKosmos)**
  - **Problems of concept formation based on meanings of language expressions**
  - **Why it is important to know**
- **Thesaurus RuThes – linguistic ontology**
  - **the largest Russian ontological resource**
  - **How to do a linguistic ontology more formal, more ontological**
- **Conclusions for formal ontologies developers and advocates of formal approaches**

Are ontologies fully independent from a natural language?

- **In all known ontologies the words are used to represent concepts**
  - **Names of the concepts in ontologies are often formulated in natural language**
  - **Wilks et. al .: The symbols in representation languages are fundamentally based on the natural language**
  - **That which can not be captured by words cannot be represented in an ontology**
- **In specific domains most ontologies are based on domain-specific terminology**
  - **The knowledge is hidden in texts**

Linguistic Ontology?

- **Linguistic ontology is an ontology, concepts of which are considerably related to the meanings of linguistic units, the terms of the subject field**
- **A linguistic ontology can be considered as a special kind of a lexical database and a special type of an ontology**
- **They belong to the “terminological” ontologies according to J. Sowa**
- **The role of the ”linguistic ontologies” increases in applications related to natural language processing**
- **Examples: WordNet, MicroKosmoc, information-retrieval thesauri**

General Principles of Concept Formation in Ontologies

- **One needs to distinguish the concept and its name - synonyms do not represent different classes**
- **A child concept should be distinctly different from its parent**
- **A concept should be clearly distinguished from the concepts at the same level**

- **Concepts of linguistic ontologies?**
  - **relations between the concepts and lexical meanings are quite complex**
  - **creation of any ontology deals more or less with lexical or terminological meanings**

Units in Linguistic Ontologies

- **WordNet synsets – sets of synonyms**
  - **Synonyms can substitute each other in sentences**

- **MikroKosmos**
  - **distinction of an ontology and a lexicon**
  - **Small ontology, large lexicon**
  - **A lexicon entry can contain a reference to an ontology concept and features of a particular lexical unit**

- **Information-Retrieval Thesauri**
  - **based on the terms of a subject field,**
  - **national and international standards,**
  - **an indexing term is a concept presentation, preferably in the form of a noun or a noun phrase**

Problems of Concept Formation in Linguistic Ontologies

- **It is not easy to distinguish a concept and its names**
- **How to represent closely-related senses of ambiguous words with distinguishable concepts**
- **How to represent senses of near-synonyms with distingishable concepts**
- **Is it possible to do a linguistic ontology more ontological, to form distinguishable concepts?**

Confusion of concepts and its names in linguistic ontologies

- **different hierarchies for different parts of speech (*adorn* – *adornment*)**
- **different synsets to describe**
  - **the old and new names,**
  - **the names of concepts in different dialects of the language, in different text genres**

  WordNet: additional synset for "nose"

  - ***beak, honker, hooter, nozzle, snoot, snout, schnozzle, schnoz -- (informal terms for the nose)***
- **Confusion is based on definition of 'synset' as a representative of synonyms**

Confusion of Concepts and its names in WordNet: another example

- ***franc -- (the basic monetary unit in many countries; equal to 100 centimes)***
- **similarity between different monetary units is only their names**
- **they are different in value**
- **concepts should be introduced for such entities as Swiss franc, French franc, American dollar, Canadian dollar**

Representation of near-synonyms in linguistic ontologies

- **Near-synonyms – words with similar meanings**
- **Near-synonyms can differ in many features: denotative content, language register, evaluation, dialect, collocations, etc.**
- **In another language the corresponding set of near-synonyms is characterized by its own system of parametric differences**
- **Example of near-synonyms: e*rror, fault, omission, oversight, blunder, mistake, miss, screw-up, dereliction, defect***
- **How can such a set be subdivided into distinguishable concepts?**

Near- synonyms in WordNet

- **four different synsets denoting *likeness, similarity***

- **each next synset is a hyponym of the previous:**
- ***sameness*** **–**
  (the quality of being alike)
- ***similarity*** **–**
  (the quality of being similar)
- ***likeness, alikeness, similitude*** **–**
  (similarity in appearance or character or nature between persons or things)
- ***resemblance*** **–**
  (similarity in appearance or external or superficial details)

Near-synonyms in MikroKosmos ontology

- **The main principle: to unite near-synonyms under the same concept**
- **All the "change" verbs are assigned to the same concept of *CHANGE_EVENT*.**
- **Features of the words are described in lexicon entries**
  - ***increase*** **- concept *CHANGE_EVENT***
    - ***theme: scalar value***
    - ***the value is changed to a larger one***
  - ***zionist*** **- concept *POLITICAL_ROLE***
    - **in vocabulary*: AGENT_OF a SUPPORT_EVENT,***
    - **the theme of which *is Israel*.**
  - ***to asphalt - concept COVER_EVENT***
    - **instrument of which *is asphalt concept***

Near-synonyms in MikroKosmos ontology: problems

- **The Ontology is small, the vocabulary is large**
- **Problems:**

1) Inconsistency: *CHANGE_EVENT*

- **In vocabulary*: acclimatization, commerzialization, contamination, damage, deteriorate, improve***
- **Separate concepts (subclasses): *ADJUST, CORRECT-EVENT, DIVIDE, INTEGRATE, RESTRUCTURE***

2) *Acclimatization, commerzialization, contamination* are presented as synonyms!?

3) Domain-specific applications:

- **the relatively small size of the ontology leads to the introduction of additional concepts even for words that are already included in the lexicon**

Representation of similar meanings of ambiguous words

- **Automatic disambiguation of lexical senses is a difficult task**
- **WordNet: differences of senses are too fine-grained for many applications**
  - **No relations between similar senses**
  - **Experiments on sense clustering**
- **MikroKosmos ontology:**
  - **it is necessary to unite as many meanings as possible**
- **Is it possible to cluster similar senses to reduce the disambiguation problem?**
- **Ambiguity is an important problem for any approach in ontology developing**

Problems of clusterization of similar senses

- 1) **WordNet: intensive research on methods for clustering senses**
  - **Ch.Fellbaum: the sense clustering can be based on a variety of alternative criteria**
  - **Appropriate sense clustering depends on an application**
- **2) Clusterization of similar senses leads to violation of the ontological structure, e.g. to confusion of their relations**
  - **N. Guarino: several ontologies treat the *Window* concept as an artifact (*The workers mounted a window)* and**
  - **as an opening (*A man looked out of the window)* at the same time.**

Concept formation in linguistic ontologies: alternatives

- **WordNet**
  - **Synsets – confusion of concepts and their names**
  - **A lot of related senses of ambiguous words**
  - **No relation between such senses**
  - **Near-synonyms are arbitrarily split to synsets**

- **MikroKosmos**
  - **the ontology and the vocabulary are divided**
  - **Related senses of ambiguous words are generalized to the same concept**
  - **Near-synonyms are generalized to the same concept**
  - **But: overgeneralization, violation of the ontological structure, problems in applications**

Thesaurus RuThes as a linguistic ontology

- **RuThes is a hierarchical net of concepts**
- **Most concepts are based on meanings of real language expressions**
- **RuThes includes more than 50 thousand concepts and more than 140 thousand Russian words and multiword expressions.**
- **It was translated into English and comprises almost 130 thousand English words and expressions**

- **RuThes is used in information-retrieval applications:**
  - **conceptual indexing;automatic text categorization, document clustering, automatic text summarization,**
  - **question-answering.**

Unit of RuThes is not a Synset

- **Main principles**
  - **Distinguishable concepts,**
  - **Concept should have a clear, nonambiguous and concise name**
  - **Text entries should be equivivalent in respect to concept relations**

- **A concept unites the following language expressions:**
  - **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, the meanings of which correspond to this concept**

Examples of concept text entries

Concept *JUDICIAL COURT:*

court
court authorities
court instance
court of judiciary
court of jurisdiction
court of justice
court of law

judicature
judicial bodies
judicial court
judicial organ
judicial tribunal
law court
tribunal

Concept *PRIVATIZATION:*

privatisation
privatise
privatize
privatization of public assets

privatization of state property
privatization process, transfer to public ownership

How to subdivide a set of near-synonyms to distinguishable concepts

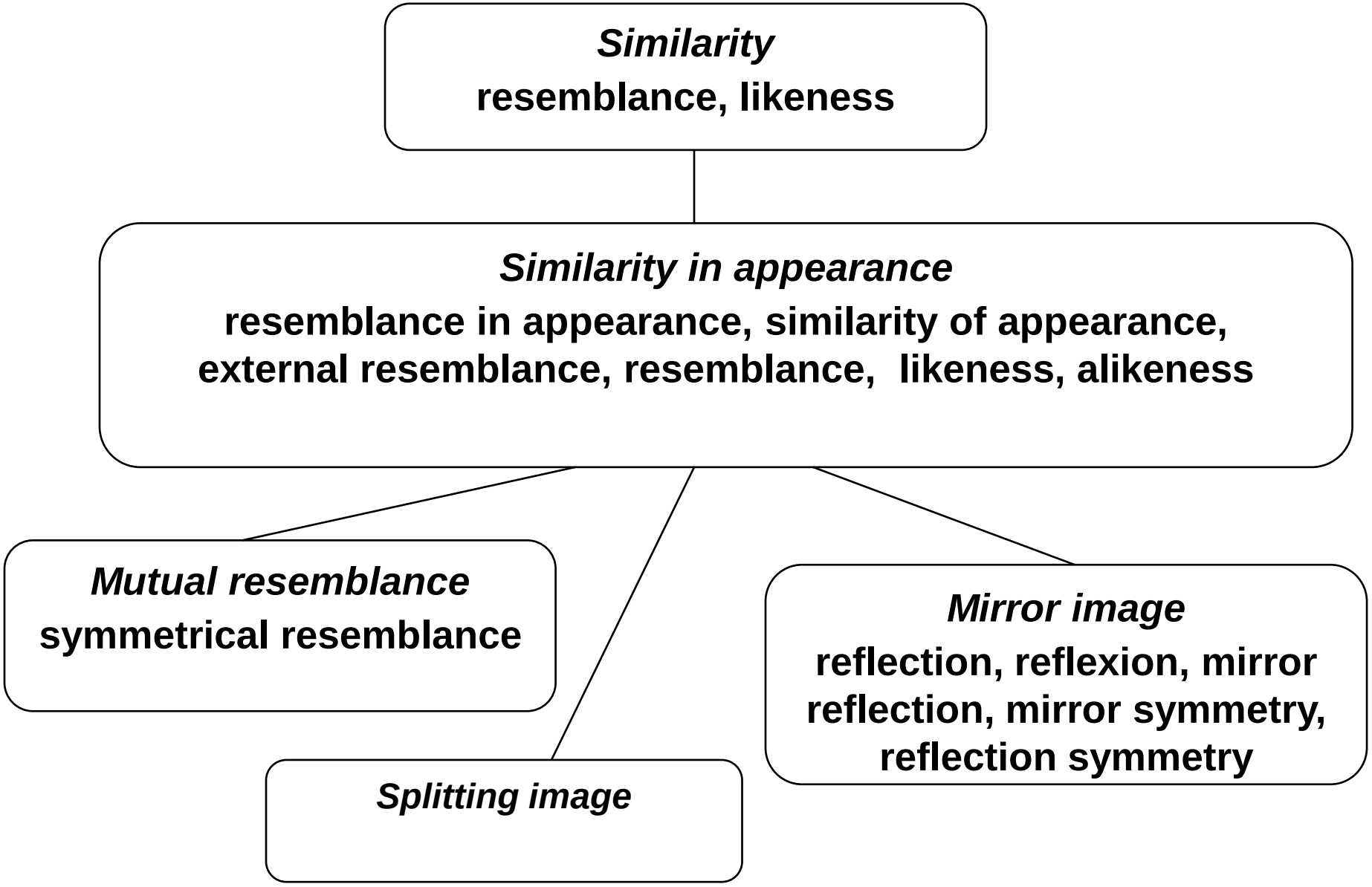
Main principles:

- **to find important features of near-synonyms,**
- **to find existing unambiguous language expressions (usually multiword expressions) with the corresponding meaning,**
- **to introduce a new concept with unambiguous name**
- **to assign language expressions to the introduced concept**

‘Similarity’ near-synonyms

Steps:

- **Identify important components of the meaning: *similarity of external characteristics***
- **Formulate the name of a new concept – *Similarity in appearance***
- **Find various ways to express this concept: *resemblance in appearance, similarity of appearance, external resemblance***
- **words with vague meanings as *resemblance* and *likeness should be related to concepts Similarity (general) and Similarity in appearance***


Similarity
resemblance, likeness
Similarity in appearance
resemblance in appearance, similarity of appearance, external resemblance, resemblance, likeness, alikeness
Mutual resemblance
symmetrical resemblance
Mirror image
reflection, reflexion, mirror reflection, mirror symmetry, reflection symmetry
Splitting image

How to represent closely related senses of ambiguous words

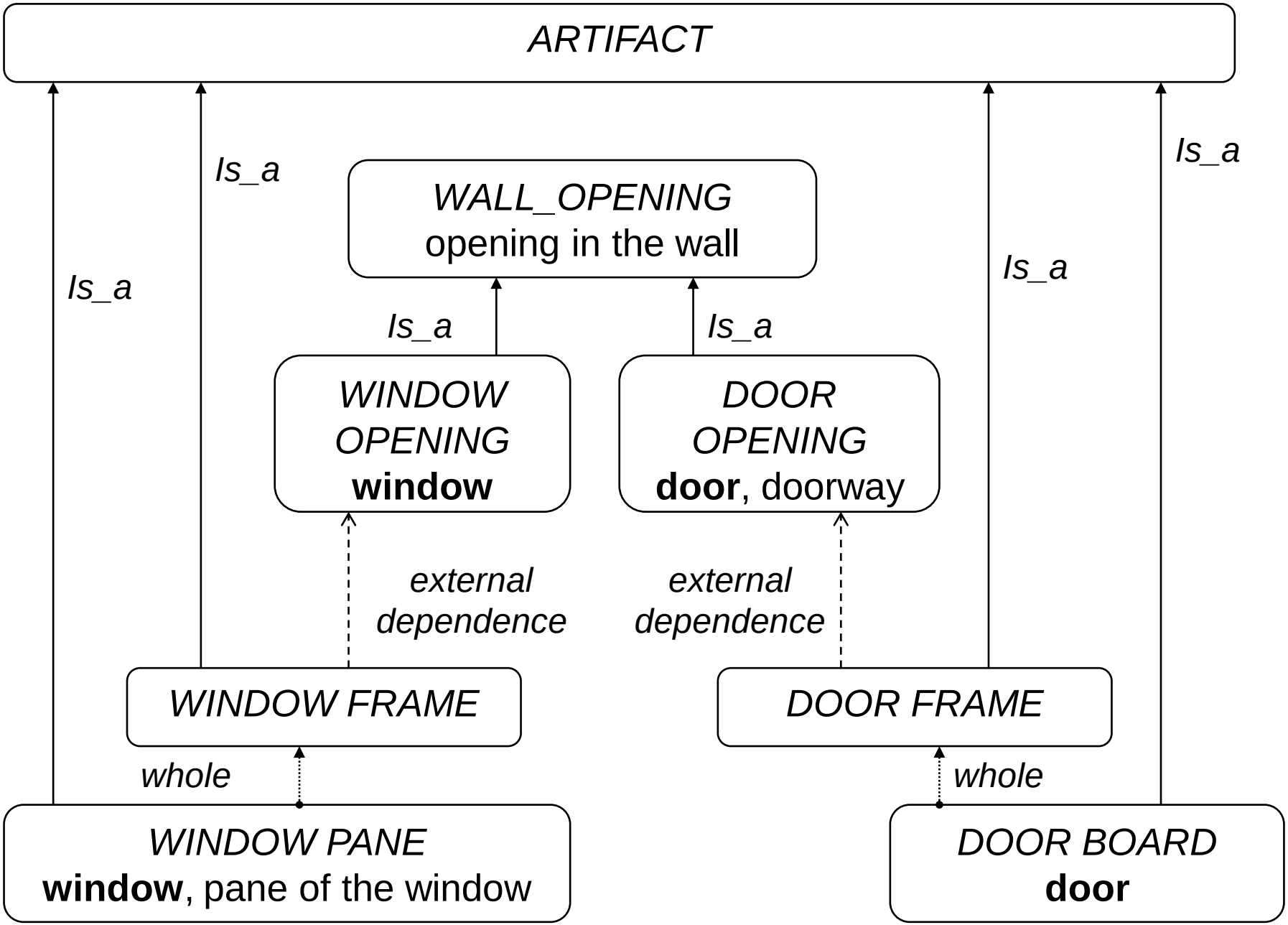
Main principles:

- **Try not to cluster senses, but to create distinguishable concepts**
- **Try to find clear, not dependent on the context distinctions**
  - **Unambiguous synonyms (e.g. multiword expressions)**
  - **Specific relations**
- **Try to provide various text entries to facilitate disambiguation**
- **To describe ontological relations between closely-related senses**

How many senses should window have

- ***Window (opening)*** **and** ***Window (artifact)***
  - **distinctly different entities**
  - **arise and exist for some time independent of each other**
- **Window (Opening) can be expressed as *window opening* (351,000 pages in Google),**
- **Window (artifact) can be expressed as *window pane* (697,000 pages in Google).**
- **If to cluster senses**
  - ***window opening*** **becomes a synonym of *window pane***
  - **the relations of these different entities are confused**

Window in Thesaurus RuThes



Formal Concept Analysis and Linguistic Ontologies

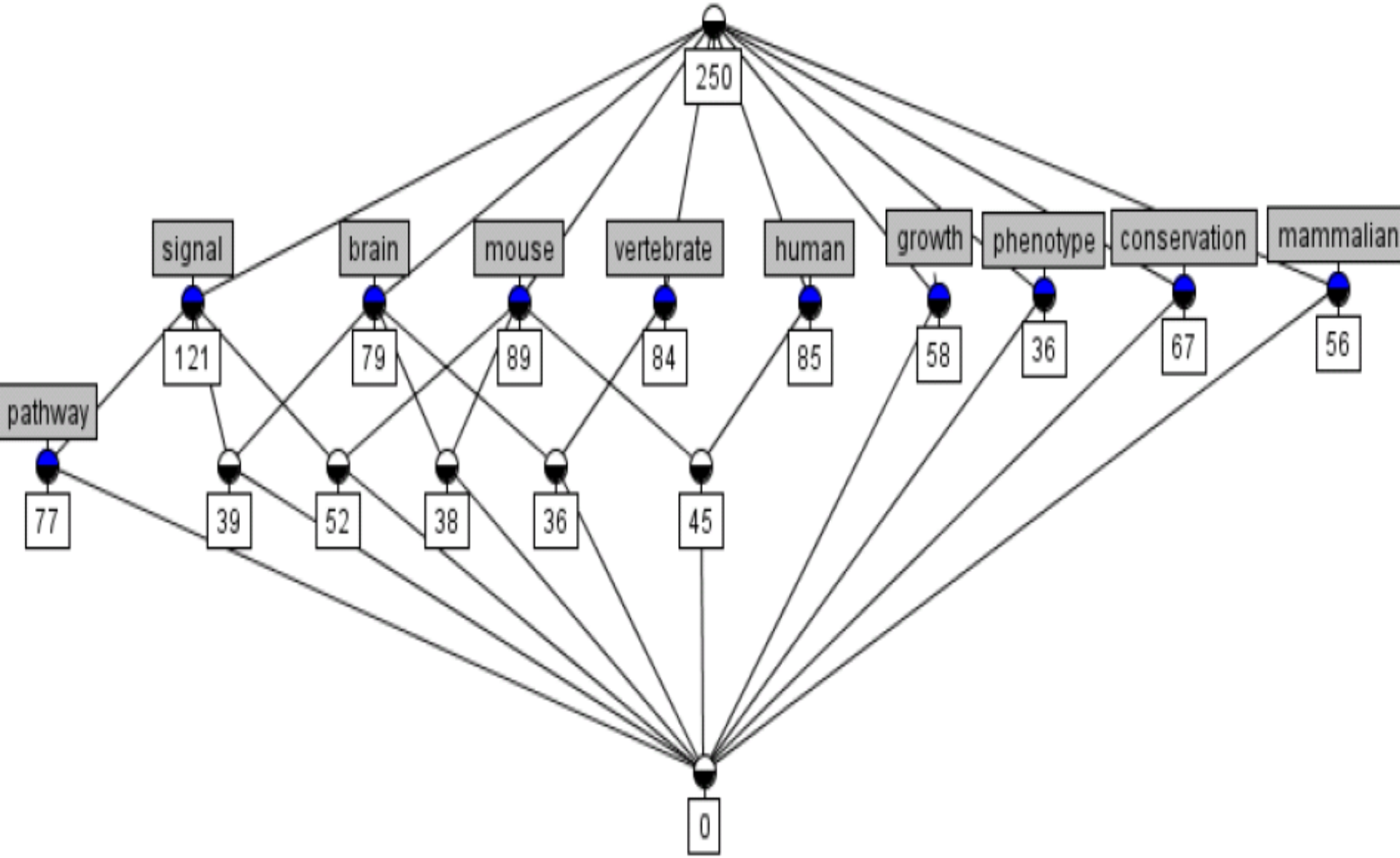
- **Huge number of formal concepts obtained from formal contexts. How to select the most important formal concepts of a lattice?**
- **It is necessary to distinguish different entities behind the same ambiguous word**
- **Unstable attributes are not a good basis for formal concepts formation**

How to select important concepts?

- **An additional condition of choosing the most important concepts:**
  - **existence of words or multiword expressions with corresponding meanings.**
  - –
- **Procedure:**
  - **Take intermediate concepts of a lattice,**
  - **Generate candidate phrases**
  - **Check their usage using Internet search engines**

- **Interesting experiment (?): To compare mathematical methods of concept selection and existence of linguistic names**

Intermediate concepts have linguistic names



- *BRAIN SIGNAL (39), MOUSE BRAIN (38), SIGNAL PATHWAY (77), VERTEBRATE BRAIN (36)*

Ambiguous words in formal concept analysis

- **It is necessary to distinguish different entities behind the same ambiguous word**
- **A lattice should not be constructed on ambiguous entities**
- **Synonymic phrases can help to reveal ambiguity.**
- **In different contexts word *nation* can denote**
  - **Political nation (*'state, country'*)**
  - **Ethnic nation (*ethnicity)***

Two concepts behind the same ambiguous word

Conclusion-1

- **Ontology developers can hardly avoid the influence of linguistic meanings, linguistic polysemy**
  - **The names of concepts and relations in ontologies have mnemonic names**
  - **The knowledge in many subject fields is hidden in texts**

- **It is important for ontology developers to understand the problems related to the formation of concepts on the basis of linguistic meanings, namely:**
  - **The problem of distinguishing the concept and its name**
  - **The problem of presenting closely related meanings of ambiguous words**
  - **The problem of splitting meanings of near-synonyms into concepts**

Conclusion-2

- **In developing the RuThes as a linguistic ontology we are trying to adhere to two contradictory criteria.**
  - **we form concepts of the thesaurus as close as possible to the meanings of linguistic units**
  - **we try to introduce distinguishable concepts**
- **Exploitation of really existing unambigous multiword expressions helps us mitigate these contradictory requirements.**