Semiautomatic Building and Extension of Terminological Thesaurus for Land Surveying Domain

Adam Rambousek, Aleš Horák, Vít Suchomel,
Vít Baisa, Lucia Kocincová

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

Introduction

Surveying Uncharted Lands

- project for Czech Office for Surveying, Mapping and Cadastre (CUZK)
- integrating terminological dictionary with specialized domain corpus
  - corpus building
  - automatic term extraction
  - terminological thesaurus editor
- methodology reusable for other domains

Corpus Building

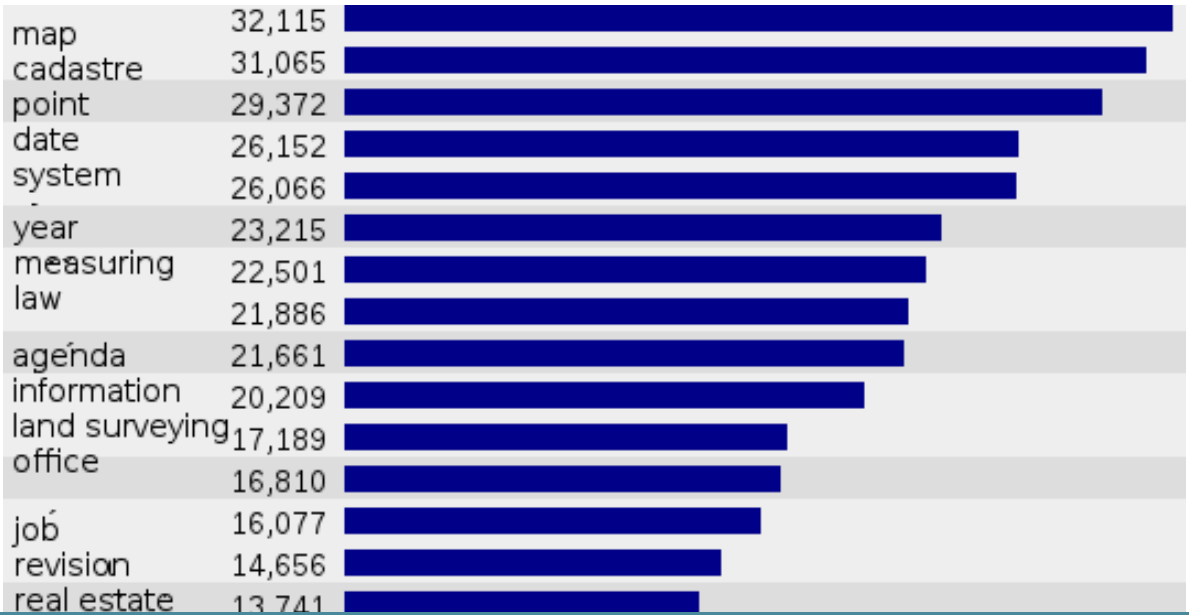
Claiming the Data

- first step – download content from public websites in the domain
  - `cuzk.cz, vugtk.cz, zemeric.cz,...`
  - over 25,000 unique documents, over 7,500,000 unique tokens
- second step – download more websites based on *root content*
- remove non-textual and low quality content (Justext)
- remove duplicate documents (Onion)

Term Extraction

Field Survey

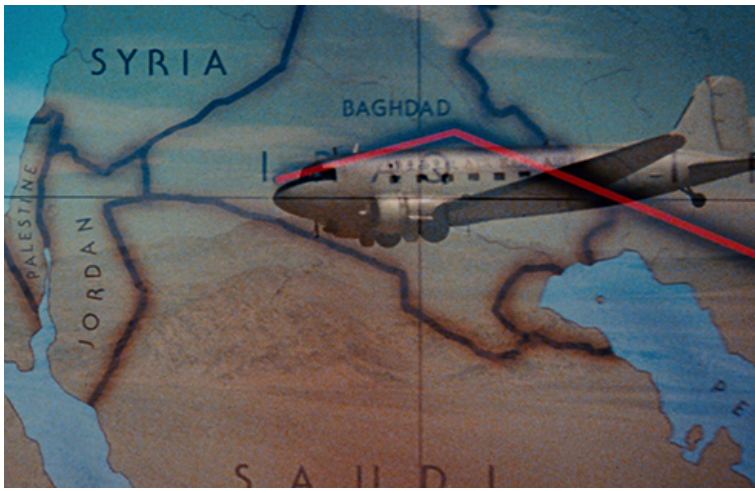
- detect *candidate terms* in the corpus as proposal to include in thesaurus
- comparing frequency of words and named-entities from specialized corpus with the biggest NLPC Czech corpus (czTenTen12)

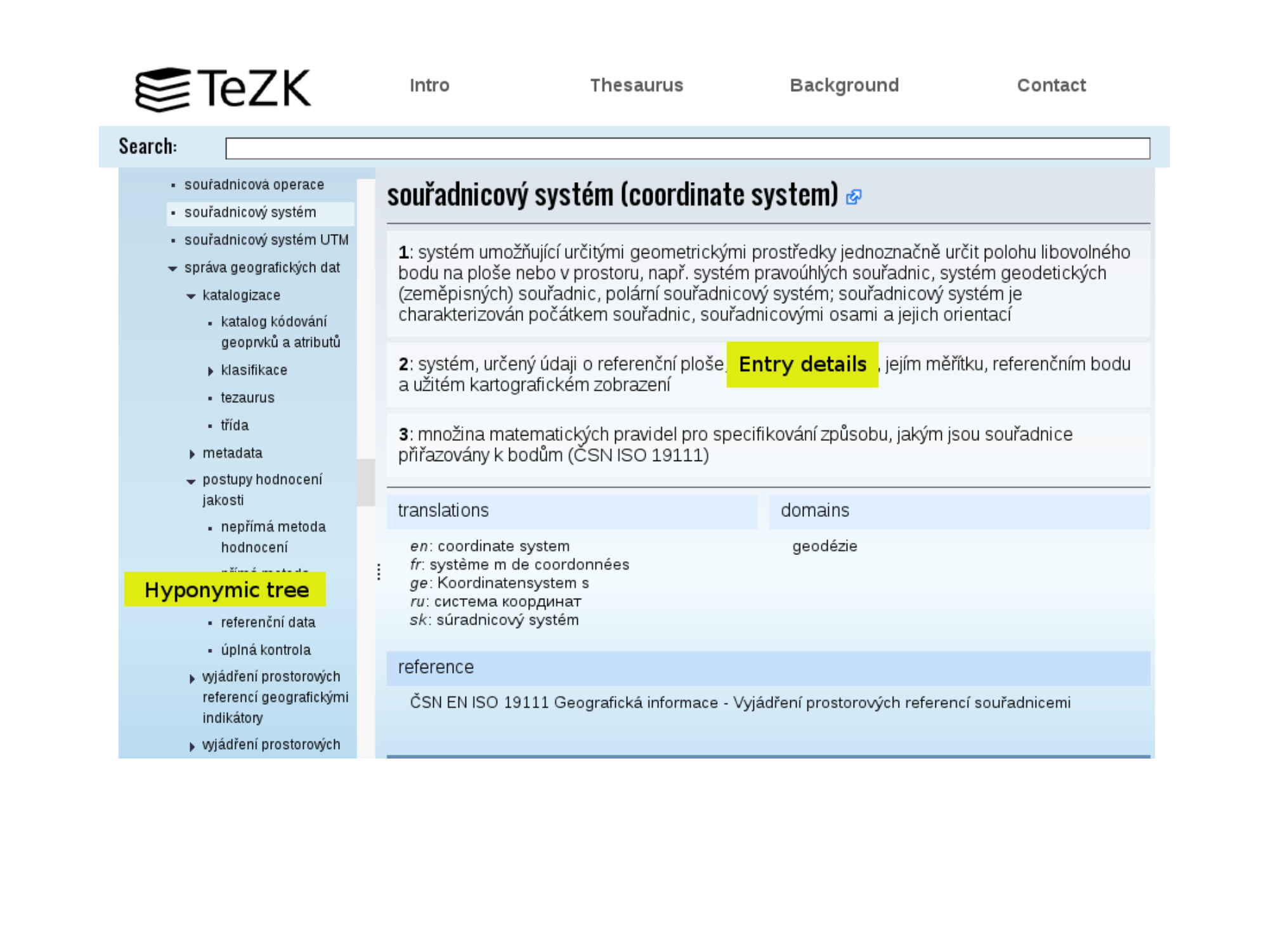


Thesaurus Editor

Observing the Dictionary

- based on DEB II platform
- client-server
- server – providing API for integration in third-party applications
- client – web application
- corpus integration (eg. metadata for `geoportal.cuzk.cz`)

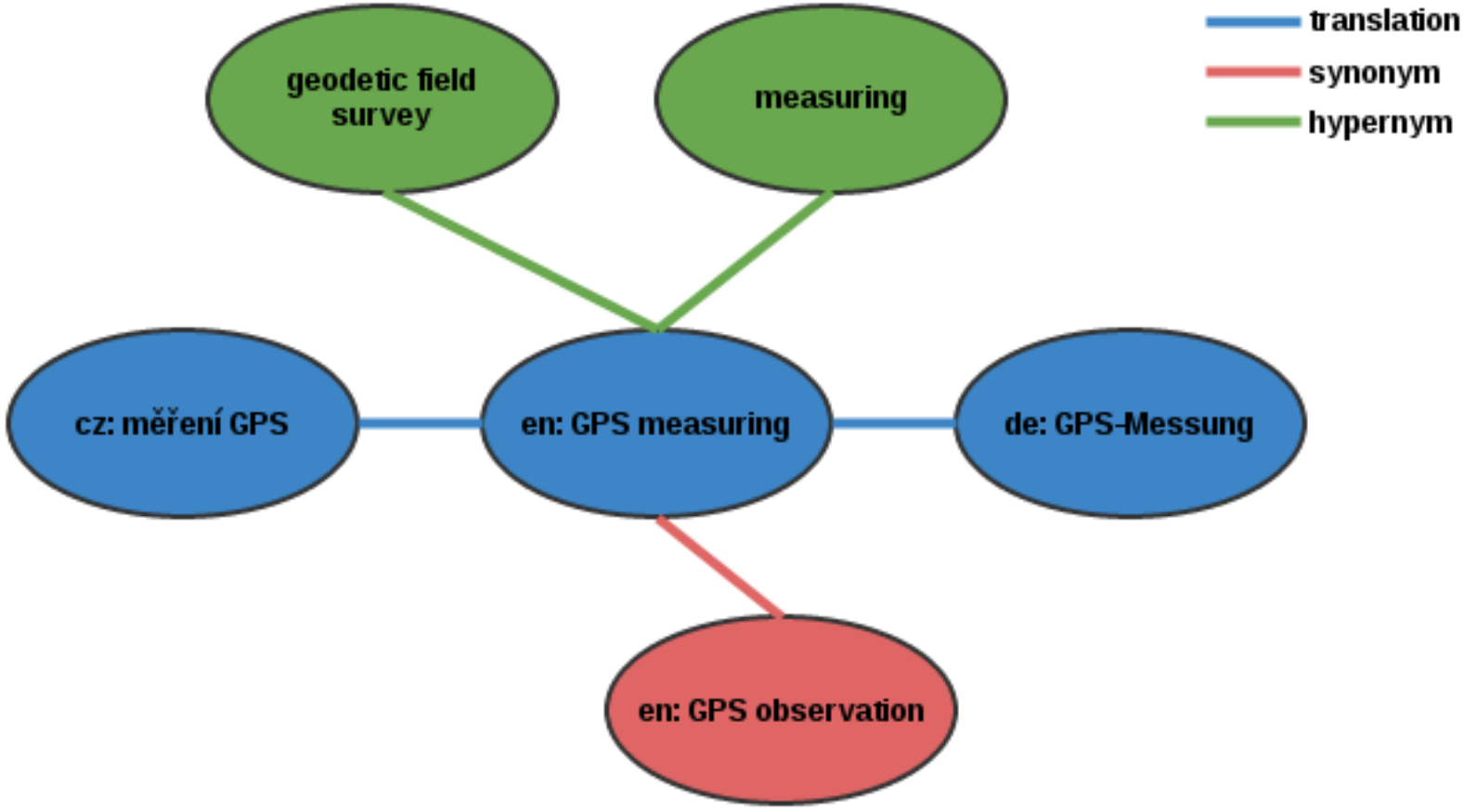



TeZK
Intro
Thesaurus
Background
Contact
Search:
souřadnicová operace
souřadnicový systém
souřadnicový systém UTM
správa geografických dat
katalogizace
katalog kódování geoprvků a atributů
klasifikace
tezaurus
třída
metadata
postupy hodnocení jakosti
nepřímá metoda hodnocení
Hyponymic tree
referenční data
úplná kontrola
vyjádření prostorových referencí geografickými indikátory
vyjádření prostorových
souřadnicový systém (coordinate system)
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
Entry details
, 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 m de coordonnées
ge: Koordinatensystem s
ru: система координат
sk: súradnicový systém
domains
geodézie
reference
ČSN EN ISO 19111 Geografická informace - Vyjádření prostorových referencí souřadnicemi

Land Surveying Thesaurus

Drawing the Term Map

- first step – combining existing separate data
  - authorized terminological dictionary – almost 4,000 terms (no relations)
  - hypero/hyponymic tree – over 6,800 entries (relations, but no detailed information)
- more resources in final version
  - parts of GEMET (General Multilingual Environmental Thesaurus)
  - regularly updated Registry of Territorial Identification (RUIAN)
  - automatically extracted multi-lingual terms
  - users' suggestions


translation
synonym
hypernym
geodetic field survey
measuring
cz: měření GPS
en: GPS measuring
de: GPS-Messung
en: GPS observation

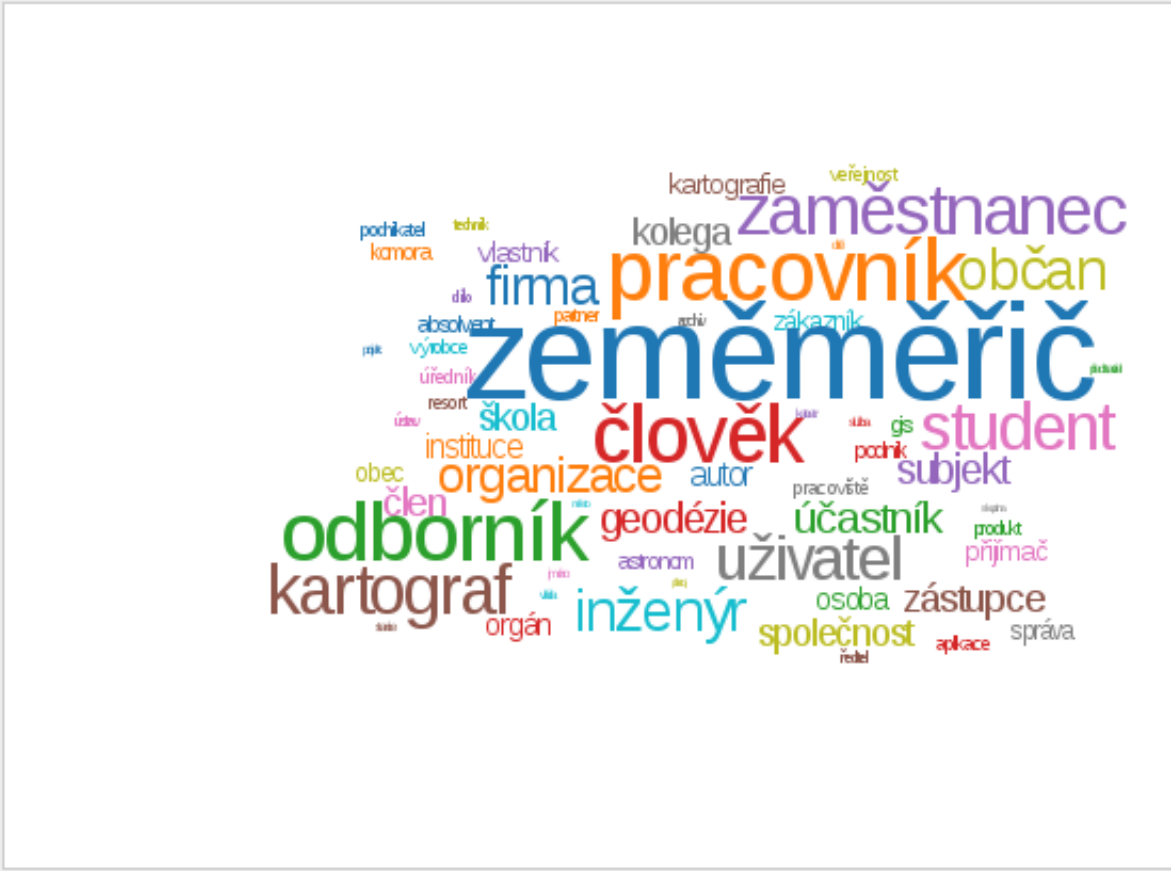
Query **teodolit** **966** (74.7 per million)

Page 1 of 49 Go Next | Last

csgk.fce.vutbr.cz	Trimble, Nikon, TS a **teodolity** pro stavebnictví
czechmaps.cz	Autor těchto řádků si s lehkou nostalgií uvědomil, že se všemi vystavenými historickými počítadly a **teodolity** kdysi pracoval.
vugtk.cz	Družicové komory a proměřování snímků První pozorování sovětských družic 2. (1957 β) a 3. (1958 δ1, 1958 δ2) byla na Pecném konána vizuálně, pomocí širokoúhlého hledáčku namontovaného na **teodolitu** Wild T2, a veškerá „elektronika" spočívala v záznamu času na chronografu Favag [7].
vugtk.cz	My jsme při měření k redukci světelného signálu používali síťkové clony, o celkem 4 velikostech, které se nasazovaly buď přímo na objímku dalekohledu **teodolitu** nebo je držel pomocník před objektivem.
vugtk.cz	A Ing. Weber, Foto 6: Vynášení spodní části astronomického universálu WT4 z tábora astronomické nejzdatnější „nosič" naší skupiny na Fatra Kriváň (1671 m) na krosně od **teodolitu** WT3.
vugtk.cz	Podstatné zlepšení přesnosti přineslo jejich pozorování pomocí speciálního lomeného hledáčku, který opatřil Ing. Rükl a který se připevnil k dalekohledu **teodolitu** WT3.
vugtk.cz	Ideální by však bylo podle získaných zkušeností zcela předělat spodní část přístroje se stavěcími šrouby umístěnými zcela mimo obvod této části přístroje, tak jak je tomu např. u **teodolitů** WT3.
vugtk.cz	Tak jako se asi nikdo nenechá operovat televizním divákem seriálu Nemocnice na kraji města, který si před rokem koupil skalpel, tak proč si nechá vytyčit svůj pozemek osobou, která má **teodolit** a stativ.
vugtk.cz	Tak jako se asi nikdo nenechá operovat televizním divákem seriálu Nemocnice na kraji města, který si před zákrokem koupil skalpel, tak proč si lidé nechávají vytyčit svůj pozemek osobou, která má doma **teodolit** a stativ.
vugtk.cz	Nabízí se třeba humor za **teodolitem**

geodet TECU – Zeměměřický webový korpus freq = 4,360 (337.2 per million)

Lemma	Score	Freq
zeměměřič	0.317	4,934
pracovník	0.219	3,343
odborník	0.205	1,455
člověk	0.204	3,468
zaměstnanec	0.189	2,461
kartograf	0.188	1,646
student	0.179	1,766
uživatel	0.171	6,434
občan	0.17	1,490
inženýr	0.17	1,114
firma	0.169	6,987
organizace	0.152	5,234
účastník	0.142	4,314
geodézie	0.138	5,496
subjekt	0.137	1,304
člen	0.137	5,377
zástupce	0.133	2,521
společnost	0.132	9,072
kolega	0.131	853
škola	0.131	3,975
instituce	0.13	1,767
autor	0.128	3,139
osoba	0.127	10,493
orgán	0.125	6,020
přijímač	0.123	2,866
vlastník	0.122	5,630
kartografie	0.119	5,627
správa	0.118	21,661
obec	0.117	7,404
zákazník	0.117	1,408



Future Work

New Expeditions

- build corpora in more languages – English, French, German
- automatically extracted terminology – propose translations
- integrating editor in other applications
- periodic semi-automatic imports from authorized sources
- support for Terminological Committee
- filter terms based on reliability

Thank you for your attention!

More information:
http://deb.fi.muni.cz
deb@fi.muni.cz