Using Domain Corpora for Semi-automatic Building of a Multilingual Terminology Thesaurus

Aleš Horák, Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
http://deb.fi.muni.cz
deb@fi.muni.cz

Introduction

- NLP Centre for the Czech Office for Surveying, Mapping and Cadastre (CUZK)
- building and extending a specialized terminology thesaurus for the related domain
- 6 languages (Czech, English, German, French, Russian and Slovak)

TeZK Editing System

- web application for browsing and editing, based on DEB platform
- browsing the hypernymic tree
- detailed information for entry, enriched with corpus examples or related terms
- export entries to XLS for batch editing
- API to connect external services

Advanced features

- automatic extraction of new terms
- automatic term relations identification
- translation candidates extraction

Specialized Corpora

- corpora of domain-specific texts for each language
- SpiderLing + justext + WebBootCat
- seed terms for web search from older terminology dictionary
- corpora managed by Manatee/SketchEngine

Statistics detailing the web-crawled domain corpora

Language	Documents	Tokens	Web domains
English	8,149	40,225,064	4,946
French	5,326	15,789,761	3,291
German	3,373	9,744,313	2,220
Russian	2,914	19,015,734	1,770
Slovak	2,943	10,252,449	1,528
Czech	27,389	12,689,548	1,061

Automatic Extraction of New Term Candidates

- detecting “candidate terms”
- proposals to be checked by terminology experts and easily added to the thesaurus
- comparing the relative frequency of noun phrases in domain corpus with its relative frequency in reference corpus
- TenTen corpus family used as a reference corpora

Automatic Term Relations Identification

- identification of candidate hypernyms/broader terms
- pattern extraction from a domain corpus
- identifying lexically similar terms
- evaluation: 56% accuracy (correct hypernym among the top three candidates)

Translation Candidates Extraction

- entries organized around Czech as pivot language
- suggesting possible translations to foreign languages
- non-parallel corpora
- equivalent terms share similar contexts
- evaluation: correct term translation appears in top 20 suggestions
    - English 34%, German 40%, French 21%, Russian 24%, Slovak 47%

Initial Thesaurus Data

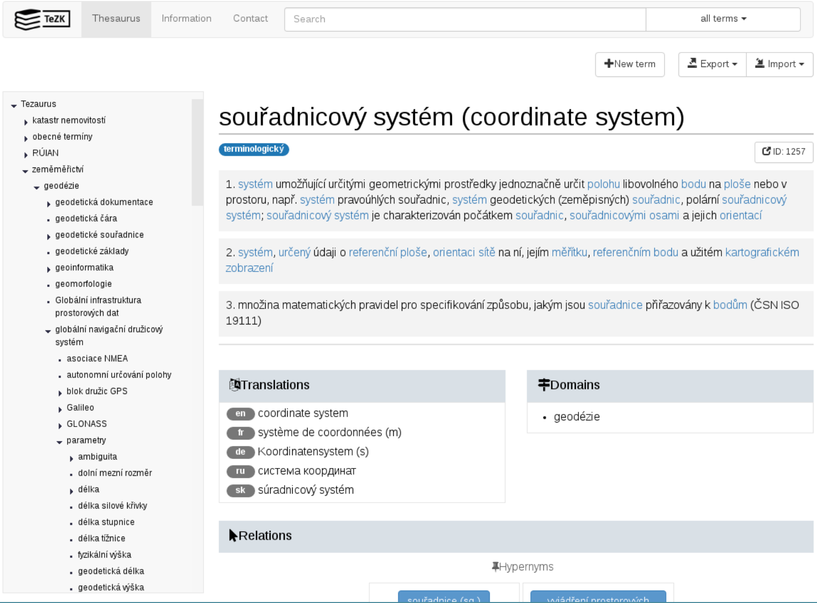
- combining existing resources
    - current Czech authoritative terminology dictionary (term definitions and translations)
    - hyper/hyponymic tree (without detailed term information)
- import module for HTML, CSV, TXT → XML
- term entries with both detailed term information and term relations

Thesaurus size statistics

	Number of entries
total entries	8,427
hyper/hyponymic relations	8,827
explanations provided	4,117
entries categorized to domain	3,905
total number of translations	24,973
English translation	9,073
German translation	4,513
Slovak translation	3,751
Russian translation	3,068
French translation	4,568

Conclusions, Future Work

- project is completed, to be installed at CUZK
- future research, investigate the techniques for candidate translations identification



- Tezaurus
  - katastr nemovitostí
  - obecné termíny
  - RÚIAN
  - zeměměřictví
    - geodézie
      - geodetická dokumentace
      - geodetická čára
      - geodetické souřadnice
      - geodetické základy
      - geoinformatika
      - geomorfologie
      - Globální infrastruktura prostorových dat
      - globální navigační družicový systém
        - asociace NMEA
        - autonomní určování polohy
        - blok družic GPS
        - Galileo
        - GLONASS
        - parametry
          - ambiguita
          - dolní mezní rozměr
          - délka
          - délka silové křivky
          - délka stupnice
          - délka tížnice
          - fyzikální výška
          - geodetická délka
          - geodetická výška

souřadnicový systém (coordinate system)

terminologický

ID: 1257

1. systém umožňující určitými geometrickými prostředky jednoznačně určit polohu libovolného bodu na ploše nebo v prostoru, např. systém pravoúhlých souřadnic, systém geodetických (zeměpisných) souřadnic, polární souřadnicový systém; souřadnicový systém je charakterizován počátkem souřadnic, souřadnicovými osami a jejich orientací

2. systém, určený údaji o referenční ploše, orientaci sítě na ní, jejím měřítku, referenčním bodu a užitém kartografickém zobrazení

3. množina matematických pravidel pro specifikování způsobu, jakým jsou souřadnice přiřazovány k bodům (ČSN ISO 19111)

Translations

en	coordinate system
fr	système de coordonnées (m)
de	Koordinatensystem (s)
ru	система координат
sk	súradnicový systém

Domains

- geodézie

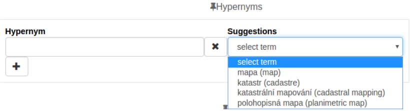
Relations

Hypernyms

souřadnice (sg.)

vyjádření prostorových

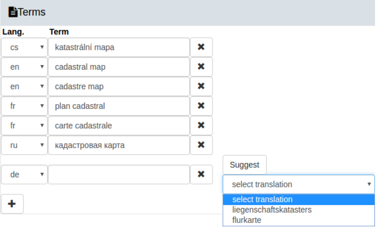
broader term candidates and translation candidates offered within the editor



Hypernyms

Hypernym	Suggestions
✖	select term ▾
✚	

select term
mapa (map)
katastr (cadastre)
katastrální mapování (cadastral mapping)
polohopisná mapa (planimetric map)



Terms

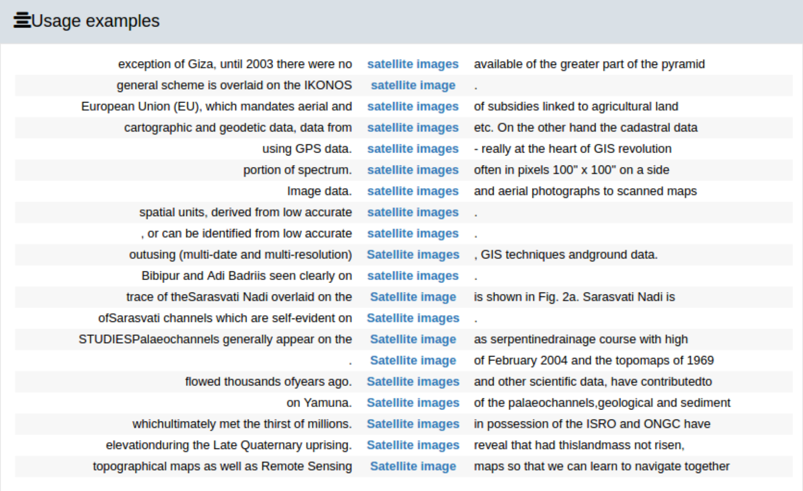
Lang.	Term	
cs ▾	katastrální mapa	✖
en ▾	cadastral map	✖
en ▾	cadastre map	✖
fr ▾	plan cadastral	✖
fr ▾	carte cadastrale	✖
ru ▾	кадастровая карта	✖
de ▾		✖

✚

Suggest

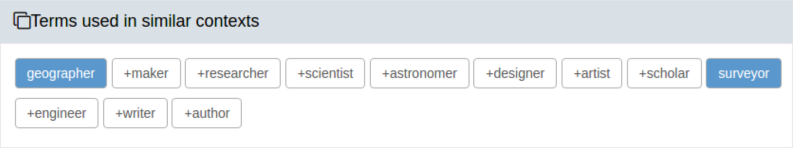
select translation ▾

select translation
liegenschaftskatasters
flurkarte



Usage examples

exception of Giza, until 2003 there were no	satellite images	available of the greater part of the pyramid
general scheme is overlaid on the IKONOS	satellite image	.
European Union (EU), which mandates aerial and	satellite images	of subsidies linked to agricultural land
cartographic and geodetic data, data from	satellite images	etc. On the other hand the cadastral data
using GPS data.	satellite images	- really at the heart of GIS revolution
portion of spectrum.	satellite images	often in pixels 100" x 100" on a side
Image data.	satellite images	and aerial photographs to scanned maps
spatial units, derived from low accurate	satellite images	.
, or can be identified from low accurate	satellite images	.
outusing (multi-date and multi-resolution)	Satellite images	, GIS techniques andground data.
Bibipur and Adi Badriis seen clearly on	satellite images	.
trace of theSarasvati Nadi overlaid on the	Satellite image	is shown in Fig. 2a. Sarasvati Nadi is
ofSarasvati channels which are self-evident on	Satellite images	.
STUDIESPalaeochannels generally appear on the	Satellite image	as serpentinedrainage course with high
.	Satellite image	of February 2004 and the topomaps of 1969
flowed thousands ofyears ago.	Satellite images	and other scientific data, have contributedto
on Yamuna.	Satellite images	of the palaeochannels,geological and sediment
whichultimately met the thirst of millions.	Satellite images	in possession of the ISRO and ONGC have
elevationduring the Late Quaternary uprising.	Satellite images	reveal that had thislandmass not risen,
topographical maps as well as Remote Sensing	Satellite image	maps so that we can learn to navigate together



Terms used in similar contexts