ArtsSemNet: From Bilingual Dictionary To Bilingual Semantic Network

Ivanka Atanassova  
University of Veliko Tarnovo “St. Cyril and Methodius”

Svetlin Nakov  
Sofia University “St. Kliment Ohridski”

Preslav Nakov  
University of California at Berkley
Introduction

- Computer dictionaries replace the paper ones
  - Electronic dictionaries
  - Thesauri
- Semantic networks – the next evolutionary step
  - Incorporate dictionaries and thesauri
- Our work progressed in the same fashion:
  - We started with a bilingual electronic dictionary.
  - Later we transformed it into a semantic network with various terminological relations – polysemy, homonymy, synonymy, antonymy and hyponymy.
**ArtsDict: Bilingual Terminological Dictionary**

- Tool for creation, editing and usage of parallel bilingual electronic terminological dictionaries
- Used to build domain-specific terminological dictionaries
  - Languages supported:
    - Bulgarian
    - Russian
  - Dictionary entries consist of:
    - Terms
    - Glosses explaining term sense(s)
  - The most complete terminological dictionaries of fine arts (not only electronic but also in general)
ArtsDict: Structure

- ArtsDict workspace is split horizontally to separate the two parallel dictionaries
- Each of the dictionaries is split horizontally:
  - Terms stay on the left side
  - Glosses stay on the right side
- Terms are:
  - Single-term words
  - Multiple-term words
- Terms listed along with their:
  - Origin, singular, doublets, variants, stylistic synonyms
- Support for polysemy
### Български речник:

<table>
<thead>
<tr>
<th>Дума</th>
<th>Значение на думата</th>
</tr>
</thead>
<tbody>
<tr>
<td>Гравюр</td>
<td>1. В изобразителното изкуство: вид графика, който обхваща произведения, получени чрез отпечатване от гравирана плоча; 2. Отделно произведение от този вид графика.</td>
</tr>
<tr>
<td>Гравюрска игла, Литографска игла</td>
<td></td>
</tr>
<tr>
<td>Гравюра (рус. гравюра, от фр. gravure)</td>
<td></td>
</tr>
<tr>
<td>Гравюра на дърво, Ксилиография,</td>
<td></td>
</tr>
<tr>
<td>Гравюра на камък</td>
<td></td>
</tr>
<tr>
<td>Гравюра на линолеум, Линогравюра</td>
<td></td>
</tr>
</tbody>
</table>

### Руски речник:

<table>
<thead>
<tr>
<th>Дума</th>
<th>Значение на думата</th>
</tr>
</thead>
<tbody>
<tr>
<td>Тематическая художественная выс</td>
<td>1. Темперная краска - красочный материал для живописи с эмульсионным составом связующего вещества, преимущественно в виде смеси масла или масляного лака с водным клеевым раствором; 2. Темперная техника - техника живописи, выполняемая темперными красками; 3. Произведение живописи, выполненное темперными красками в соответствующей технике.</td>
</tr>
<tr>
<td>Темная гамма</td>
<td></td>
</tr>
<tr>
<td>Темные тона (ед. ч. тон)</td>
<td></td>
</tr>
<tr>
<td>Темные цвета (ед. ч. цвет), Темные</td>
<td></td>
</tr>
<tr>
<td>Темпера (ит. tempera)</td>
<td></td>
</tr>
<tr>
<td>Темпера казеиново-масляная, Казе</td>
<td></td>
</tr>
</tbody>
</table>

**Клавиатурни комбинации:**

- **F1** - добавя българска дума
- **F2** - добавя руска дума
- **Enter** - добавя дума
- **F3** - търси дума
- **F4** - променя думата
- **Tab** - сменя речника
- **Alt-F4** - изход
ArtsSemNet – Terminological Semantic Network

- ArtsSemNet is a lexical reference system for the fine arts terminology in Bulgarian and Russian.
- The terms are organized into a semantic network on the base of several important lexical relations:
  - Polysemy
  - Homonymy
  - Synonymy
  - Antonymy
  - Hyponymy
- Provides a specialised term browser for search and navigation between the terms and the corresponding terminological relations.
ArtsSemNet: Bulgarian Fine Arts Terminology

- The Bulgarian lexical information consists of:
  - 2900 lexemes
  - 276 hyponym chains
  - 157 antonym chains
  - 483 absolute synonym chains
  - 136 relative synonym chains
  - 14 homonyms

- Polysemous terms distribution:
  - 2571 with 1 sense
  - 273 with 2 senses
  - 49 with 3 senses
  - 4 with 4 senses
  - 2 with 5 senses
  - 1 with 6 senses
ArtsSemNet: Russian Fine Arts Terminology

- The Russian lexical information consists of:
  - 2664 lexemes
  - 283 hyponym chains
  - 134 antonym chains
  - 458 absolute synonym chains
  - 114 relative synonym chains
  - 6 homonyms

- Polysemous terms distribution:
  - 2313 with 1 sense
  - 263 with 2 senses
  - 56 with 3 senses
  - 9 with 4 senses
  - 2 with 5 senses
  - 1 with 7 senses
**ArtsSemNet: Creation**

- We used *ArtsDict* for the extraction of homonyms, synonyms and polysemous terms.
- For the extraction of hyponyms and antonyms we used two techniques:
  - A *formal technique* for extraction of hyponyms sharing a common term-element (suffix/stem, affix, …)
    - Given a target term-element *ArtsDict* generates a list of terms from the dictionary that contain it
    - The list is manually examined afterwards
  - A *semantic technique* (based on LSA) for extraction of hyponyms/antonyms with no common term-element
Latent Semantic Analysis

- *Latent Semantic Analysis (LSA)*
  - Popular statistic technique for indexing, retrieval and analysis of textual data
  - Assumes mutual latent dependency between the term and its context

- LSA works in two stages
  - *Learning*: Given a collection of texts LSA builds a real-valued vector for each term and for each text
  - *Analysis*: The proximity between two texts/terms is calculated as the dot product between their normalised vectors (i.e. cosine).
Finding Hyponyms and Antonyms: The Semantic Technique

- Semantic technique for hyponyms extraction
  - LSA is given as target a hypernym or its glosses
  - LSA produces a ranked list of terms sorted according to the semantic proximity to the target hypernym
  - The list is manually examined for hyponyms

- Passing to LSA the terms only, the glosses only or both are several variants we found useful
**ArtsSemNet: Functionality**

- **Searching for terms:**
  - Exact and inexact searching

- **Browsing the terms:**
  - Term glosses list
  - Homonyms list
  - Absolute synonyms list, relative synonyms list
  - Antonym chains
  - Hyponym chains (with target term hypernym or cohyponym)

- **Supports changes between languages:**
  - Russian and Bulgarian are supported
Query results for word "Надълъжна гравюра":

Word:
Надълъжна гравюра

Meaning:
Техническа разновидност на гравюрата на дърво, при която влажната на дървеното блокче, върху което се изпълнява гравюрата, са успоредни на повърхността му.

Homonyms:
No homonyms found.

Absolute synonyms:
Надълъжна гравюра – Обрезна гравюра

Relative synonym:
No relative synonyms found.

Antonyms:
Надълъжна гравюра, Обрезна гравюра – Напречна гравюра, Щорцов гравюра

Hyponym chains:
Гравюра на дърво, Ксилография, Дърворез (техническа разновидност) – Кироскуро – Надълъжна гравюра. Обрезна гравюра – Напречна гравюра, Щорцов гравюра.
**Related Work**

- **WordNet**
  - *WordNet* is a general lexical reference system for English language organised as a semantic network.
  - The lexemes in *WordNet* are represented as *synsets*.
  - A *synset* groups a term along with some of its synonyms which as a whole represent some particular sense of that term.
  - Different senses of the same term have different synsets.
  - Relations between *synsets* include: hyponymy, meronymy, synonymy, antonymy and others.

- **EuroWordNet and BalkaNet**
  - Similar systems for European and Balkan languages.
  - Support cross-language synset index.
**ArtsSemNet and WordNet**

- **WordNet:**
  - Terms are represented as synsets, no polysemy
  - General English lexica
  - Supports more relations: e.g. meronymy

- **ArtsSemNet:**
  - Term-centered, partial synset support
  - Domain-specific terminology
  - Supports polysemy
  - Distinguishes between absolute and relative synonyms
  - Supports homonymy
  - The user interface supports hyperlink navigation
Availability of *ArtsDict* and *ArtsSemNet*

- Both *ArtsDict* and *ArtsSemNet* are freely available for download from:
  - [http://cs.berkeley.edu/~nakov/artssemnet/](http://cs.berkeley.edu/~nakov/artssemnet/)

- The lexical database is available in two variants:
  - As a Microsoft Access .mdb file
  - As a set of SQL-scripts for creating the database schema and populating the data in the tables

**Future Work**

- Create cross-language index
- More powerful user interface, e.g. hierarchical visualisation, editor for terms and relations, etc.