Semiautomatic Building and Extension of Terminological Thesaurus for Land Surveying Domain

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

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xrambous,hales,xsuchom2,lucyia}@fi.muni.cz

Abstract. This paper describes the methodology and development of tools for building and presenting a terminological thesaurus closely connected with a new specialized domain corpus. The thesaurus multiplatform application offers detailed information on each term, visualizes term relations, or displays real-life usage examples of the term in the domain-related documents. Moreover, the specialized corpus is used to detect domain specific terms and propose an extension of the thesaurus with new terms. The presented project is aimed at the terminological thesaurus of land surveying domain, however the tools are re-usable for other terminological domains.

Keywords: corpus building, thesaurus, terminological dictionary, term extraction, DEB platform

1 Introduction

Specialists in every field of work use their own domain-specific vocabulary and it is desirable to share the same terminology amongst the professionals. Detailed domain terminology is not usually included in general language dictionaries, thus specialized terminological dictionaries are needed. With the need to share information unambiguously in different languages, terminological dictionaries link original terms to their translations. The taxonomical ordering of the terminology is described by term relations such as synonymy or hyperonymy and hyponymy. The information is presented and visualized in a way that helps the readers (both specialists and general public) to understand the term meaning and usage in contexts. If the data are encoded properly, the system enables automatic processing and integration of the data in third-party applications.

Natural language is still evolving and new words keep appearing or the usage and meaning of words is changing. This evolution is even more noticeable in specialized vocabularies [1]. The thesaurus system thus can employ sophisticated methods of detecting emerging words and distinct new terms in the given domain by processing synchronous domain-oriented corpora.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 129–137, 2014. © NLP Consulting 2014

The Natural Language Processing Centre (NLP Centre) at the Faculty of Informatics, Masaryk University in cooperation with the Czech Office for Surveying, Mapping and Cadastre is developing a system for building and extensions of specialized terminological thesaurus for the domain of land surveying and land cadastre. The project consists of two interconnected parts – an application to create, edit, browse and visualize the terminological thesaurus, and the tools to build large corpus of domain oriented documents with the possibility to detect newly emerging terms, or terms missing from the thesaurus. Already available tools for corpus building and term extraction and the platform for dictionary applications are utilized. During the project, we are enhancing the corpus tools (mainly to support parallel multilingual corpora), building the thesaurus web application (not limited to single domain), and developing methods to inter-connect the domain corpus with the terminological thesaurus.

The project is currently in its first phase. We have built the Czech corpus of land surveying oriented documents and we are able to detect domain specific terms. We have also developed the multiplatform web-based editor and browser thesaurus application based on the dictionary writing platform DEB. Although this project aims to build and manage the terminological thesaurus of land surveying domain, the tools may be re-used for any other domain dictionary, thus stimulating the sharing of information and general awareness of the selected domain.

2 Specialized Corpus and Term Extraction

To build the specialized corpus for land surveying and geoinformation domain, we have followed the principles designed for creation of large corpora extracted and processed from web data. The data for the corpus were gathered from publicly available online resources utilizing two different methods developed by NLP Centre.

Firstly, a set of main websites related to the land surveying, the cadastre of real estates, and related topics was enlisted. See Table 1 for details regarding the sources.

Secondly, based on the content of these *root websites* a broader set of documents from 1,063 websites utilizing the WebBootCat tool [3] was obtained.

Table 1: Website resources for the specialized corpus

Website	Documents	Tokens	Unique documents	Unique tokens
www.cuzk.cz	16,405	3,137,795	15,289	340,943
www.vugtk.cz	4,659	6,419,950	3,212	4,386,238
csgk.fce.vutbr.cz	241	77,255	198	58,561
www.kgk.cz	417	44,814	414	29,890
www.sfdp.cz	192	35,287	106	11,279
www.czechmaps.cz	94	108,506	90	98,914
www.zememeric.cz	8,634	6,100,751	6,200	2,638,308

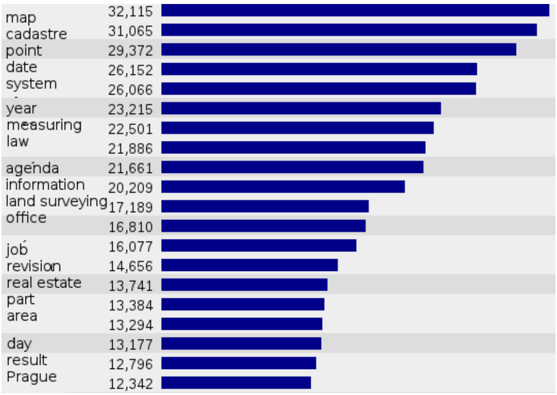


Fig. 1: Most frequent nouns in the land surveying corpora.

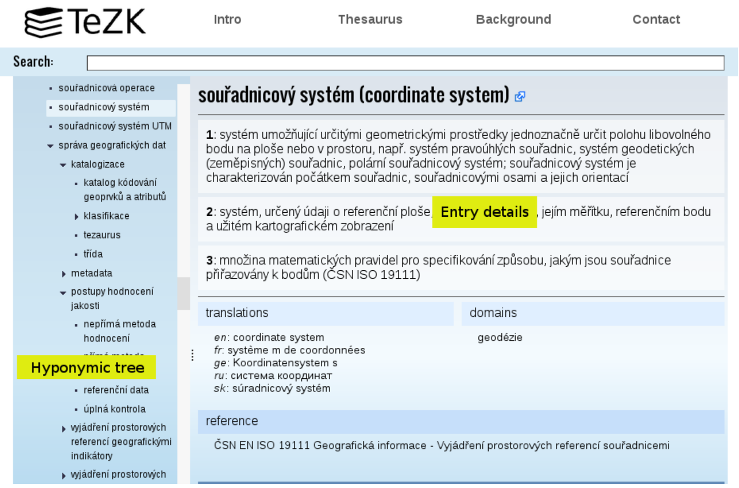
This method needs a set of "seed words" to search the web for relevant documents. We used the main domain terms obtained from existing publicly available terminological dictionary [12] as seed words. The resulting corpus is used for extraction of new suggested terms for inclusion in the thesaurus. See Table 2 for detailed information on downloaded documents and their distribution amongst different sub-domains (as divided in the available terminilogical dictionary) covered by the thesaurus.

Non-textual and low quality content was removed from the downloaded documents, utilizing the Justext tool [2]. Subsequently, duplicate documents or parts (e.g. paragraphs) of the documents were purged with the Onion tool [2].

Following the corpus creation, a list of "candidate terms" (proposals to include into the thesaurus) was prepared. The candidate terms were extracted from the specialized land surveying and geoinformation corpus by employing the process of corpora comparing and keywords extraction [4,5]. Frequencies

Table 2: WebBootCat resources for the specialized corpus

Domain	Documents	Tokens	Unique documents	Unique tokens
GPS system	118	250,833	117	221,315
metrology	144	867,156	144	619,482
photogrammetry	42	244,212	42	227,731
geographical information	55	805,059	55	550,681
mapping	213	858,575	212	722,080
cartography	368	1,358,973	365	1,124,708
cadastre of real estates	260	970,951	259	776,497
geodesy	190	575,381	189	483,679
theory of errors	75	258,345	75	218,809
instrumental technology	115	187,106	113	173,984
engineering surveying	114	286,846	113	242,857



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 , 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	domains
en: coordinate system *fr*: système m de coordonnées *ge*: Koordinatensystem s *ru*: система координат *sk*: súradnicový systém	geodézie

reference

ČSN EN ISO 19111 Geografická informace - Vyjádření prostorových referencí souřadnicemi

Fig. 2: Browsing the thesaurus, with detailed information for one term.

of words and named-entities in the specialized corpora are compared to the frequencies of the same phrases in a general language corpus (in this case, the biggest Czech corpus developed in NLPC – czTenTen12 [6]). The best candidate terms have the highest frequency quotient [7].

3 DEB Platform

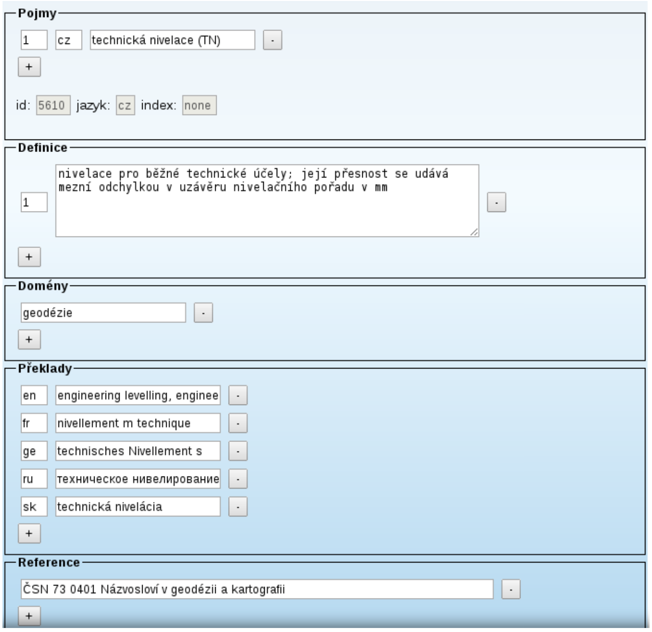
Utilizing the experience from several lexicographic projects, we have designed and implemented universal dictionary writing system that can be exploited in various lexicographic applications to build large lexical databases. The system is called Dictionary Editor and Browser, or DEB [8], and has been used in many lexicographic projects, e.g. for development of the Czech Lexical Database [9], or currently running Pattern Dictionary of English Verbs [10], and Family names in UK [11].

The DEB platform is based on client-server architecture, which brings along a lot of benefits. All the data are stored on a server and considerable part of functionality is also implemented on the server, while the client application can be very lightweight.

This approach provides very good tools for editor team cooperation; data modifications are immediately seen by all the users. Server also provides authentication and authorization tools.

4 Thesaurus Building

Although the main aim of the thesaurus development is publishing the authorized specialized terminology and its updates both to the experts, and general



Pojmy

1 cz technická nivelace (TN)

id: 5610 jazyk: cz index: none

Definice

1 nivelace pro běžné technické účely; její přesnost se udává mezní odchylkou v uzávěru nivelačního pořadu v mm

Domény

geodézie

Překlady

en	engineering levelling, enginee
fr	nivellement m technique
ge	technisches Nivellement s
ru	техническое нивелирование
sk	technická nivelácia

Reference

ČSN 73 0401 Názvosloví v geodézii a kartografii

Fig. 3: Editing the term entry.

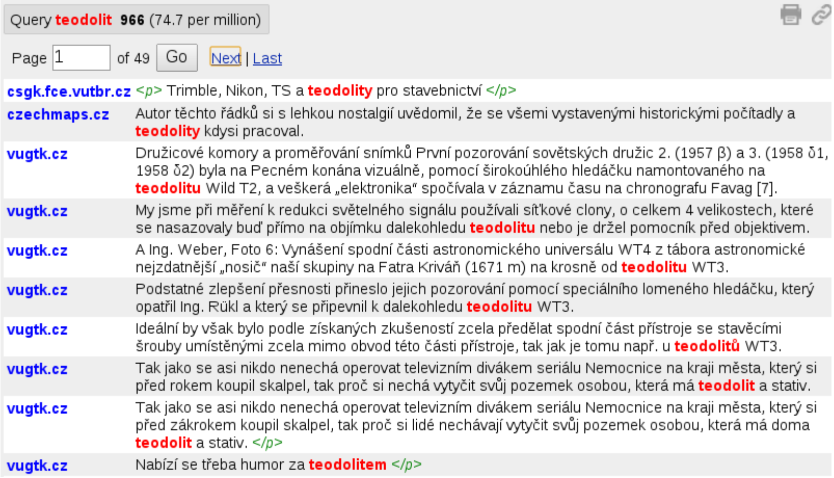
public, the thesaurus will contain broad vocabulary of related terms. Users may search even for unofficial terms and thanks to the relations between the terms and the detailed information on the source of given term, user will find the related terms and links to the recommended official variant.

To build the thesaurus covering broad domain vocabulary, several resources are combined. In the first stage, the current authorized terminological dictionary [12] (containing almost 4,000 terms' definitions and translations, but does not offer the taxonomy network) was combined with the hypero/hyponymic tree of over 6,800 entries (containing hyponymic relations, but no detailed information about terms) and by 450 candidate terms extracted from the domain corpus.

The first two resources were available in HTML form, tagging some parts of entry structure, but still leaving a lot of text in unstructured format. It was necessary to tidy up the data and convert resources to the unified XML format

Table 3: Thesaurus size statistics

total number of terms	8,783	English translations	8,873
hyponymic relations	10,020	German translations	3,936
meaning explanations	4,124	Slovak translations	3,511
		Russian translations	2,762
		French translations	3,936



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

Page 1 of 49 Go Next | Last

csgk.fce.vutbr.cz	<p> Trimble, Nikon, TS a **teodolity** pro stavebnictví </p>
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. </p>
vugtk.cz	Nabízí se třeba humor za **teodolitem** </p>

Fig. 4: Corpus evidence for usage of the selected term (*teodolit*).

for the database storage. Some of the terms were shared by both dictionaries, thus combined term entries were created, containing both detailed information on terms, and the term relations. See Table 3 for more details regarding the current size of the thesaurus.

In the next stage, the thesaurus will be expanded even more by including several resources:

- appropriate parts of the GEMET[1] (General Multilingual Environmental Thesaurus),
- regularly updated Registry of Territorial Identification (RUIAN)[2],
- automatically extracted multi-lingual terms,
- suggestions from the public users.

5 Editing Tool

The thesaurus editing tool is implemented as a client-server application, with DEB server providing the database and management back-end. The client-side application is a multiplatform web application accessible in any modern browser, built utilizing open-source technologies – JQuery[3] and SAPUI5[4] libraries for graphical interface. The client and the server communicate using standardized interface over HTTP, currently JSON format is supported and support for SOAP web-service protocol will be added in the final version.

[1] http://www.eionet.europa.eu/gemet

[2] http://www.cuzk.cz/ruian/

[3] http://jquery.com/

[4] https://sapui5.netweaver.ondemand.com/

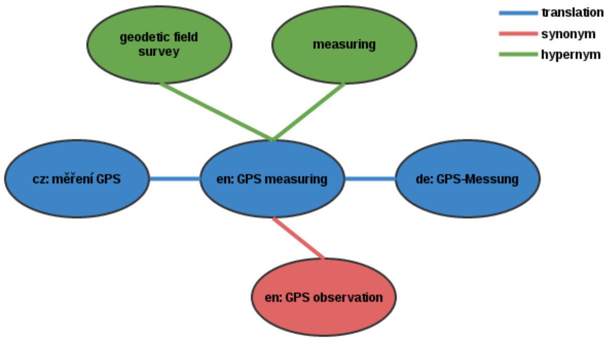


Fig. 5: Entry relations visualized.

The standardized application interface also allows an integration of third-party applications that would like to re-use the thesaurus data. One of the intended use-cases is the integration into the Geoportal[5], where the terms are to be used for the document metadata and categorization.

The thesaurus web application itself provides a graphical interface for browsing the hyponymic tree (see Figure 2). Out of the several possible visualizations of the tree, the expanding multi-level tree was selected, although it may not display all the relations in a proper graph form, it is much more intuitive for the users. If the term has more hyponyms, it is displayed multiple times in the tree structure. To graphically visualize the relations of a term, a graph of hypernyms, synonyms, and other related terms is displayed (see Figure 5).

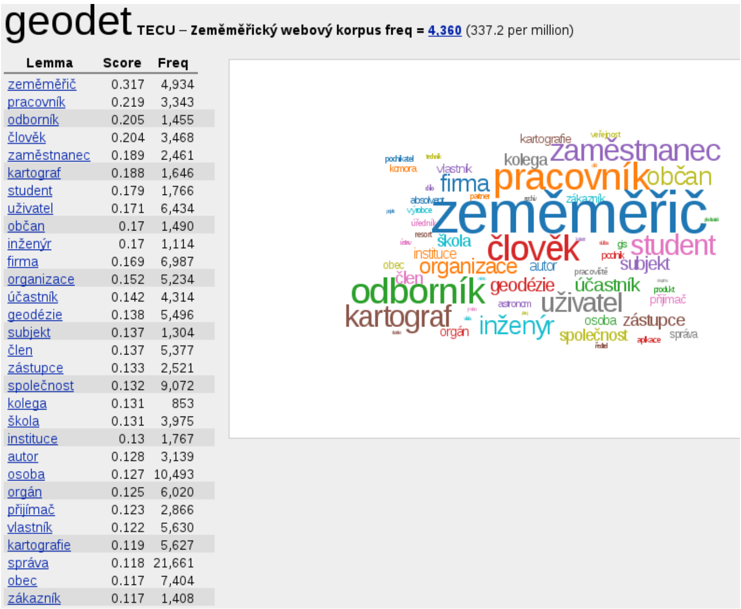
For each term, a detailed description is given, including meaning explanation, translations, or accepted variants. When more sources are incorporated in the thesaurus, the reliability of each source and revision history will be presented to the users. Source reliability follows the rating scale of the Office for Surveying – the most reliable are terms authorized by the terminological committee, followed by terms used in scientific journals, with the terms made up by general public at the bottom of the scale. Users or third-party applications may decide which sources or terms they prefer to work with.

To get a better picture of the term and its usage, extended information from the corpus are presented. Users may consult full examples (see Figure 4) or related words from the corpus (see Figure 6).

6 Conclusions and Future Work

In the next phase of the project, we plan to extend multi-lingual and multi-source aspects of the thesaurus.

[5] `http://geoportal.cuzk.cz/`



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

Fig. 6: Related words for the selected term (*geodet*).

Based on the successful evaluation of the Czech specialized corpus for the land surveying domain, we will build corpora in more languages – English, French, German, with the possibility of other languages, depending on the availability of source documents. We will provide automatically extracted terminology from these corpora as the suggestions for terminology translation.

Hand in hand with adding more sources for the thesaurus terms, the editing and browsing application will offer options for filtering the terms based on the source reliability and authorization status and periodic semi-automatic imports from authorized sources.

Acknowledgements This work has been partly supported by TACR in the project TB02CUZK004.

References

1. Fischer, R.: Lexical change in present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms. Volume 17. Gunter Narr Verlag (1998)
2. Pomikálek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Phd thesis, Masarykov university, Faculty of Informatics (2011)
3. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P.: Webbootcat: instant domain-specific corpora to support human translators. In: Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation, Oslo,

The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway) (2006) 247–252

4. Kilgarriff, A.: Comparing corpora. International journal of corpus linguistics **6**(1) (2001) 97–133
5. Kilgarriff, A.: Simple maths for keywords. In: Proc. Corpus Linguistics. (2009)
6. Suchomel, V.: Recent czech web corpora. In Aleš Horák, P.R., ed.: 6th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2012) 77–83
7. Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V.: Finding terms in corpora for many languages with the sketch engine. In: Proceedings of the Demonstrations at the 14th Conferencethe European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, The Association for Computational Linguistics (2014) 53–56
8. Horák, A., Rambousek, A.: DEB Platform Deployment – Current Applications. In: RASLAN 2007: Recent Advances in Slavonic Natural Language Processing, Brno, Czech Republic, Masaryk University (2007) 3–11
9. Horák, A., Rambousek, A.: PRALED – A New Kind of Lexicographic Workstation. In: Computational Linguistics. Springer (2013) 131–141
10. Hanks, P.: Corpus pattern analysis. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud (2004)
11. Hanks, P., Cullen, P., Draper, S., Coates, R.: Family names of the United Kingdom. (2014)
12. Hánek, P.: Terminologický slovník zeměměřictví a katastru nemovitostí. Výzkumný ústav geodetický, topografický a kartografický, v.v.i. (2012)