

RASLAN 2017
Recent Advances in Slavonic
Natural Language Processing

A. Horák, P. Rychlý, A. Rambousek (Eds.)

RASLAN 2017

**Recent Advances in Slavonic Natural
Language Processing**

**Eleventh Workshop on Recent Advances
in Slavonic Natural Language Processing,
RASLAN 2017**

**Karlova Studánka, Czech Republic,
December 1–3, 2017
Proceedings**

**Tribun EU
2017**

Proceedings Editors

Aleš Horák
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: hales@fi.muni.cz

Pavel Rychlý
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: pary@fi.muni.cz

Adam Rambousek
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: rambousek@fi.muni.cz

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the Czech Copyright Law, in its current version, and permission for use must always be obtained from Tribun EU. Violations are liable for prosecution under the Czech Copyright Law.

Editors © Aleš Horák, 2017; Pavel Rychlý, 2017; Adam Rambousek, 2017

Typography © Adam Rambousek, 2017

Cover © Petr Sojka, 2010

This edition © Tribun EU, Brno, 2017

ISBN 978-80-263-1340-3

ISSN 2336-4289

Preface

This volume contains the Proceedings of the Eleventh Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2017) held on December 1st–3rd 2017 in Karlova Studánka, Sporthotel Kurzovní, Jeseníky, Czech Republic.

The RASLAN Workshop is an event dedicated to the exchange of information between research teams working on the projects of computer processing of Slavonic languages and related areas going on in the NLP Centre at the Faculty of Informatics, Masaryk University, Brno. RASLAN is focused on theoretical as well as technical aspects of the project work, on presentations of verified methods together with descriptions of development trends. The workshop also serves as a place for discussion about new ideas. The intention is to have it as a forum for presentation and discussion of the latest developments in the field of language engineering, especially for undergraduates and postgraduates affiliated to the NLP Centre at FI MU.

Topics of the Workshop cover a wide range of subfields from the area of artificial intelligence and natural language processing including (but not limited to):

- * text corpora and tagging
- * syntactic parsing
- * sense disambiguation
- * machine translation, computer lexicography
- * semantic networks and ontologies
- * semantic web
- * knowledge representation
- * logical analysis of natural language
- * applied systems and software for NLP

RASLAN 2017 offers a rich program of presentations, short talks, technical papers and mainly discussions. A total of 13 papers were accepted, contributed altogether by 20 authors. Our thanks go to the Program Committee members and we would also like to express our appreciation to all the members of the Organizing Committee for their tireless efforts in organizing the Workshop and ensuring its smooth running. In particular, we would like to mention the work of Aleš Horák, Pavel Rychlý and Marie Stará. The \TeX expertise of Adam Rambousek (based on \LaTeX macros prepared by Petr Sojka) resulted in the extremely speedy and efficient production of the volume which you are now holding in your hands. Last but not least, the cooperation of Tribun EU as a printer of these proceedings is gratefully acknowledged.

Brno, December 2017

Karel Pala

Table of Contents

I Electronic Lexicography and Language Resources

New features in DEBVisDic for WordNet Visualization and User Feedback	3
<i>Adam Rambousek, Aleš Horák, David Klement, and Jiří Kletečka</i>	
Preliminary Thoughts on Issues of Modeling Japanese Dictionaries Using the OntoLex Model	11
<i>Louis Lecailliez</i>	
Wikilink – Wikipedia Link Suggestion System, Its Problems and Possible Solutions	21
<i>Vojtěch Mrkývka</i>	

II Semantics and Language Modelling

The Ordered-triple Theory of Language: Its History and the Current Context	29
<i>Aleš Horák and Karel Pala</i>	
Property Modifiers	37
<i>Marie Duží and Michal Fait</i>	
Multilinguality Adaptations of Natural Language Logical Analyzer	51
<i>Marek Medved', Terézia Šulganová, and Aleš Horák</i>	
Idiomatic Expressions in VerbaLex	59
<i>Zuzana Nevěřilová</i>	

III NLP Applications

Recognition of Invoices from Scanned Documents	71
<i>Hien Thi Ha</i>	
Enlargement of the Czech Question-Answering Dataset to SQAD v2.0	79
<i>Terézia Šulganová, Marek Medved', and Aleš Horák</i>	
Semantic Similarities between Locations based on Ontology	85
<i>Moiz Khan Sherwani, Petr Sojka, and Francesco Calimeri</i>	

IV Text Corpora

Language Code Switching in Web Corpora	97
<i>Vladimír Benko</i>	
KernelTagger – a PoS Tagger for Very Small Amount of Training Data	107
<i>Pavel Rychlý</i>	
Manipulative Propaganda Techniques	111
<i>Vít Baisa, Ondřej Herman, and Aleš Horák</i>	
Subject Index	119
Author Index	121

Part I

**Electronic Lexicography and
Language Resources**

New features in DEBVisDic for WordNet Visualization and User Feedback

Adam Rambousek, Aleš Horák, David Klement, Jiří Kletečka

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
deb@aurora.fi.muni.cz
<http://deb.fi.muni.cz>

Abstract. This paper presents two new features that help with wordnet management and presentation in DEBVisDic. The first is the new interface to gather user feedback about errors in wordnet and streamlined management of revisions approval and possible updates to the wordnet database. The second feature is the new visualization interface, providing both textual and graphical representation of wordnet data, with emphasis on user-friendly and responsive design. New visualization interface will be included in the DEBVisDic editor and also published as a stand-alone web application.

Key words: DEB platform, crowdsourcing, wordnet, visualization, error checking

1 Introduction

Projects to build a large ontology or semantic network usually do not include capacities needed for long-term management and updates. Although there are various automatic ontology consistency checkers [1,2,3], manual reviews are always the most reliable method for database fixes and updates. However the process is time consuming and cannot be completely finished by a small group of linguists. General audience often discovers mistakes in the published version, it is thus very useful to take user feedback into consideration.

In the following text, we present a new interface for error checking and reporting in wordnets. The interface is developed with the idea of crowdsourcing by wide public which should speed up the process of errors discovery and correction. The tool is developed within the DEB (Dictionary Editor and Browser [4,5]) framework and connected to the backend database of DEBVisDic used for developing a number of national wordnets.

For users, simple and user-friendly method for discovering wordnet data is an important aspect. After reviewing existing visualization tools, we have decided to develop new text and graph based interface within the DEBVisDic tool, which we believe offers a best solution to combined user needs in the network exploration process.

2 Wordnet Development and Issues

Issues in wordnet may be divided to two main categories:

- surface errors – issues with synset description, e.g. spelling errors in literals or definitions,
- structural errors – issues with semantic relations, appropriate literal selection, varying subtrees depth and granularity, or orphaned synsets.

Two general methodologies defined during the EuroWordNet project [6] are usually used to build new wordnets:

- *Expand model* – with this approach, Princeton WordNet (or its part) is translated to a new language, keeping the semantic relations mostly intact. Some projects translated the synsets semi-automatically, which may introduce surface errors if the results are not verified properly.
- *Merge model* – new wordnet is created either from scratch, or based on existing dictionary, which does not contain semantic relations and entries are not grouped to synsets. Wordnets utilizing this method contain more structural errors.

Many of the errors may be prevented during the wordnet development phase. Important part is to design and follow detailed guidelines [7,2]. Software tools may help significantly. Wordnet editing software should check for a range of errors, from spellchecking to semantic relations completeness [8]. Some projects also use periodical heuristic testing to check recently added or updated synsets [9].

3 Crowdsourcing in Linguistics

In linguistics and NLP research, crowdsourcing is usually used to manually annotate large datasets with semantic or syntactic information [10], word sense disambiguation [11], or to evaluate the results of automatic tools [12], but may even help to detect epidemics outbreak [13].

The results of crowdsourcing experiments in NLP research were evaluated multiple times, concluding that combining annotation by several "unskilled" annotators may result in cheaper and faster annotation. Study by [14] concluded that on average 4 non-expert annotations achieve the equivalent inter-annotator agreement as a single expert. Another experiment [15] evaluated machine translation using crowdsourcing and concluded that combination of many non-expert evaluations provides equivalent quality as experts.

In the field of lexicography, Wiktionary,¹ a sister project of Wikipedia, is one of the most prominent crowdsourced resource. The goal of Wiktionary is to create a freely available "dictionary of all words in all languages" [16] edited by volunteers. Several analysis [17,18,19] found Wiktionary to be a useful linguistic resource, however, entry quality varies from well-crafted to unreliable.

¹ <http://www.wiktionary.org/>

Road, route [n]

Definition: [an open way \(generally public\) for travel or transportation](#)

Domain: [town_planning](#)

Sumo: [StationaryArtifact](#)

Sumo type: +

Usages [Add](#)

Synonyms [Add](#)

[road 1](#)

[route 2](#)

Relations [Add](#)

[ENG20-01895340-v:route:2](#)

[ENG20-01897936-v:route:1](#)

[Cancel](#) [Save](#)

Fig. 1. User feedback form to provide synset data suggestions.

4 Crowdsourcing Tool and Review Process

Czech WordNet (CzWN) was published as a part of the EuroWordNet and Balkanet projects [6,20] and since then CzWN was mostly just maintained. However, there are several versions with various amount of edits, together with version semi-automatically extended using English-Czech translation dictionary [21]. NLP Centre (the CzWN developer) is currently running a project to integrate all updates to Czech WordNet and publish new Open Czech WordNet linked to Collaborative Interlingual Index [22].

Czech WordNet was developed using the expand model, translating the English wordnet synsets. Most notable example of errors caused by this approach are the synsets containing words that are not exactly synonyms, or only rare in the Czech language, but present in the Czech WordNet because of the translation from English. For example, the English synset *cabriolet:1, cab:2* has the equivalent Czech synset *kabriolet:2, dvoukoloový jednosprężní povoz:1, koňská drožka:1* (cabriolet, two-wheeled one horse cart, horse-drawn carriage). Although the translation is correct, this sense of *cabriolet* in Czech is very archaic, in current language the only sense used in spoken language is the convertible car. Another problem is

House [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Synonym	-- home	<input checked="" type="checkbox"/> <input type="checkbox"/>
Car, auto, automobile, machine, motorcar. [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Remove"/> Synonym	motorcar --	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Edit"/> Synonym	machine -- motorcar	<input checked="" type="checkbox"/> <input type="checkbox"/>
Love, passion [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Usage	-- love is in the air	<input checked="" type="checkbox"/> <input type="checkbox"/>
Cat, true cat [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Edit"/> Domain	zoology -- catology	<input checked="" type="checkbox"/> <input type="checkbox"/>
Cat, true cat [n]	16/11/2017	<input checked="" type="checkbox"/> <input type="checkbox"/>
<input type="button" value="Add"/> Relation	-- ENG20-02352256-n.[n] paw:1	<input checked="" type="checkbox"/> <input type="checkbox"/>

Fig. 2. Administrator view of suggested changes.

the inclusion of multiword expressions in the synset, which may be justified in some cases, but these are not fixed lexical units in the Czech language.

However, during the integration we will not have enough resources and lexicographers to check all synsets and relations in the Czech WordNet. We have developed new software tool that allows any wordnet user to report issues they spot in the data. Although we are testing the tool on the Czech WordNet, it is language-independent and available for all wordnets developed using the DEBVisDic editor.

The tool is not directly integrated into DEBVisDic editor, but rather uses the DEBVisDic server API to access wordnet data. On the other hand, all available synset representation (editor, simplified browser, API call) will enable users to easily move to the error reporting application. Users are presented with the data from the synset they were browsing and may update any data value – change existing value, add a new one if some part of synset is missing, or remove an unwanted value. See Figure 1 for an example of the user feedback form. Updates are stored in a separate database as suggestions. Each value (e.g. gloss or relation) is stored as a single suggestion.

Any member of the editing team with access permissions to the given wordnet may browse all user suggestions (or filter them by reporting user, information type, or review status). The editor may approve or reject any single suggestion, or approve/reject all suggestions for any synset at once. Of course, it is also possible to approve/reject all suggestions based on the selected filter. Before deciding, the editor may compare user feedback with previously

The screenshot shows the DEBVisDic interface for the word "house". At the top left, there is a search bar with "house" entered and a dropdown menu set to "English wordnet". To the right of the search bar are icons for a menu and a user profile. The main content area is titled "house¹" and includes a definition: "a dwelling that serves as living quarters for one or more families". Below the definition are sections for "Paths to word" (entity → object, physical object → whole¹, whole thing¹, unit¹ → artifact¹, artefact¹ → structure¹, construction¹ → building¹, edifice¹), "Word properties" (n, ENG20-03413667-n), and "Semantic relations". The "Semantic relations" section is divided into three columns: "hyponym" (stash house¹, ranch house¹, solar house¹, chalet¹, safe house¹), "near_antonym" (empty), and "part" (study², porch¹, loft¹, attic¹, garret¹, library¹). On the left side, under "Found synonym rings (synsets):", there are several lists of related terms, such as "(n) house²", "(n) family¹, household¹, house¹, home¹, menage¹", "(n) house¹", "(n) sign of the zodiac¹, star sign¹, sign¹, mansion¹, house¹, planetary house¹", "(n) house¹²", "(n) theater¹, theatre¹, house¹", "(v) house², put up¹, domicile¹", and "(n) firm¹, house¹, business firm¹".

Fig. 3. DEBVisDic synset information in the text mode.

approved or rejected updates for the selected synset. See Figure 2 for preview of editor's interface.

All approved suggestions are immediately transferred to the development version of the wordnet database and presented to users. When a user's feedback is rejected by the editor, the information is kept in the database and future users trying to suggest the same update are notified about the previous refusal.

5 New DEBVisDic Visualization Interface

The goal of the new DEBVisDic wordnet interface is to facilitate visualization of wordnets in both textual and graphical forms to the widest possible audience. The tool aims to provide access to wordnets in a platform independent way so that user is not bound to use only e.g. desktop computer or a mobile phone to access the network data. This was achieved by developing the tool as a web application which allows user to utilize it on any device that is equipped by a web browser. For good usability on any device, the tool needs to be responsive and to adapt itself to any reasonable size of a user's screen. The responsive design goes in hand with the other goal of the tool which is a visually appealing and modern style. This is essential when using the application for educational purposes as wordnets offer a rich basis for language trainings of both children and adults considering the correspondence of the semantic network structure with the presumed human brain organization. Another goal achieved by developing the tool as a web application was its broad accessibility. With the only requirement of having a web browser installed, the tool can be accessed on almost any device and there is no need for complicate preparation of e.g. school computers for the usage in classes.

From the technical point of view, the new DEBVisDic interface is an HTML document partly generated on the client side by JavaScript according to the data sent by the server in the JSON format. The look of the document is defined by

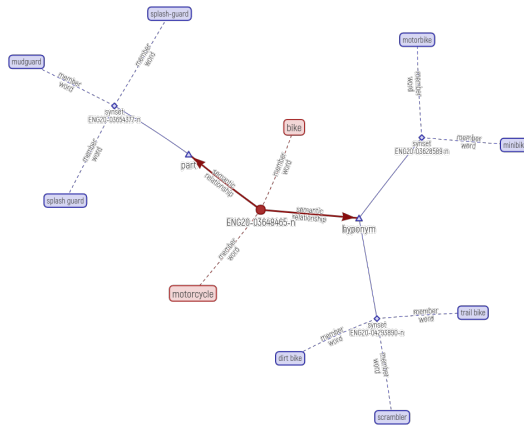


Fig. 4. Graph representation of synset relations in DEBVisDic.

CSS, the content being a result of what was sent by the server. The server sends synsets which contain a searched word or have an ID identical to the one that was requested by the client. The server response contains all the synsets that comply with the searched term and the interface displays by default the first one that was in data and the rest is offered to the user to choose in a sidebar menu on the left (or on the top if the screen is too small for the sidebar). The right (or bottom) part of the page then contains either textual or graph representation of the presented synset. The two views can be switched by two buttons on the top of the sidebar.

In the text mode, the interface displays all available information i.e. the synonymic set contained in the synset (the literals), the hypero-hyponymic path leading to the synset and last but not least the synset semantic relations. These are displayed as a column for each relation with the connected synsets. See Figure 3 for an example of the synset information in text mode.

The graph mode is an alternative representation of relations to which the synset belongs. The central node of the graph is the displayed synset with edges leading to its literals and then to nodes representing all connected semantic relations. An example of the graph-based synset visualization can be found in Figure 4.

6 Conclusion

We have presented a new tool for crowdsourcing reporting of wordnet errors. The process workflow takes into account all the needed phases of lexical database updates and enhancement. After thorough public testing with the

Czech Wordnet, we plan to release the tool for all wordnets developed within the DEB platform.

To present wordnet data in a visually attractive and understandable way for users, we have developed new DEBVisDic Visualization Interface providing both textual and graphic mode for synset preview. This interface is both integrated with the DEBVisDic editor, and as a standalone web application.

Acknowledgements. This work was partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0915/2016).

References

1. Alvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., Rigau, G.: Complete and consistent annotation of wordnet using the top concept ontology. In: Proceedings of LREC 2008. (2008)
2. Tufiş, D., Cristea, D.: Methodological issues in building the romanian wordnet and consistency checks in balkanet. In: Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation. (2002) 35–41
3. Rath, H.H.: Technical issues on topic maps. In: Proceedings of Metastructures 99 Conference, GCA (1999)
4. Rambousek, A., Horák, A.: DEBVisDic: Instant WordNet Building. In: Proceedings of the Eighth Global WordNet Conference, GWC 2016. (2016) 25–29
5. Horák, A., Vossen, P., Rambousek, A.: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In: Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing, Haifa, Israel, Springer-Verlag (2008) 1–15
6. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer (1998)
7. Pociello, E., Agirre, E., Aldezabal, I.: Methodology and construction of the basque wordnet. *Language Resources and Evaluation* 45(2) (May 2011) 121–142
8. Horák, A., Pala, K., Rambousek, A., Povolný, M.: DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Proceedings of the Third International WordNet Conference - GWC 2006, Jeju, South Korea, Masaryk University, Brno (2006) 325–328
9. Čapek, T.: SENEQA - System for Quality Testing of Wordnet Data. In Fellbaum, C., Vossen, P., eds.: 6th International Global Wordnet Conference Proceedings, Matsue, Japan, Toyohashi University of Technology (2012) 400–404
10. Grác, M.: Rapid Development of Language Resources. PhD thesis, Faculty of Informatics, Masaryk University (2013)
11. Rumshisky, A.: Crowdsourcing word sense definition. In: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics (2011) 74–81
12. Nevěřilová, Z.: Paraphrase and Textual Entailment Generation in Czech. PhD thesis, Faculty of Informatics, Masaryk University (2014)
13. Munro, R., Gunasekara, L., Nevins, S., Polepeddi, L., Rosen, E.: Tracking Epidemics with Natural Language Processing and Crowdsourcing. In: AAAI Spring Symposium: Wisdom of the Crowd. Volume SS-12-06., AAAI (2012)

14. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics (2008) 254–263
15. Callison-Burch, C.: Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics (2009) 286–295
16. Wikipedia: Wiktionary — Wikipedia, The Free Encyclopedia (2017) [Online; accessed 2-October-2017].
17. Hanks, P.: Corpus evidence and electronic lexicography. In Granger, S., Paquot, M., eds.: *Electronic Lexicography*. Oxford University Press, Oxford (2012) 57–82
18. Meyer, C.M., Gurevych, I.: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S., Paquot, M., eds.: *Electronic Lexicography*. Oxford University Press, Oxford (2012) 259–291
19. Fuertes-Olivera, P.A.: The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective Free Multiple-Language Internet Dictionary. In Bergenholtz, H., Nielsen, S., Tarp, S., eds.: *Lexicography at a Crossroads*. Peter Lang, Bern (2009) 103–120
20. Christodoulakis, D.: *Balkanet Final Report*, University of Patras, DBLAB (2004) No. IST-2000-29388.
21. Blahuš, M., Pala, K.: Extending Czech WordNet using a bilingual dictionary. In Fellbaum, C., Vossen, P., eds.: *6th International Global Wordnet Conference Proceedings*, Matsue, Japan, Toyohashi University of Technology (2012) 50–55
22. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: CILI: the Collaborative Interlingual Index. In Barbu Mititelu, V., Forascu, C., Fellbaum, C., Vossen, P., eds.: *Proceedings of the Eighth Global WordNet Conference*, Romanian Academy (2016) 50–57

Preliminary Thoughts on Issues of Modeling Japanese Dictionaries Using the OntoLex Model

Louis Lecailliez

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
louis.lecailliez@outlook.fr

Abstract. Recent works aiming at making Linked Data dictionaries make use of the Lemon or OntoLex models. Application to existing dictionaries revealed the need for extensions to the model to properly deal with lexicographic data without loss of information. These works however focus on languages found in Europe, and thus let the issue of Est-Asian lexicography for future exploration. This paper provides a small typology of existing dictionaries in Japan and exposes issues in existing related works that could form the ground of new modules for OntoLex.

Key words: Linked Data, Lemon model, dictionary, e-lexicography, Japanese

1 Introduction

After a first wave of dictionary computerization, where they have been structured in a way close to their existing paper embodiment [13], the next step appears to be graph-based dictionaries [16]. In this trend, dictionaries are made or being converted using formalisms and the technology stack of the Linked Data [5]. In particular, the OntoLex model [12] – based on the Lemon model – initially created to lexicalize ontologies is under development to support more information coming from traditional (electronic) dictionaries. However, as Bosque-Gil [1] note it: “Future steps include the analysis of dictionaries in languages that are underrepresented in the LLOD cloud (e.g. Japanese) to identify further representation challenges”.

The present work aims to start identifying a few of these issues. At first, we expose a concise typology of existing Japanese dictionaries in order to pin down the lexicographic landscape of the Japanese language. In this paper, we focus on the Japanese language but some parts are equally applicable to other languages as well. That’s why we will indicate when a class of dictionary is relevant to the East-Asia area as a whole. Secondly, existing works related to the problem of modelling dictionaries and encoding lexical knowledge of Chinese characters and the Japanese language will be presented. These works are not directly incorporable in OntoLex [18] as such because of an initial divergent goal but form a very viable base of reflexion for a dedicated module.

2 Concise Typology of Japanese Dictionaries

In addition to unilingual and bilingual dictionaries, Japanese features some original kind of dictionaries: some are specific to Japan while others are also found in the whole Sinosphere. Most of these dictionaries emerged because of the characteristics of the Chinese writing system, which was imported in Japan and subsequently derived to write the autochthon language. Others, such as the accent and katakana dictionaries arose respectively from features of Japanese and its vernacular writing system.

Table 1 lists the most common dictionaries found in Japan. For each class, it is specified if it also exists in other countries of the Sinosphere, or if it can be found in Japan only. A dash means it is not specific to either regions. The last column describes the form (writing system and number of characters) of headwords compiled in a given dictionary type, if applicable. Kana encompass hiragana and katakana.

2.1 Chinese Characters¹ Related Dictionaries

The first class of specific dictionary to be found in Japan is the *kanwa-jiten* (漢和辞典). Literally a “Chinese-Japanese Dictionary”, it focusses on explaining Chinese characters and compound words made from Chinese morphemes that were borrowed in Japanese. This is different from the *ch-unichi-jiten* (中日辞典) that are bilingual dictionary of contemporary Mandarin to modern Japanese language. The class may be further split between *kanji-jiten* and *kango-jiten*. The former lists and describes sinograms and the afferent readings, meanings and compounds while the latter focusses on providing lists of compound words that use a given Chinese character. Chinese character dictionaries can also be found in Korea under the generic term of *hanja-sajeon* (漢字辭典).

Another type of kanji dictionary exists that is targeted at non-Japanese people: the bilingual sinogram dictionary. In them, characters are headwords, meaning is given in a foreign language and readings are written in a romanization. The *New Nelson Japanese-English Character Dictionary* [7] for the Japanese-English pair and the *Dictionnaire Ricci de caractères chinois* [17] for Mandarin-French are example of such dictionaries.

2.2 Proverb Dictionaries

Two types of dictionary exist for proverbs. A *kotowaza-jiten* is a book listing idioms: any proverb used in Japanese can fit in this kind of dictionary, regardless of its form. The *yojijukugo-jiten* on the other hand only lists idiotisms made of four Chinese characters. Most of the time entries in such a dictionary feature

¹ In the rest of the paper, the terms “Chinese character” and “sinogram” are used in an interchangeable way to denote the Chinese characters and their use in the whole Sinosphere. The Japanese term “kanji” (漢字) is used only when speaking of sinograms in a Japanese context.

a Japanese reading using Japanese native words and have an idiomatic value, hence the compilation in a different kind of dictionary than a *kango* or *kotowaza* dictionary. Proverb dictionaries are also found in China under the name of *chengyu-cidian* (成語詞典). As most proverbs of Mandarin are made of four characters, there is only one type of dictionary for them in Chinese.

2.3 Specialized Dictionaries Specific to Japan

The written language of Japan was for centuries modeled on literature of ancient times. The diglossia between the spoken and the written languages made it so that dictionaries are well needed for understanding the classical language. This is the *raison d'être* of the classical language dictionary: *kogo-jiten* (literally old language dictionary).

The contemporary Japanese language features a pitch accent that is not uniform between locations. There is a class of dictionary (accent dictionary) made to indicate the “proper” accentuation of words, modelled on the Tokyo dialect. One of them is available on the web [15].

Foreign words of non-Chinese origin are written with the Japanese katakana syllabary. Some dictionaries compile such words. Contrary to other dictionary types, these dictionaries make use of the latin alphabet in the definitions in order to display word in their original writing.

Finally, the various styles that were used to write sinograms through the ages give birth to the need for *jitai-jiten* (style dictionary) that compile writing of characters in different styles. That kind of work is different from the others in that the information it encodes cannot be stored in text form. These dictionaries are used by calligraphers or readers of literary work in manuscript form.

Table 1. Most common and specialized type of dictionaries found in Japan

Dictionary type	Japanese name	Romanized name	Specific to	Headword
Unilingual Dictionary	国語辞典	Kokugo jiten	—	Kana
Bilingual Dictionary	XY辞典	XY jiten	—	Kana or Alphabet
Etymology Dictionary	原語辞典	Gengo jiten	—	Mixed
Sinogram Dictionary	漢字辞典	Kanji jiten	Sinosphere	One kanji
Chinese Compound Dictionary	漢語辞典	Kango jiten	Sinosphere	Kanji
Proverb Dictionary	ことわざ辞典	Kotowaza jiten	—	Mixed
“Four Character Compound Dictionary”	四字熟語辞典	Yoji-jukugo jiten	Sinosphere	Four kanji
Accent Dictionary	アクセント辞典	Akusento jiten	Japan	Kana
Classical Language Dictionary	古語辞典	Kogo jiten	Japan	Kana
Dictionary of Words in Katakana	カタカナ語辞典	Katakanago jiten	Japan	Katakana
Style Dictionary	字体辞典	Jitai jiten	Sinosphere	Kanji

2.4 Modeling Problematics

It is clear from the list of dictionaries listed in Table 1 that one module cannot encode by itself all information required by dictionaries that address such a variety of concerns. As a starting point, the main dictionaries to be modeled are unilingual and *kanwa* dictionaries. Moreover, the different dictionaries target distinct demographics and needs. In particular, the native speaker population and the non-native one have different concerns when it comes to searched entries. A native is probably more interested in checking the meaning of a word or the way to write it in *kanji*, while a foreign learner may be more concerned about the reading of a character or a word and its translation in his native language. Both unilingual and *kanji* dictionaries are used in conjunction even by natives to find the meaning of an unknown word written in *kanji* [2].

This duality is expressed in the dictionaries themselves: unilingual dictionaries will compile entries listed in a phonetic way and sub-entries may be distinguished by graphical forms. On the other hand, *kanji* dictionaries head-words are characters for which multiples readings are listed in a syllabary (often either in hiragana or katakana given the origin of the reading). This particular setting means, as we will see in section 4.2, that a graphical form (if represented by the *lemon:Form* class) cannot be considered as depending only of a lexical entry but needs to be linked somehow to a concept, otherwise some information is lost.

Using the translation module mentioned by Gracia [6], bilingual (characters) dictionaries can be constructed by linking unilingual Japanese to other language lexicons. With this framework in mind, issues can be categorized in three sets related to: (1) Chinese character modeling, (2) representing unilingual Japanese dictionaries without loss of information, and (3) interaction between Chinese characters and other lexical entries of any other kind of dictionary, which is also needed to solve (2).

3 Modeling Chinese Characters

In the cultures that use Chinese characters as their main script or as part of their writing, sinograms are a lexicographic object *per se*. As such, a way to model them is needed.

The complete modeling of Chinese characters, the relation between them and to the language that use them is complex because it requires representing various phenomena and multiple many-to-many relationships. Modeling the sinograms for the Japanese language causes almost all the same problematics as modeling them for the Mandarin language, but additional phenomena needs to be taken in account. For example, additional information exists about the distinction between the type of reading (Sino-Japanese or pure Japanese) and the historical periods from which the readings were borrowed.

Related works on the matter include Hantology [3], an ontology derived from the Chinese writing. It was later expanded to include *kanji* as well [9].

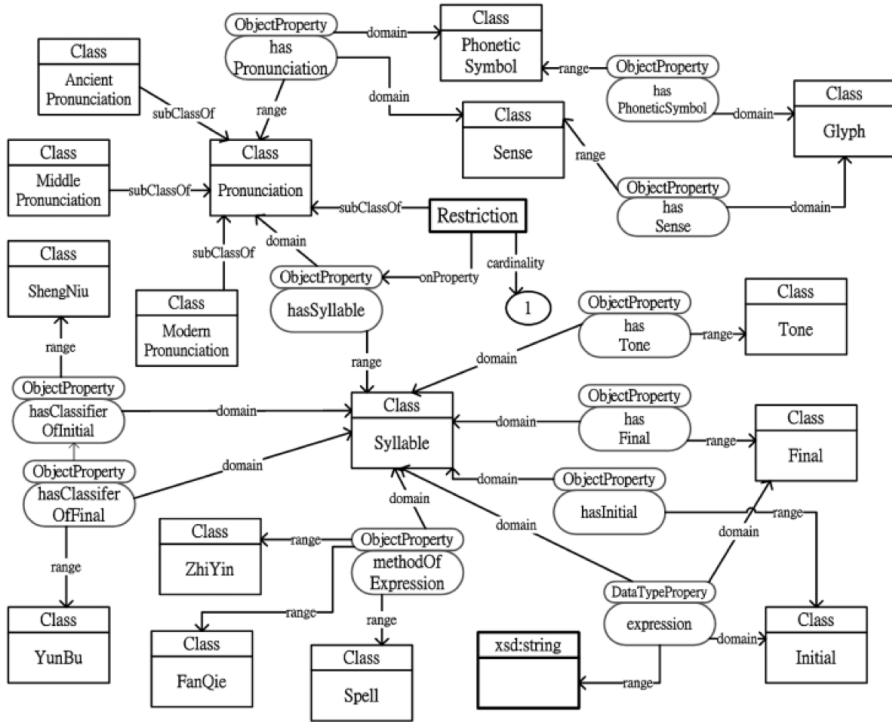


Fig. 1. Sinogram pronunciation description modeled in OWL from [3].

The ontology is based on the character decompositions and definitions given by the *Shuowen Jiezi* (說文解字) dictionary [19]. While their goal was not to encode the content of this dictionary nor any others, the meticulousness of their modelling and its implementation in Web Ontology Language [18] are a suitable base for later work. The whole ontology is detailed, schematized and explained in [3]. The sub-graph of the model dedicated to pronunciation description is reproduced as Figure 1.

A great attention has been made to the problem of character variants in Hantology. While there exist a sheer amount of characters – more than 50,000 referenced in the *Daikanwa Jiten* (大漢和辭典) – most are actually variants that were used at a given time and place. This information is important for automatic processing where the number of entries matters but at the same time introduce noise when they are not needed.

Link between variants is also an important point for linking sinograms between languages that may use different glyphs for the same character: it can act as a link between data resources of various languages and thus increase the connectivity in the Linguistic Linked Open Data (LLOD) cloud.

4 Modeling the Japanese Lexicon

4.1 The JLP-O Model

Linked Data modelling of the Japanese lexicon already went under scrutiny in various work of Joyce and Hodošček [10,11] about building an ontology of lexical properties of Japanese: the JLP-O model. Although the aim of these works is not to represent dictionaries *per se*, advance on this front is particularly interesting as the chosen model is built on Lemon. Joyce and Hodošček [10] derive five classes from lemon's *LexicalEntry* class to fit the perceived needs for representing the Japanese lexicon. Beside the complexity of the Japanese writing system, the Japanese lexicon feature a great deal of compound words and three of the classes (*BoundUnit*, *SimpleWord*, *ComplexWord*) were created specifically to deal with it.

The Figure 2 reproduces the figure 3 from [10]. It illustrates how the word “yomu” (to read) is handled in the JLP-O model. The model regroups the various graphical forms of a word under the *canonicalForm* and *orthographicForm* properties of a main lexical entry. It raises two issues. First, forms that carry a slightly different meaning are not separated in their own lexical entry. While this is not a problem in a tradition paper dictionary, it may cause wrong inference while using the graph (see section 4.2 below). In this example, both “読む” and “詠む” are grouped in the same entry despite the latter being only used in the context of reading poetry and song [8]. It means that the compound verb “読み始める” (to start reading) that links the “yomu” verb is also indirectly linked to forms that are actually not be used to write it. The second issue is related to the first: each form is wrapped in a blank node. RDF blank nodes are anonymous and thus cannot be reused as the object or subject of other properties. The first stated issue thus arises and cannot be solved because the precise forms that needs to be linked are not directly accessible.

4.2 Making Wrong Inferences

The *yomu* example shows how the pronunciation of a word and its graphical forms are intertwined in relation their meaning. A given word may feature very different senses if written with different characters. Reciprocally, a given written form can be used to write word of very different meaning. For example, “十分” read as *juubun* means enough; read as *juppun* or *jippun* it means ten minutes.

It is thus essential to link a couple of (form, reading) to a meaning. Failure to do so would allow a program using the graph to make inferences that are not true. For example, let have a *LexicalEntry* instance with form F_1 (meaning S_1) and F_2 (meaning S_2) which is linked to ontology entities denoting concepts S_1 and S_2 . A program reading the graph could make the respective inferences by transitivity: F_1/F_2 can be used to express concept S_1/S_2 but also that F_2 can denote S_1 and F_1 means S_2 . The two later predicates are false.

The way the class *Form* is linked to a *LexicalEntry* independently of a the *LexicalConcept* and *LexicalSense* class is a problem in modelling Japanese dictionaries. In addition, a similar problem exists in modelling *kanji*.


```

jlpo:読む_動詞-一般
a jlpo:SimpleWord ;
lemon:canonicalForm [
  lemon:writtenRep "読む"@ja ;
  jlpo:decomposition (
    [ jlpo:Character jlpo:読_character ]
    [ jlpo:Character jlpo:む_character ] ) ;
  jlpo:use [ jlpo:frequency 23324 ; jlpo:corpus "BCCWJ" ] ] ;
jlpo:orthographicForm [
  lemon:writtenRep "読む"@ja ;
  jlpo:decomposition (
    [ jlpo:Character jlpo:読_character ]
    [ jlpo:Character jlpo:む_character ] ) ;
  jlpo:use [ jlpo:frequency 20382 ; jlpo:corpus "BCCWJ" ] ] ;
jlpo:orthographicForm [
  lemon:writtenRep "よむ"@ja ;
  jlpo:decomposition (
    [ jlpo:Character jlpo:よ_character ]
    [ jlpo:Character jlpo:む_character ] ) ;
  jlpo:use [ jlpo:frequency 322 ; jlpo:corpus "BCCWJ" ] ] ;
jlpo:orthographicForm [
  lemon:writtenRep "詠む"@ja ;
  jlpo:decomposition (
    [ jlpo:Character jlpo:詠_character ]
    [ jlpo:Character jlpo:む_character ] ) ;
  jlpo:use [ jlpo:frequency 653 ; jlpo:corpus "BCCWJ" ] ] ;
# [... 9 other orthographicForms ...]

```

Fig. 2. Reproduction of figure 3 from [10]: Part of the RDF representation for the SimpleWord lexical entry '読む' in Turtle format.

5 Silex: Towards a Lemon Module for Chinese Characters

Chinese characters are an important lexicographic object for the Japanese language processing. It must be tackled first because almost every other kind of dictionary will link them from other lexical entries. Unilingual dictionaries typically feature annotations that need such an atomic decomposition to be fully encoded.

Sinograms also pose an issue of many-to-many relationships between readings and meanings of the same character. A similar problem exists for the Japanese orthography as a whole, thus solving the problem at the character level will provide a template to solve it a higher level.

Finally, as sinograms are used or were used in a variety of East-Asian language, it makes sense to model them in the most language agnostic way. That allows reuse of entities in a multilingual context, increasing the connectivity within the Linguistic Linked Open Data cloud. From this point of view, a language agnostic term for Chinese character should be chosen for the main entity. The term sinogram answers this problem elegantly by burying the reference to the Chinese culture in a root from latin and avoiding the use of

localized term such as *kanji* (Japanese), *hanzi* (Chinese), *hanja* (Korean), *hán tự* (Vietnamese). And it abbreviated nicely as Silex, the Sinogram Lexical module.

Acknowledgments. This paper was written with the support of the MSMT-10925/2017-64-002 scholarship grant from the Czech Ministry of Education, Youth and Sports.

References

1. Bosque-Gil, J., Gracia, J., & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. *DICTIONARY News*, 7.
2. Breen, J. (2004). Multiple Indexing in an Electronic Kanji Dictionary. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries* (pp. 1-7). Association for Computational Linguistics.
3. Chou, Y. M., & Huang, C. R. (2005). Hantology: An ontology based on conventionalized conceptualization. In *Proceedings of the Fourth OntoLex Workshop. A workshop held in conjunction with the second IJCNLP*. October (Vol. 15).
4. Chou, Y. M., & Huang, C. R. (2013). Hanzi zhishi de xingshi biaoda [汉字知识的形式表达]. *Dangdai yuyanxue* [当代语言学], 15(2), 142-161.
5. Gracia, J., Kernerman, I., & Bosque-Gil, J. (2017). Toward Linked Data-native Dictionaries. In Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., & Baisa, V. (eds) *Proceedings of eLex 2017 conference*, September 19-21, Leiden, Netherlands. Lexical Computing CZ s.R.O, Brno, Czech Republic.
6. Gracia, J., Villegas, M., Gómez-Pérez, A., & Bel, N. (2016). The apertium bilingual dictionaries on the web of data. *Semantic Web*, (Preprint), 1-10.
7. Haig, J., Nelson, A. (1997). *The new Nelson Japanese-English character dictionary*. C.E. Tuttle Co.
8. Hayashi, S., Nomoto, K., Minami, F., & Kunimatsu, A. (1992). *Sanseid-o Reikai Shinkokugo Jiten*, 3rd edition [三省堂例解新国語辞典 第三版].
9. Huang, Y. M., Huang, C. R., & Hong, J. F. (2008). The Extended Architecture of Hantology for Kanji. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D. (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, May 28-30, Marrakech, Morocco. European Language Resources Association (ELRA).
10. Joyce, T., & Hodošček, B. (2014). Constructing an ontology of Japanese lexical properties: Specifying its property structures and lexical entries. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)* (pp. 174-185).
11. Joyce, T., Masuda, H., & Hodošček, B. (2016). Constructing a Database of Japanese Lexical Properties: Outlining its Basic Framework and Initial Components. *Tama University Global Studies Department Bulletin Paper*, 8, 35-60.
12. McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., & Baisa, V. (eds) *Proceedings of eLex 2017 conference*, September 19-21, Leiden, Netherlands. Lexical Computing CZ s.R.O, Brno, Czech Republic.
13. Měchura, M. (2016). Data Structures in Lexicography: from Trees to Graphs in Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pp. 97-104, 2016.
14. Nagasawa, K. (1991). *Sanseid-o Jiten* 4th edition [三省堂漢和辞典 第四版].

15. Nakamura, I., Minematsu, N., Suzuki, M., Hirano, H., Nakagawa, C., Nakamura, N., Tagawa, Y., Hirose, K., Hashimoto, H. (2013). Development of a web framework for teaching and learning Japanese prosody: OJAD (Online Japanese Accent Dictionary). *Proceedings of INTERSPEECH*, pp.2254-2258.
16. Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4), 396-418.
17. Ricci Institute. (2001). Grand dictionnaire Ricci de la langue chinoise. Desclée de Brouwer, Paris.
18. W3C Ontology Lexicon Community Group. (2016). Final Model Specification. https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
19. Xu, S. (121). Shuowen Jiezi [說文解字].

Wikilink – Wikipedia Link Suggestion System, Its Problems and Possible Solutions

Vojtěch Mrkývka

Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno, Czech Republic
421310@mail.muni.cz

Abstract. In my bachelor thesis I have created a tool, which was able to analyse paragraphs from the given Wikipedia article and suggested the editors internal links, that they could add into this article.

In this paper I return to my thesis and to the tool and evaluate its procedures. I offer ways to solve the problems associated with it, ways which would lead to overall improvement and acceleration of the tool.

Key words: lemmatization, Wikipedia, text mining

1 Introduction

Wikipedia is without doubts the largest encyclopedia in the world [1]. The difference between Wikipedia and its competitors is the flexibility or in the other words possibility to quickly react to current events and discoveries.

Wikipedia is an encyclopedia where anyone can add some information or edit the existing one. But in practice only a tiny percent of the users actively creates new content. Nevertheless there are active efforts to increase the number of editors with regards to conservation of the same or better quality of the content. When I was creating my tool I tried to follow the same principle. I hoped that it would ease the users' editing work and therefore it would increase quality of the target contribution.

2 About Wikipedia

Wikipedia is a free internet encyclopedia, which allow its users to read its content and participate on its development without need to pay any subscription fee or for any licence [2].

It was launched on 15 January 2001 by its founders Jimmy Wales and Larry Singer. Although Wikipedia was meant only as a side-project to encyclopedia Nupedia, which included only articles verified by renowned experts (Wikipedia should have acted like a source of topics and drafts for new articles), its popularity raised rapidly. Today Wikipedia contains more than 5.5 million of articles in its English mutation only. Although the number of visitors is astonishing, the problem is only a small percent of registered users is also helping as editors (by the recent data only about 0.4% edited at least once during the last month) [3].

3 Editing Wikipedia

One of the problems associated with the number of editors was the way of editing. Until 2013 the only possibility to create or change any article was to use wikitext, special markup language used for saving Wikipedia articles. This was inconvenient for less technically competent users. Because of it there was a great unbalance between different fields of study. To attract new editors is what Wikipedia Usability Initiative took as its aim. In its five-year-plan for years 2010–2015 proposed creation of new rich-text editor, which would, according to their idea, help to increase the number of Wikipedia contributors to 200,000 (from original number of about 80,000) [4]. In 2012 new editor called VisualEditor was presented in the Wikimedia project. One year later the gradual inclusion of this tool into different language mutations of Wikipedia started [5].

4 Linking Wikipedia

Because of the format of Wikipedia, the internet site, the individual links between Wikipedia articles are made as hypertext links. All of the Wikipedia's links can be separated into three groups. Internal links are those, which are linking pages within Wikipedia or its sister projects (Wiktionary, Wikidata, Wikibooks and others) [6], external links are targeting some external page [7]. Internal links can be further divided to truly-internal, which keeps link within single project and its single language mutation (for example English Wikipedia), and semi-internal, which do not meet the previous criteria.

Insertion of internal and external links is driven by different rules. External links usually connect the article with the source of extended information about the topic. For this reason they are usually situated in reference part of the article and not in the article's body [7]. The internal links on the other hand can help reader to understand the unknown parts of the text by exploring the related topics. They fulfil the role of classic *see something*, which is often present in paper encyclopedias [6]. Even their way of writing is different when the wikitext editation is used.

To keep the article readable, there is a rule on Wikipedia which states that only the first occurrence of some topic in the article's text (information boxes and similar do not count) should be linked also only the articles where is probability to help user with extending knowledge of this articles topic or can bring extending information about it should be linked [6].

5 Wikilink – the Link Suggestion Tool

In my bachelor thesis I designed and created a system, which suggested non-present truly-internal links to the editor of the article. In the beginning I wanted to create an automatic tool (bot), which would insert the links independent on any user. The problem was, create a tool, which would be precise enough to insert only correct links (within some small tolerance), would be task, which

was at the time far superior to my knowledge. Because of this I created the tool, which only suggested possible links to the user and let him to make the final decision.

Wikilink, as I named the tool, consisted of two parts – the client part and the server part. The client part was written in Javascript with jQuery. In the beginning the purpose of the client part was to send data to the server (processing) part and display the results back on the Wikipedia website. Due to the same-origin policy, which prevents AJAX response from another domain, the final version of the client part only send data without receiving any [8]. I used VisualEditor as the source of the data due to the assumption that new users would use it rather than wikitext editing. I avoided editing the result after saving because it would create more unnecessary versions of the article.

Data gained from the client part were processed by the server part. The server part, written in Python¹, removed undesirable elements like information boxes or references. The following processes split individual paragraphs into tokens, lemmatized them and searched for potential links in the reference file.

The reference file was made from Wikipedia backup database dump [9], more specifically from the list of all pages in the main namespace² which is generated multiple times every month. This file had to be lemmatized and sorted by lemma for faster functioning.

Output of the tool consisted of the text of every paragraph followed by the table. This table consisted of alphabetically sorted pairs of the link text and the article name to which it should be connected. Because of the first occurrence rule I removed those lines, where the suggestion was done in any of the previous paragraphs.

6 Shortcomings of the First Version

Although the Wikilink was principally working, meaning it suggested new links, which were not present in the article, it contained number of shortcomings, which limited its release and possible wide spreading.

First problem was solely based on user-friendliness. When I was writing my bachelor thesis I wasn't able to integrate the button, which started the whole analysis, into the VisualEditor interface. More specifically I wasn't able to force any Javascript code to start only after the VisualEditor was loaded. Due to this the button to start the Wikilink's process had to be situated outside the editor interface. Because VisualEditor is sometimes loaded without page reload (for example from article itself), the button had to be displayed also on these pages. In this case, the button only displayed special alert message.

Other shortcomings occurred in the server part during analysis. Analytic process took lemmatized tokens as an input and tried to find them in the reference list. Output contained two lists, where lemmatized token:

¹ version 2.6

² The main namespace contains articles, other namespaces contain portals, help pages, categories and others [10].

- Fully corresponded to lemmatized article name in the reference list (full match)
- Partially corresponded to lemmatized article match in the reference list (partial match)

Union of these lists was the input for the second run of the analytic process. In the second run, the bi-grams were analysed and so on. When the run returned empty result in both of the lists, the cycle of analytic processing stopped. Because of size of the reference list, even the bisection search³ returned results very slowly (see Table 1). Furthermore I kept links from full match list even if there was found longer string on the same place. I believed, that there are cases, where the link on shorter term can be in the context more specific than on the longer term. In practice, however, I wasn't able to prove this claim, so I think that if there is special case where this concept is right, it occurs so rarely, that there is no need to take it into account.

The last shortcoming was the output format. As I stated before, initially I wanted to display the results back on Wikipedia, ideally right inside the editing window, but due to the same-origin policy, which I wasn't able to overcome at the time, I created special output page where the analysis results were displayed. Thinking of clarity improvement, I ordered the list of results under every paragraph alphabetically, but due to high number of the false results, the solution was far from perfect.

7 Possible Solutions

Shortcomings presented in the previous chapter can be sorted into three overlapping categories – accuracy, overall speed and user-friendliness.

From the perspective of accuracy it is necessary to filter links to the articles which are invalid (or common) by its nature. Example of the very common suggestion can be article *Comma*, to which redirects article name *.*. It is very unlikely that any of the Wikipedia articles wouldn't have any punctuation.

³ using *bisect* package [11]

Table 1. Statistical representation of Wikilink results on random articles. Relevant links column is purely opinion-based, but it can outline the precision of the tool (however links already present in the article are skipped, so the true number would be probably higher).

Article	Word count	Avg. run time	Suggested	Relevant
<i>Bergelmir</i>	89 words	5.45 s	34 links	2 links
<i>Pavel Suchý</i>	420 words	10.44 s	95 links	11 links
<i>SARS</i>	749 words	19.53 s	183 links	59 links
<i>Spolek přátel Rumburku</i>	1,171 words	47.09 s	319 links	34 links
<i>Brno</i>	13,842 words	393.00 s	1,978 links	329 links

Concurrently can be assumed, that only a small number of articles should truly link to the *Comma*. These cases could be removed by application of the simple stop-list, or regular expression. The more complicated are cases where the lemmatization fails and for the specific string is suggested for example some abbreviation. Stop-list could be used to remove the most common mistakes, but not all of them. Because the most of these mistakes link to particular parts of speech and common words the filter could follow them. More specifically the system should filter links made only by preposition, conjunction or one of the verbs *to be* or *to have*.

Considering my previous statement, that shorter strings are usually worse than longer on the same place, the analytic phase of the application can be changed from the multiple runs to the single one. This could improve not only accuracy, but also the overall speed, the second condition for better usability. Speed can be further improved by focusing more on preprocessing of the reference list, which would lower the user requirements. For example the reference list can be rebuild into form of the tree, which can be saved in the JSON format. Crawling the tree would find longer links significantly faster than using the old method.

As for user-friendliness, the previous form of the input and the output was influenced by two mentioned problems – VisualEditor implementation and especially the same-origin policy, which made AJAX requests across different domains impossible. Since my bachelor thesis I found the ways, how to overcome both of these inconveniences. On the website of MediaWiki, whose editing team created the VisualEditor, could be found parts of code and tutorials to operate with editor's interface [12]. The same-origin policy could be evaded by the HTTP header *Access-Control-Allow-Origin* [13]. The results then could be displayed straight inside the VisualEditor interface.

8 Conclusion

Wikilink as a tool did in principle what it was designed for. Although it was far from practical tool due to its speed and accuracy, it outlined some trends which can be followed in further development. In this article I tried to present the problems of the tool with possible solutions, but they are only some of the improvements by which can be the tool enhanced. There are other possibilities which can be explored, for example separate AJAX request, which would suggest possible links continuously. But the aspects of further improvements must be evaluated before any judgements.

Acknowledgments. This work was supported by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0915/2016).

References

1. Wikipedia: Wikipedia:size comparisons — wikipedia, the free encyclopedia (2017) [Online; accessed 29-October-2017].
2. Wikipedia: Wikipedia:about — wikipedia, the free encyclopedia (2017) [Online; accessed 29-October-2017].
3. Wikipedia: Special:statistics — wikipedia, the free encyclopedia (2017) [Online; accessed 29-October-2017].
4. Strategic Planning: Product whitepaper — strategic planning, (2011) [Online; accessed 29-October-2017].
5. MediaWiki: Visualeditor — mediawiki, the free wiki engine (2017) [Online; accessed 29-October-2017].
6. Wikipedia: Wikipedia>manual of style/linking — wikipedia, the free encyclopedia (2017) [Online; accessed 27-October-2017].
7. Wikipedia: Wikipedia:external links — wikipedia, the free encyclopedia (2017) [Online; accessed 27-October-2017].
8. Mozilla Foundation: Same-origin policy - web security | mdn (2017) [Online; accessed 20-October-2017].
9. Wikimedia Foundation: Wikimedia downloads (2017) [Online; accessed 20-October-2017].
10. Wikipedia: Wikipedia:namespace — wikipedia, the free encyclopedia (2017) [Online; accessed 20-October-2017].
11. Python Software Foundation: 8.5. bisect — array bisection algorithm — python 2.7.14 documentation (2017) [Online; accessed 20-October-2017].
12. MediaWiki: Visualeditor/gadgets — mediawiki, the free wiki engine (2017) [Online; accessed 20-October-2017].
13. Mozilla Foundation: Cross-origin resource sharing (cors) - http | mdn (2017) [Online; accessed 20-October-2017].
14. Mrkývka, V.: Návrh chybějících interních odkazů v české wikipedii (2016) [Online; accessed 31-October-2017].

Part II

Semantics and Language Modelling

The Ordered-triple Theory of Language: Its History and the Current Context

Aleš Horák and Karel Pala

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

Abstract. In this paper, we recall the historical perspectives of the Ordered-Triple Theory of Language (OTT) whose authors are Materna, Pala and Svoboda. The Ordered-Triple Theory, as the title suggests captures three fundamental components of a language system, i.e. syntax, semantics and pragmatics, and is fully comparable with similar linguistic theories. It became a starting point for further interconnection of logic, linguistics and informatics thanks to the intensive mutual cooperation of Pala and Materna at the newly established Faculty of Informatics from 1995.

We show the subsequent milestones related to OTT and its realisation by means of the transparent intensional logic (TIL) in relation to the natural language processing (primarily Czech).

Key words: ordered-triple theory, theory of language, syntax, semantics, pragmatics, procedural grammar of Czech, transparent intensional logic, TIL

1 Introduction

The authors of the Ordered-triple Theory of Language (further OTT) are Pavel Materna, Karel Pala a Aleš Svoboda who proposed it at FF UJEP¹ during 1976–79 and published in *Brno Studies in English* 12 [1] and 13 [2].

In a sense, OTT was a reaction to the two language theories that were influential by this time, particularly to Chomsky's generative approach [3] and the Prague functional generative framework (FGD) by Sgall et al [4]. The main difference is that OTT had not been conceived explicitly as generative but it had allowed to deal with both recognoscative and generative devices (see below). Prague's FGD from the beginning contained semantic component in the form of the tectogrammatical level which was based on a set of actants (semantic roles). However, it did not use any logical formalism in contrast with OTT. The relevant feature of OTT was a consistently grasped semantic component while, for instance, Chomsky's generative grammars were primarily based on syntax.

¹ The Faculty of Arts at the former Jan Evangelista Purkyně University, current Masaryk University, in Brno

Later, attempts appeared to mate the generative approach with Montague's intensional logic [5,6].

In the following text we would like to characterize OTT briefly and present the main interesting results that have been reached within this framework.

2 The Ordered-triple Theory

The Ordered-triple Theory offers a theoretical framework for a formal natural language description which considers all the basic components of any semiotic system, i.e. it captures relations between language user and real world and relations between language and language user. In other words, the proposed framework consequently takes into account the syntactic, semantic and pragmatic component of language (in Morris' sense [7]) and takes them as a unified system. The assumption is that natural language expressions consist of syntactic, semantic and pragmatic constituents and thus can be described as ordered triples comprising:

⟨semantic component, formal language expression, pragmatic component⟩

2.1 Part I – Semantics

The first part of OTT is mainly about semantics. It is conceived as consequently intensionalistic, thus at the beginning we pay attention to the problems of the extensional approach to semantics in which language expressions denote what we call extensions, i.e. mainly individuals, classes, relations and truth values. Then we give reasons for a different approach that does not suffer from the non intuitive consequences of the extensionalism – extensional analysis does not allow to distinguish empiric sentences from non-empiric ones and understanding from verification. The intensionalistic approach allows to handle also other referential phenomena like individual roles, propositional attitudes, or episodic verbs.

The logical analysis of natural language in OTT relies on the transparent intensional logic (further TIL) in the form of the system, whose author was in 1970s P. Tichý, who after August 1968 left ČSSR,² spent short time in UK and then moved to New Zealand where he started to work in Dunedin at the Otago University [8].

We define the basic concepts of the intensional semantics consisting of the epistemic base constituted by four sets: the universe (ι , set of individuals), the set of truth values (θ), the set of possible worlds (ω) and the (continuous) set of time moments (τ). The simple type theory is used to produce derived entities and the most typical intensions are given using the operation called *intensional descent*, i.e. the application of a (possible) *world* w and a *time moment* t to arrive from an intension (an $((\alpha\tau)\omega)$ -function) to the corresponding extension (an α -object, where α denotes an extensional type).

² The Czechoslovak Socialist Republic

Further relevant concepts are introduced: constructions (atoms, applications, abstractions) and the relation between language expression, construction and intension. Also class of what is called language constructions is distinguished among all other constructions. Class of the language constructions is to be understood as a class of the constructions that can be expressed by natural language expressions. In this respect, a grammar of language can be taken as a set of rules enabling to derive constructions reflecting these expressions from the structure of the language expressions: simple examples of such rules are given. Such grammar may contain syntactic rules having the form of context-free rules and semantic rules operating on the output of the grammar and providing formulae of the λ -calculus as a result.

Consequently, it is shown how to extend the sets consisting of the universe, truth values and possible worlds with further sets. Thus the set of the time moments allows to perform a more subtle semantic analysis capturing time characteristics including grammatical tenses. Similarly, the semantics of locational adverbs can be examined if we add to them the set of space points.

The next task is to handle deictic (indexical) expressions as e.g. personal pronouns, and to establish relations between what we call external pragmatics and semantics. Finally, the attention is paid to the semantic relations of expressing, denoting and constructing on one hand and to the pragmatic relations of demonstrating (internal pragmatics) and determining (external pragmatics) on the other.

2.2 Part II – Computer Tools for the OTT

So far, we have characterized the general framework of the OTT, now we would like to deal with its computer application (model). In Chomsky's and Sgall's approaches mentioned above the language levels are used. Within the OTT, we can also have syntactic and semantic component including morphology and, moreover, we are interested in their algorithmic description leading to a computer model for OTT.

In the first version of OTT, a context-free like grammar was used, particularly, it was a set of procedural rules for Czech [9] called a procedural grammar inspired by T. Winograd for English [10]. It was the first and only procedural grammar for Czech implemented in programming language LISP 1.5, containing 34 LISP functions, and tested on the mainframe TESLA 200. It has to be remarked that in 1976 no standard morphological analyzer for Czech existed, therefore a morphological input for the Czech procedural grammar was prepared manually and had a form of a file of word forms with corresponding parts of speech and lists of grammatical features. This "syntactic dictionary" is, in fact, almost identical with the output of the present-day morphological analyzers including our Majka [11,12].

The analyser produced a syntactic structure of a sentence in the form of the labelled tree graph which could serve as an input for the semantic analyser [13] as well as for rules handling attitudes of language users in the internal pragmatics framework (this is missing both in Chomsky's and Prague approaches). The

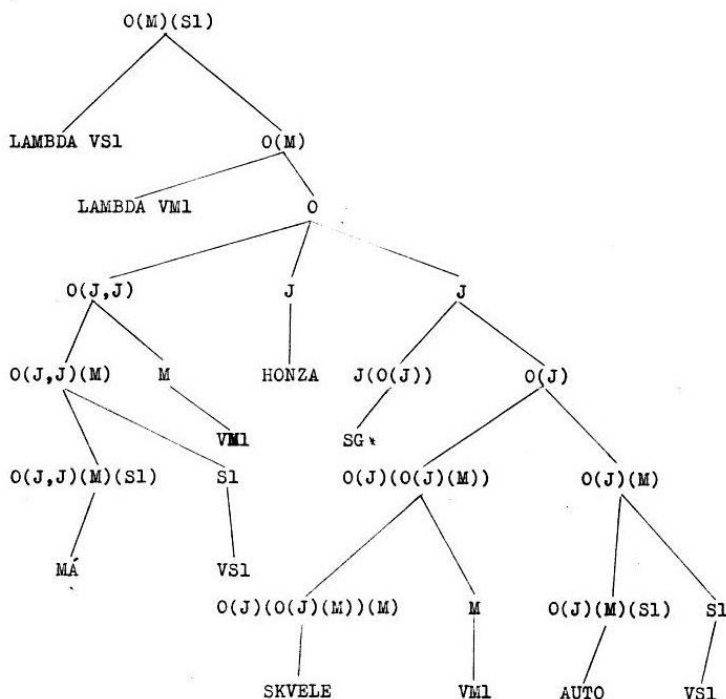


Fig. 1. An example of a semantic tree from [13].

semantic analyser was implemented in LISP 1.5 and tested on TESLA 200 as well. It produced a tree structure representing a λ -calculus formulae obtained from the syntactic tree of the Czech sentence produced by the procedural syntactic analyser (see Figure 1 for an example). λ -calculus formulae are corresponding to the natural language constructions and can be understood as semantic representations of the analyzed natural language (in this case Czech) sentences.

2.3 Other Theoretical Approaches

With contentment, we can say that the mentioned results obtained in OTT (procedural grammar of Czech, semantic analyser, both written in LISP) were in their time fully comparable with the Sgall's FGD framework and with the similar results obtained in the area of transformational grammars. It has to be stressed that OTT has proved itself as a consistent starting point for computer processing of the natural language, particularly Czech, in contrast with American results oriented primarily towards English.

A historical remark [14]: above we have been dealing mainly with the results achieved until 1989. After this A. Svoboda had to move to Opava University,

Materna and Pala finally started to teach at the Faculty of Informatics at Masaryk University (FI MU), where the work on OTT in a sense further continued.

3 Continuation of the OTT at FI MU

The Faculty of Informatics MU was established in 1994, Pala started teaching there in 1995, as well as P. Materna who offered lectures about TIL to the students of informatics in the same year. Thanks to this, Materna got in touch with students who were interested in logical analysis of natural language. One of them was in 1995 A. Horák, who became strongly attracted by the one of the central topic of AI, particularly, by the analysis of the natural language and especially the analysis of meaning. He investigated the problems of the logical and semantic analysis of language in his diploma thesis and also dissertation. The purpose of this work was to make a progress in TIL, which is an important part of the OTT, and continue with the computer implementation enabling to translate standard natural language sentences (Czech, in the first step) into the constructions of the intensional logic. As we have said above, first steps in this direction were presented in [13] and also [15].

However, Horák's research went further and brought a new original result published as the Normal Translation Algorithm (NTA [16]) containing the syntactic analyser Synt [17,18,19] which employs context actions translating syntactic trees of standard Czech sentences to corresponding intensional constructions (expressed as λ -formulae). These results became a base for a further cooperation with P. Materna and later also with M. Duží which led to further development of TIL [20] and the corresponding approaches within several grant projects (GAČR 2005–2007, 2010–2012, 2015–2017), firstly taking place at the FI MU and then also at the Technical University of Ostrava.

The verification of the TIL logical analysis theory included building a large corpus of TIL logical constructions [21] suitable for explicating various language phenomena in common Czech texts. The corpus consists of more than 5,000 sentences that were semi-automatically analysed and translated according to NTA and used for checking by human logicians. An example logical analysis from this corpus is displayed in Figure 2.

3.1 Valency Frames and the OTT

One of the recent results in the NLP Centre is a valency database for Czech language named VerbaLex [22,23]. It contains approx. 10,500 Czech verbs and it is the largest valency database for Czech. Since the verb valency frames represent verbs as predicates with their arguments they can be linked with the linguistic constructions in the TIL [24]. From this point of view we can relate the VerbaLex to the OTT and to exploit it in our further research.

Družice zaznamenaly zrod třetího přechodně trvajícího radiačního pásu Země.

$\lambda w_1 \lambda t_2 [P_{t_2},$ $[Onc_{w_1},$ $\lambda w_3 \lambda t_4 (\exists x_5) (\exists x_6) (\exists i_7) ($ $[Do_{w_3 t_4},$ $x_5,$ $[Perf_{w_3, x_6}]$ $]$ $\wedge x_5 \subset \text{družice}_{w_3 t_4}$ $\wedge [$ $[Of,$ $[Numerize, \text{zrod}, \text{třetí}],$ $[$ $[\text{přechodně}, \text{trvající}],$ $[Of,$ $[\text{radiační}, \text{pás}],$ země $]$ $]$ $]_{w_3 t_4},$ i_7 $]$ $\wedge x_6 = [\text{zaznamenat}, i_7]_{w_3}$ $)$ $],$ Anytime $] \dots \pi$	družice ... $(oi)_{\tau\omega}$ zaznamenat ... $((o(o\pi))(o\pi))_{\omega i}$ zrod ... $(oi)_{\tau\omega}$ třetí ... τ Numerize ... $((oi)_{\tau\omega}(oi)_{\tau\omega}\tau)$ přechodně ... $((oi)_{\tau\omega}(oi)_{\tau\omega})$ trvající ... $((oi)_{\tau\omega}(oi)_{\tau\omega})$ radiační ... $((oi)_{\tau\omega}(oi)_{\tau\omega})$ pás ... $(oi)_{\tau\omega}$ země ... $(oi)_{\tau\omega}$ Of ... $((oi)_{\tau\omega}(oi)_{\tau\omega}(oi)_{\tau\omega})$ družice ... $(oi)_{\tau\omega}$ Anytime ... $(o\tau)$ Onc ... $((o(o\tau))\pi)_{\omega}$ Do ... $(o(oi)(o(o\pi)))_{\tau\omega}$ Perf ... $((o(o\pi))(o(o\pi)(o\pi)))_{\omega}$ P ... $((o(o(o\tau))(o\tau))\tau)$ (verbal object) $x_6 \dots (o(o\pi)(o\pi))$
--	--

Fig. 2. An example TIL logical analysis in the corpus of TIL constructions. The English equivalent of the analysed sentence is “*Satellites recorded the birth of the third temporarily-sustaining radiation belt of the Earth.*”

4 Conclusions and Future Directions

To conclude, we may say that the OTT as such provides a complex theoretical framework for the NLP research within which all relevant components (syntax, semantics, pragmatics) are present. Moreover, the OTT is open to the methodological variability in the present-day NLP since some of its parts rely on rule-based techniques (particularly TIL) while in other parts statistical methods and machine learning (parsing) can be used. In this respect, the OTT may be characterized as a hybrid approach.

We would like to stress one more point: thanks to the TIL formalism, the OTT can serve as a formal tool for handling knowledge representations and inference. In this respect, the OTT can be regarded as a basis for future knowledge-rich question answering systems based on full logic of the underlying discourse.

Acknowledgements. This work has been partly supported by the Czech Science Foundation under the project GA15-13277S.

References

1. Svoboda, A., Materna, P., Pala, K.: An ordered-triple theory of language. *Brno Studies in English* 12 (1976) 159–186
2. Svoboda, A., Materna, P., Pala, K.: The ordered-triple theory continued. *Brno Studies in English* 13 (1979) 119–165
3. Chomsky, N.: *Aspects of the Theory of Syntax*. Volume 11. MIT press (2014)
4. Sgall, P., Nebeský, L., Goralčíková, A., Hajičová, E.: *A Functional Approach to Syntax: in Generative Description of Language*. Elsevier, New York (1969)
5. Dowty, D.R.: *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Reidel (1979)
6. Montague, R.: *Formal Philosophy*. Yale University Press, New Haven (1974)
7. Morris, C.: *Logical Positivism, Pragmatism and Scientific Empiricism*. Number 449 in *Exposés de philosophie scientifique*. Hermann et Cie (1937)
8. Tichý, P.: *The foundations of Frege's logic*. Walter de Gruyter, New York (1988)
9. Palová, I.: The syntactic analyzer for Czech. In: *Papers at the Conference on Cybernetics, Prague* (1976)
10. Winograd, T.: Understanding natural language. *Cognitive psychology* 3(1) (1972) 1–191
11. Šmerk, P.: Fast morphological analysis of Czech. In: *RASLAN 2009, Recent Advances in Slavonic Natural Language Processing*. (2009) 13–16
12. Jakubiček, M., Kovář, V., Šmerk, P.: Czech morphological tagset revisited. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*. (2011) 29–42
13. Cihánek, P.: The semantic analyzer for Czech. In: *Papers at the Conference on Cybernetics, Prague* (1976)
14. Horák, A., Pala, K.: Pavel Materna a tříšložková teorie jazyka (in Czech, Pavel Materna and the Ordered-Triple Theory). *Studia Philosophica* 62(2) (2015) 114–119
15. Chrz, T.: A system for knowledge representation based on intensional logic. *Computers and Artificial Intelligence* (3, 4) (1984) 193–209, 305–317
16. Horák, A.: *The Normal Translation Algorithm in Transparent Intensional Logic for Czech*. PhD thesis, Faculty of Informatics, Masaryk University, Brno (2002)
17. Kadlec, V.: *Syntactic analysis of natural languages based on context-free grammar backbone*. PhD thesis, Masaryk University, Faculty of Informatics (2008)
18. Horák, A.: *Computer Processing of Czech Syntax and Semantics*. *Tribun EU* (2008)
19. Jakubiček, M.: *Extrakce strukturních informací z běžného textu na základě syntaktického analyzátoru* (in Czech, Structural Information Extraction from Common Texts Based on Syntactic Parser). Master's thesis, Masaryk University, Faculty of Informatics (2008)
20. Duží, M., Jespersen, B., Materna, P.: *Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic*. Volume 17 of *Logic, Epistemology and the Unity of Science*. Springer, Berlin (2010)
21. Kovář, V., Horák, A., Jakubiček, M.: How to analyze natural language with transparent intensional logic? *RASLAN 2010, Recent Advances in Slavonic Natural Language Processing* (2010) 69–76

22. Hlaváčková, D.: Databáze slovesných valenčních rámců VerbaLex (in Czech, VerbaLex – the Database of Verb Valency Frames). PhD thesis, Masaryk University, Faculty of Arts (2008)
23. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the verbalex verb valency lexicon in the syntactic analysis of czech. In: Proceedings of Text, Speech and Dialogue 2006, Brno, Czech Republic, Springer-Verlag (2006) 79–85
24. Horák, A., Pala, K., Duží, M., Materna, P.: Verb Valency Semantic Representation for Deep Linguistic Processing. In: Proceedings of the Workshop on Deep Linguistic Processing, ACL 2007, Prague, Czech Republic, the Association for Computational Linguistics (2007) 97–104

Property Modifiers

Marie Duží and Michal Fait

VSB-Technical University Ostrava, Department of Computer Science FEL,
17. listopadu 15, 708 33 Ostrava, Czech Republic
marie.duzi@vsb.cz, michal.fait@vsb.cz

Abstract. In this paper, we deal with property modifiers defined as functions that associate a given root property P with a modified property $[MP]$. Property modifiers typically divide into four kinds, namely intersective, subsective, privative and modal. Here we do not deal with modal modifiers like alleged, which appear to be well-nigh logically lawless, because, for instance, an alleged assassin is or is not an assassin. The goal of this paper is to logically define the three remaining kinds of modifiers. Furthermore, we introduce the rule of pseudo-detachment as the rule of left subsectivity to replace the modifier M in the premise by the property M^* in the conclusion, and prove that this rule is valid for all kinds of modifiers. Furthermore, it is defined in a way that avoids paradoxes like that a small elephant is smaller than a large mouse.

Key words: Property modifier, subsective, intersective, privative, the rule of pseudo-detachment, Transparent intensional logic, TIL, intensional essentialism

1 Introduction

We introduce a logic of property modifiers modelled as a mapping from properties to properties, such that the result of the application of a modifier to a property is another property. This is because the result of modification does not depend on the state of the world, nor on time. For instance, if one applies the modifier *Skilful* to the property *Surgeon*, they obtain the property of being a skilful surgeon. The conception of modifiers presented here goes along the lines introduced in Duží et.al. [2, §4.4]. The novel contribution of this paper is a new definition of subsective and privative modifiers in terms of *intensional essentialism*.

As a starting point, here is a standard taxonomy of the three kinds of modifiers, with rigorous definitions coming afterwards. Let the extension of a property P be $|P|^1$, M standing for a modifier, M^* for the property corresponding to a

¹ Extensionalization of properties will be explained below; it corresponds to the application of a property to empirical indexes such as world and time.

modifier².

Intersective. “If a is a *round* peg, then a is round and a is a peg.”

$$\begin{aligned} M_i P(a) & \therefore M^*(a) \wedge P(a). \\ \text{Necessarily, } |M_i P| & = |M^*| \cap |P|. \end{aligned}$$

Necessarily, i.e. in all worlds and times, the set of round pegs equals to the intersection of the sets of round objects and pegs.

Note that we cannot transfer M_i from the premise to the conclusion. The reason is that a modifier cannot also occur as a predicate; these are objects of different types. Hence M^* instead of just M_i .

Subsective. “If a is a *skilful* surgeon, then a is a surgeon.”

$$\begin{aligned} M_s P(a) & \therefore P(a). \\ \text{Necessarily, } |M_s P| & \subseteq |P|. \end{aligned}$$

Necessarily, i.e. in all worlds and times, the set of skilful surgeons is a subset of the set of surgeons.

The major difference between subsective and intersective modification is that this sort of argument: $M_s P(a), Q(a) \therefore M_s Q(a)$ is *not* valid for subsective modifiers. Tilman may be a skilful surgeon, and he may be a painter too, but this does not make him a skilful painter. Scalar adjectives like ‘small’, ‘big’ or ‘skilful’ represent subsective modifiers. On the other hand, to each intersective modifier M_i there is a unique ‘absolute’ property M^* such that if a is an $M_i P$ then a is M^* not only as a P but absolutely.

Privative. “If a is a *forged* banknote, then a is not banknote.”

$$\begin{aligned} M_p P(a) & \therefore \neg P(a). \\ \text{Necessarily, } |M_p P| \cap |P| & = \emptyset. \end{aligned}$$

Necessarily, i.e. in all worlds and times, the intersection of the set of forged banknotes and banknotes is empty.

Modifiers are intersective, subsective and privative *with respect to a property* P . One and the same modifier can be intersective with respect to a property P and privative with respect to another property Q . For instance, a wooden table is wooden and is a table, but a wooden horse is not a horse. We leave aside the question whether there are modifiers privative with respect to any property. Most probably, yes, modifiers like *faked*, *forged*, *false* appear to be privative with

² The corresponding property M^* is defined below by the rule of pseudo-detachment. It is the property M (*something*), where in case of intersective modifiers M^* is an ‘absolute’ property. Hence a round peg is round not only as a peg, but absolutely. See Jespersen [6] for details.

respect to any property. Yet this issue is irrelevant to the main goal of this paper, which is to define the *rule of pseudo-detachment (PD)* and prove its validity for *any* kind of modifiers.

The rest of the paper is organised as follows. Section 2 introduces the fundamentals of our background theory TIL necessary to deal with property modifiers, which is the issue we deal with in Section 3. Here in Section 3.1 the difference between non-subsective and subsective modifiers is defined, followed by the rule of pseudo-detachment defined in Section 3.2. Concluding remarks can be found in Section 4.

2 Basic Notions of TIL

Tichý's TIL comes with *procedural semantics*, which means that we explicate meanings of language expressions as abstract procedures encoded by the expressions. Tichý defined six kinds of procedures as the so-called *constructions*³. Here we need only four of them, leaving aside Single and Double Execution.

Definition 1 (*construction*).

- (i) *Variables* x, y, \dots are *constructions* that construct objects (elements of their respective ranges) dependently on a valuation v ; they *v-construct*.
- (ii) Where X is an object whatsoever (even a *construction*), 0X is the *construction Trivialization* that constructs X without any change in X .
- (iii) Let X, Y_1, \dots, Y_n be arbitrary *constructions*. Then *Composition* $[X Y_1 \dots Y_n]$ is the following *construction*. For any v , the *Composition* $[X Y_1 \dots Y_n]$ is *v-improper* if at least one of the *constructions* X, Y_1, \dots, Y_n is *v-improper*, or if X does not *v-construct* a function that is defined at the n -tuple of objects *v-constructed* by Y_1, \dots, Y_n . If X does *v-construct* such a function then $[X Y_1 \dots Y_n]$ *v-constructs* the value of this function at the n -tuple.
- (iv) $(\lambda\text{-})$ *closure* $[\lambda x_1 \dots x_m Y]$ is the following *construction*. Let x_1, x_2, \dots, x_m be pairwise distinct variables and Y a *construction*. Then $[\lambda x_1 \dots x_m Y]$ *v-constructs* the function f that takes any members B_1, \dots, B_m of the respective ranges of the variables x_1, \dots, x_m into the object (if any) that is $v(B_1/x_1, \dots, B_m/x_m)$ -*constructed* by Y , where $v(B_1/x_1, \dots, B_m/x_m)$ is like v except for assigning B_1 to x_1, \dots, B_m to x_m .
- (v) Nothing is a *construction*, unless it so follows from (i) through (iv).

In Tichý's TIL constructions are objects *sui generis*, so that we can have constructions of constructions, constructions of functions, functions, and functional values in TIL stratified ontology. To keep track of the traffic between multiple logical strata, the ramified type hierarchy is needed. The type of first-order objects includes all non-procedural objects. Therefore, it includes not only the standard objects of individuals, truth-values, sets, etc., but also functions defined on possible worlds (i.e., the intensions germane to possible-world semantics). The type of second-order objects includes constructions of

³ See Tichý [7, Chapters 4, 5] or Duží, Jespersen & Materna [2, §1.3]

first-order objects and functions with such constructions in their domain or range. The type of third-order objects includes constructions of first- and second-order objects and functions with such constructions in their domain or range. And so on, ad infinitum. Yet, for the purposes of this paper we need just the simple theory of types. Hence, we define.

Definition 2 (simple theory of types). Let B be a *base*, where a base is a collection of pair-wise disjoint, non-empty sets. Then:

- i) Every member of B is an elementary *type of order 1 over B* .
- ii) Let $\alpha, \beta_1, \dots, \beta_m$ ($m > 0$) be types of order 1 over B . Then the collection $(\alpha \beta_1 \dots \beta_m)$ of all m -ary partial mappings from $\beta_1 \times \dots \times \beta_m$ into α is a functional *type of order 1 over B* .
- iii) Nothing is a *type of order 1 over B* unless it so follows from (i) and (ii).

For the purposes of natural-language analysis, we are assuming the following base of ground types:

- \circ : the set of truth-values $\{\mathbf{T}, \mathbf{F}\}$;
- ι : the set of individuals (the universe of discourse);
- τ : the set of real numbers (doubling as discrete times);
- ω : the set of logically possible worlds (the logical space).

We model sets and relations by their characteristic functions. Thus, for instance, $(\circ\iota)$ is the type of a set of individuals, while $(\circ\iota\iota)$ is the type of a relation-in-extension between individuals. Empirical expressions denote *empirical conditions* that may or may not be satisfied at the particular world/time pair of evaluation. We model these empirical conditions as possible-world-semantic (PWS) *intensions*. PWS intensions are entities of type $(\beta\omega)$: mappings from possible worlds to an arbitrary type β . The type β is frequently the type of the *chronology* of α -objects, i.e., a mapping of type $(\alpha\tau)$. Thus α -intensions are frequently functions of type $((\alpha\tau)\omega)$, abbreviated as ' $\alpha_{\tau\omega}$ '. Extensional entities are entities of a type α where $\alpha \neq (\beta\omega)$ for any type β . Where w ranges over ω and t over τ , the following logical form essentially characterizes the logical syntax of empirical language: $\lambda w \lambda t [\dots w \dots t \dots]$.

Examples of frequently used PWS intensions are: propositions of type $\circ_{\tau\omega}$, properties of individuals of type $(\circ\iota)_{\tau\omega}$, binary relations-in-intension between individuals of type $(\circ\iota\iota)_{\tau\omega}$, individual offices (or roles) of type $\iota_{\tau\omega}$.

Modifiers of individual properties are extensional entities of type $((\circ\iota)_{\tau\omega}(\circ\iota)_{\tau\omega})$.

Logical objects like *truth-functions* and *quantifiers* are extensional: \wedge (conjunction), \vee (disjunction) and \supset (implication) are of type $(\circ\circ\circ)$, and \neg (negation) of type $(\circ\circ)$. Quantifiers $\forall^\alpha, \exists^\alpha$ are type-theoretically polymorphic total functions of type $(\circ(\circ\alpha))$, for an arbitrary type α , defined as follows. The *universal quantifier* \forall^α is a function that associates a class A of α -elements with \mathbf{T} if A contains all elements of the type α , otherwise with \mathbf{F} . The *existential quantifier* \exists^α is a function

that associates a class A of α -elements with \mathbf{T} if A is a non-empty class, otherwise with \mathbf{F} .

Below all type indications will be provided outside the formulae in order not to clutter the notation. Moreover, the outermost brackets of the Closure will be omitted whenever no confusion can arise. Furthermore, ' X/α ' means that an object X is (a member) of type α . ' $X \rightarrow \alpha$ ' means that the construction X is typed to v -construct an object of type α , if any. Throughout, it holds that the variables $w \rightarrow \omega$ and $t \rightarrow \tau$. If $C \rightarrow \alpha_{\tau\omega}$ then the frequently used Composition $[[C w] t]$, which is the intensional descent (a.k.a. extensionalization) of the α -intension v -constructed by C , will be encoded as ' C_{wt} '. Whenever no confusion arises, we use traditional infix notation without Trivialisation for truth-functions and the identity relation, to make the terms denoting constructions easier to read. Thus, for instance, instead of ' $[^0 \wedge [^0 = [^0 + ^0 2 ^0 5] ^0 7] [^0 \supset p q]]$ ' we usually simply write ' $[[[^0 + ^0 2 ^0 5] = ^0 7] \wedge [p \supset q]]$ '.

3 Property Modifiers and Intensional Essentialism

3.1 Privative vs Subjective Modifiers

The fundamental distinction among modifiers is typically considered to be one between the *subjectives* and the *non-subjectives*. The former group consists of the *pure subjectives* and the *intersectives*. The latter group consists of the modals and the privatives. Since we are not dealing with modal modifiers here, we now want to define the distinction between *subjectives* and *privatives*. At the outset of this paper this distinction between modifiers subjective (M_s) and privative (M_p) with respect to a property P has been characterized by the rules of the right subjectivity as follows:

$$M_s P(a) \therefore P(a)$$

$$M_p P(a) \therefore \neg P(a)$$

Now we have the technical machinery at our disposal to define these modifiers in a rigorous way. To this end, we apply the logic of intensions based on the notions of *requisite* and *essence* of a property, which amounts to *intensional essentialism*⁴. The idea is this. Every property has a host of other properties necessarily associated with it. For instance, the property of being a bachelor is associated with the properties of being a man, being unmarried, and many others. Necessarily, if a happens to be a bachelor then a is a man and a is unmarried. We call these adjacent properties *requisites* of a given property.

The requisite relations *Req* are a family of relations-in-extension between two intensions, so they are of the polymorphous type $(\alpha_{\tau\omega} \beta_{\tau\omega})$, where possibly α

⁴ In contrast to *individual anti-essentialism*: no individual has a non-trivial empirical property necessarily. In other words, only trivial properties like being self-identical, being identical to a or b , etc., are necessarily ascribed to an individual a . For details see Duží et al. [2, §4.2], and also Cmorej [1].

$= \beta$. Infinitely many combinations of *Req* are possible, but for our purpose we will need the following one: $Req/(o(o\iota)_{\tau\omega}(o\iota)_{\tau\omega})$; a property of individuals is a requisite of another such property. IIL embraces *partial functions*⁵. Partiality gives rise to the following complication. The requisite relation obtains analytically necessarily, i.e., for all worlds w and times t , and so the values at the $\langle w,t \rangle$ -pairs of particular intensions are irrelevant. But the values of properties are isomorphic to characteristic functions, and these functions are amenable to truth-value gaps. For instance, the property of having stopped smoking comes with a bulk of requisites like, e.g., the property of being a former smoker. If a never smoked, then the proposition that a stopped smoking comes with a truth-value gap, because it can be neither true nor false that a stopped or did not stop smoking. Thus, the predication of such a property P of a may also fail, causing $[{}^0P_{wt} {}^0a]$ to be v -improper. There is a straightforward remedy, however, namely the propositional property of being true at $\langle w,t \rangle$: $True/(oo_{\tau\omega})_{\tau\omega}$. Given a proposition v -constructed by X , $[{}^0True_{wt} X]$ v -constructs T if the proposition presented by X is true at $\langle w,t \rangle$; otherwise (i.e., if the proposition constructed by X is false or else undefined at $\langle w,t \rangle$) F . Thus we define:

Definition 3 (requisite relation between ι -properties). Let P, Q be constructions of individual properties; $P, Q \rightarrow (o\iota)_{\tau\omega}$; $x \rightarrow \iota$. Then

$$[{}^0Req Q P] = \forall w \forall t [\forall x [[{}^0True_{wt} \lambda w \lambda t [P_{wt} x]] \supset [{}^0True_{wt} \lambda w \lambda t [Q_{wt} x]]]].$$

Next, we are going to define the essence of a property. Our essentialism is based on the idea that since no purely contingent property can be essential of any individual, essences are borne by intensions rather than by individuals exemplifying intensions⁶. Hence, our essentialism is based on the requisite relation, couching essentialism in terms of a priori interplay between properties, regardless of who or what exemplifies a given property. *Intensional essentialism* is technically an algebra of individually necessary and jointly sufficient conditions for having a certain property (or other sort of intension). The $\langle w,t \rangle$ -relative extensions of a given property are irrelevant, as we said.

Definition 4 (essence of a property). Let $p, q \rightarrow (o\iota)_{\tau\omega}$ be constructions of individual properties, and let $Ess/(o(o\iota)_{\tau\omega})(o\iota)_{\tau\omega}$, i.e. a function assigning to a given property p the set of its requisites defined as follows:

$${}^0Ess = \lambda p \lambda q [{}^0Req q p]$$

Then the essence of a property p is the set of its requisites: $[{}^0Ess p] = \lambda q [{}^0Req q p]$

⁵ See Duží et al. [2, 276-78] for philosophical justification of partiality despite the associated technical complications.

⁶ By ‘purely contingent intension’ we mean an intension that is not a constant function and does not have an essential core (e.g. the property of having exactly as many inhabitants as Prague is necessarily exemplified by Prague).

Each property has (possibly infinitely) many requisites. The question is, how do we know which are the requisites of a given property? The answer requires an *analytic definition* of the given property, which amounts to the specification of its essence. For instance, consider the property of being a bachelor. If we define this property as the property of being an unmarried man, then the properties of being unmarried and being a man are among the requisites of the property of being a bachelor. Thus, the sentence “bachelors are unmarried men” comes out analytically true:

$$\forall w \forall t [\forall x [[{}^0\text{Bachelor}_{wt} x] \supset [[{}^0\text{Unmarried } {}^0\text{Man}]_{wt} x]]].$$

And since the modifier *Unmarried* is intersective, it also follows that necessarily, each bachelor is unmarried and is a man:

$$\forall w \forall t [\forall x [[{}^0\text{Bachelor}_{wt} x] \supset [[{}^0\text{Unmarried}'_{wt} x] \wedge [{}^0\text{Man}_{wt} x]]]].$$

Note, however, that *Unmarried'* / $(\text{o}\iota)_{\tau\omega}$ and *Unmarried* / $((\text{o}\iota)_{\tau\omega}(\text{o}\iota)_{\tau\omega})$ are entities of different types. The former is a property of individuals uniquely assigned to the latter, which is an intersective modifier.

With these definitions in place, we can go on to compare two kinds of *subsectives* against *privatives*⁷. Since these modifiers change the essence of the root property, we need to compare the essences, that is sets of properties, of the root and modified property. To this end, we apply the set-theoretical relations of being a *subset* and a *proper subset* between sets of properties, and the *intersection* operation on sets of properties, defined as follows.

Let $\pi = (\text{o}\iota)_{\tau\omega}$, for short, $\subseteq, \subset / (\text{o}(\text{o}\pi)(\text{o}\pi))$, and let $a, b \rightarrow_v (\text{o}\pi); x \rightarrow_v \pi$. Then

$$\begin{aligned} {}^0\subseteq &= \lambda ab [{}^0\forall \lambda x [a x] \supset [b x]] \\ {}^0\subset &= \lambda ab [[{}^0\forall \lambda x [[a x] \supset [b x]]] \wedge \neg [a = b]] \end{aligned}$$

Furthermore, the *intersection* function $\cap / ((\text{o}\pi)(\text{o}\pi)(\text{o}\pi))$ is defined on sets of properties in the usual way: ${}^0\cap = \lambda ab \lambda x [[a x] \wedge [b x]]$. In what follows we will use classical (infix) set-theoretical notation for any sets A, B ; hence instead of ' ${}^0\subseteq A B$ ' we will write ' $[A \subseteq B]$ ', and instead of ' ${}^0\cap A B$ ' we will write ' $[A \cap B]$ '.

Definition 5 (subsective vs. privative modifiers).

- A modifier M is *subsective* with respect to a property P iff

$$[{}^0\text{Ess } P] \subseteq [{}^0\text{Ess } [M P]]$$

- A modifier M is *non-trivially subsective* with respect to a property P iff

$$[{}^0\text{Ess } P] \subset [{}^0\text{Ess } [M P]]$$

⁷ Since *intersective* modification is a special kind of *subsective* modification, we are disregarding *intersectives* not to clutter the exposition. Intersectives are controlled by the same rule of *right subsectivity* that applies to the subsectives.

- A modifier M is *privative* with respect to a property P iff

$$\begin{aligned} & [[{}^0\text{Ess } P] \cap [{}^0\text{Ess } [MP]]] \neq \emptyset \wedge \\ & {}^0\exists\lambda p [[{}^0\text{Ess } P] p] \wedge [[{}^0\text{Ess } [MP]] \lambda\omega\lambda t [\lambda x \neg [p_{\omega t} x]]] \end{aligned}$$

Remark. We distinguish between *subsective* and *non-trivially subsective* modifiers, because among subsectives there are also trivial subsectives. A modifier M is *trivially subsective* with respect to P iff the modified property $[MP]$ has exactly the same essence as the property P . These modifiers are trivial in that the modification has no effect on the modified property and so might just as well not have taken place. For instance, there is no semantic or logical (but perhaps rhetorical) difference between the property of being a leather and the property of being a *genuine* leather. Trivial modifiers such as *genuine*, *real*, *actual* are pure subsectives: genuine leather things are not located in the intersection of leather things and objects that are genuine, for there is no such property as being genuine, pure and simple⁸.

Example. The modifier *Wooden* / $((\text{oi})_{\tau\omega}(\text{oi})_{\tau\omega})$ is subsective with respect to the property of being a table, *Table* / $(\text{oi})_{\tau\omega}$, but privative with respect to the property of being a horse, *Horse* / $(\text{oi})_{\tau\omega}$. Of course, a wooden table is a table, but the essence of the property $[{}^0\text{Wooden } {}^0\text{Table}]$ is enriched by the property of being wooden. This property is a requisite of the property of being a wooden table, but it is not a requisite of the property of being a table, because tables can be instead made of stone, iron, etc.

$$[{}^0\text{Ess } {}^0\text{Table}] \subset [{}^0\text{Ess } [{}^0\text{Wooden } {}^0\text{Table}]].$$

But a wooden horse is not a horse. The modifier *Wooden*, the same modifier that just modified *Table*, deprives the essence of the property of being a horse, *Horse* / $(\text{oi})_{\tau\omega}$, of many requisites, for instance, of the property of being an animal, having a bloodstream, a heartbeat, etc. Thus, among the requisites of the property $[{}^0\text{Wooden } {}^0\text{Horse}]$ there are properties like *not being a living thing*, *not having a bloodstream*, etc., which are contradictory (not just contrary) to some of the requisites of the property *Horse*. On the other hand, the property $[{}^0\text{Wooden } {}^0\text{Horse}]$ shares many requisites with the property of being a horse, like the outline of the body, having four legs, etc., and has an additional requisite of being made of wood. We have:

$$\begin{aligned} & [[{}^0\text{Ess } {}^0\text{Horse}] \cap [{}^0\text{Ess } [{}^0\text{Wooden } {}^0\text{Horse}]]] \neq \emptyset \wedge \\ & \quad [[{}^0\text{Ess } {}^0\text{Horse}] {}^0\text{Living_thing}] \wedge \\ & [[{}^0\text{Ess } [{}^0\text{Wooden } {}^0\text{Horse}]] \lambda\omega\lambda t [\lambda x \neg [{}^0\text{Living_thing } x]]] \wedge \\ & \quad [[{}^0\text{Ess } {}^0\text{Horse}] {}^0\text{Blood}] \wedge \\ & [[{}^0\text{Ess } [{}^0\text{Wooden } {}^0\text{Horse}]] \lambda\omega\lambda t [\lambda x \neg [{}^0\text{Blood } x]]] \wedge \\ & \quad \text{etc.} \end{aligned}$$

A modifier M is *privative* with respect to a property P iff the modified property $[MP]$ lacks at least one, *but not all*, of the requisites of the property P .

⁸ Iwańska [5, 350] refers to ‘ideal’, ‘real’, ‘true’, and ‘perfect’ as *type-reinforcing* adjectives, which seems to get the pragmatics right of what are semantically pleonastic adjectives.

However, in this case we cannot say that the essence of the property $[MP]$ is a proper subset of the essence of the property P , because the modified property $[MP]$ has at least one other requisite that does not belong to the essence of P , because it contradicts to some of the requisites of P . Hence, M is privative with respect to property P iff the essence of property $[MP]$ has a non-empty intersection with the essence of the property P , and this intersection is a *proper* subset of both the essences of P and of $[MP]$. For instance, a forged banknote has *almost* the same requisites as does a banknote, but it has also another requisite, namely the property of being forged with respect to the property of being a banknote.

As a result, if M_p is privative with respect to the property P , then the modified property $[M_pP]$ and the property P are contrary rather than contradictory properties:

$$\forall w \forall t \forall x [[M_pP]_{wt} x \supset \neg[P_{wt} x]] \wedge \exists w \exists t \exists x [\neg[[M_pP]_{wt} x] \wedge \neg[P_{wt} x]]$$

It is not possible for x to co-instantiate $[M_pP]$ and P , and possibly x instantiates neither $[M_pP]$, nor P .

3.2 The Rule of Pseudo-detachment

The issue we are going to deal with now is left subsectivity⁹. We have seen that the principle of left subsectivity is trivially (by definition) valid for intersective modifiers. If Jumbo is a yellow elephant, then Jumbo is yellow. Yet how about the other modifiers? If Jumbo is a small elephant, is Jumbo small? If you factor out *small* from *small elephant*, the conclusion says that Jumbo is small, period. Yet this would seem a strange thing to say, for something appears to be missing: Jumbo is a small *what*? Nothing or nobody can be said to be small or forged, skilful, temporary, larger than, the best, good, notorious, or whatnot, without any sort of qualification. A complement providing some sort of qualification to provide an answer to the question, 'a ... *what*?' is required. We are going to introduce now the rule of pseudo-detachment that is valid for all kinds of modifiers including subsective and privative ones. The idea is simple. From a is an MP we infer that a is an M -with respect to something.

For instance, if the customs officers seize a forged banknote and a forged passport, they may want to lump together all the forged things they have seized that day, abstracting from the particular nature of the forged objects. This lumping together is feasible only if it is logically possible to, as it were, abstract *forged* from a being a forged A and b being a forged B to form the new predications that a is forged (something) and that b is forged (something), which are subsequently telescoped into a conjunction.

Gamut (the Dutch equivalent of Bourbaki) claims that if Jumbo is a small elephant, then it does not follow that Jumbo is small [3, §6.3.11]. We are going to show that the conclusion does follow. The rule of pseudo-detachment (*PD*)

⁹ In this section, we partly draw on material from Duží et.al. [2, §4.4].

validates a certain inference schema, which on first approximation is formalized as follows:

$$(PD) \quad \frac{a \text{ is an } MP}{a \text{ is an } M^*}$$

where ‘ a ’ names an appropriate subject of predication while ‘ M ’ is an adjective and ‘ P ’ a noun phrase compatible with a .

The reason why we need the rule of pseudo-detachment is that M as it occurs in MP is a *modifier* and, therefore, cannot be transferred to the conclusion to figure as a *property*. So no actual detachment of M from MP is possible, and Gamut is insofar right. But (PD) makes it possible to replace the modifier M by the property M^* compatible with a to obtain the conclusion that a is an M^* . (PD) introduces a new property M^* ‘from the outside’ rather than by obtaining M ‘from the inside’, by extracting a part from a compound already introduced. The temporary rule above is incomplete as it stands; here is the full pseudo-detachment rule, SI being substitution of identicals (Leibniz’s Law)¹⁰, EG existential generalization.

- | | | |
|-----|--|------------|
| (1) | a is an MP | assumption |
| (2) | a is an (M something) | 1, EG |
| (3) | M^* is the property (M something) | definition |
| (4) | a is an M^* | 2,3, SI |

To put the rule on more solid grounds of TIL, let $\pi = (\text{oi})_{\tau\omega}$ for short, $M \rightarrow (\pi\pi)$ be a modifier, $P \rightarrow \pi$ an individual property, $[MP] \rightarrow \pi$ the property resulting from applying M to P , and let $[MP]_{wt} \rightarrow_v (\text{oi})$ be the result of extensionalizing the property $[MP]$ with respect to a world w and time t to obtain a set, in the form of a characteristic function, applicable to an individual $a \rightarrow \iota$. Further, let $= / (\text{o}\pi\pi)$ be the identity relation between properties, and let $p \rightarrow_v \pi$ range over properties, $x \rightarrow_v \iota$ over individuals. Then the *proof* of the rule is this:

- | | | |
|----|--|-------------------------|
| 1. | $[[MP]_{wt} a]$ | assumption |
| 2. | $\exists p [[Mp]_{wt} a]$ | 1, \exists I |
| 3. | $[\lambda x \exists p [[Mp]_{wt} x] a]$ | 2, λ -expansion |
| 4. | $[\lambda w' \lambda t' [\lambda x \exists p [[Mp]_{w't'} x]]_{wt} a]$ | 3, λ -expansion |
| 5. | $M^* = \lambda w' \lambda t' [\lambda x \exists p [[Mp]_{w't'} x]]$ | definition |
| 5. | $[M^*_{wt} a]$ | 4, 5, SI |

Any valuation of the free occurrences of the variables w, t that makes the first premise true will also make the second, third and fourth steps true. The fifth premise is introduced as valid by definition. Hence, any valuation of w, t that makes the first premise true will, together with the step five, make the conclusion true.

(PD), dressed up in full TIL notation, is this¹¹:

¹⁰ More precisely, substitution of identical properties.

¹¹ As mentioned above, in case of the modifier M being intersective, the property M^* is unique for any p . For details see Jespersen [6].

$$(PD) \quad \frac{[[MP]_{wt} a] \quad [M^* = \lambda w' \lambda t' [\lambda x \exists p [[Mp]_{w't'} x]]]}{[M^*_{wt} a]}$$

Additional type: $\exists / (\circ(\circ\pi))$.

Here is an instance of the rule.

- (1) a is forged banknote
- (2) forged* is the property of being a forged something
- (3) a is forged*.

The schema extends to all (appropriately typed) objects. For instance, let the inference be, “Geocaching is an exciting hobby; therefore, geocaching is exciting”. Then a is of type π , $P \rightarrow (\circ\pi)_{\tau\omega}$, $M \rightarrow ((\circ\pi)_{\tau\omega} (\circ\pi)_{\tau\omega})$, and $M^* \rightarrow (\circ\pi)_{\tau\omega}$.

Now it is easy to show why this argument must be valid:

$$\frac{\text{John has a forged banknote and a forged passport}}{\text{John has two forged things.}}$$

$$\frac{\lambda w \lambda t \exists x y [{}^0\text{Have}_{wt} {}^0\text{John } x] \wedge [{}^0\text{Have}_{wt} {}^0\text{John } y] \wedge [{}^0\text{Forged } {}^0\text{Banknote}]_{wt} x \wedge [{}^0\text{Forged } {}^0\text{Passport}]_{wt} y \wedge [{}^0 \neq x y]}{\lambda w \lambda t \exists x y [{}^0\text{Have}_{wt} {}^0\text{John } x] \wedge [{}^0\text{Have}_{wt} {}^0\text{John } y] \wedge [{}^0\text{Forged}^*_{wt} x] \wedge [{}^0\text{Forged}^*_{wt} y] \wedge [{}^0 \neq x y]} \quad \frac{\lambda w \lambda t [{}^0\text{Number_of } \lambda z [[{}^0\text{Have}_{wt} {}^0\text{John } z] \wedge [{}^0\text{Forged}^*_{wt} z]] = {}^0 2]}{\lambda w \lambda t [{}^0\text{Number_of } \lambda z [[{}^0\text{Have}_{wt} {}^0\text{John } z] \wedge [{}^0\text{Forged}^*_{wt} z]] = {}^0 2]}$$

Types: $\text{Number_of} / (\tau(\circ\iota))$; Banknote , Passport , Forged^* / π ; $\text{Have} / (\circ\iota)_{\tau\omega}$; $\text{Forged} / (\pi\pi)$.

There are three conceivable objections to the validity of (PD) that we are going to deal with now.

First objection. If Jumbo is a small elephant and if Jumbo is a big mammal, then Jumbo is not a small mammal; hence Jumbo is small and Jumbo is not small. Contradiction!

The contradiction is only apparent, however. To show that there is no contradiction, we apply (PD):

$$\frac{\lambda w \lambda t [[{}^0\text{Small } {}^0\text{Elephant}]_{wt} {}^0\text{Jumbo}]}{\lambda w \lambda t \exists p [[{}^0\text{Small } p]_{wt} {}^0\text{Jumbo}]}$$

$$\frac{\lambda w \lambda t [[{}^0\text{Big } {}^0\text{Mammal}]_{wt} {}^0\text{Jumbo}]}{\lambda w \lambda t \exists q [[{}^0\text{Big } q]_{wt} {}^0\text{Jumbo}].}$$

Types: *Small, Big* / $(\pi\pi)$; *Mammal, Elephant* / π ; *Jumbo* / ι ; $p, q \rightarrow \pi$.

To obtain a contradiction, we would need an additional premise; namely, that, necessarily, any individual that is big (i.e., a big something) is not small (the *same* something). Symbolically,

$$\forall w \forall t \forall x \forall p \ [[[{}^0\text{Big } p]_{wt} x] \supset \neg [[{}^0\text{Small } p]_{wt} x]].$$

Applying this fact to Jumbo, we have:

$$\forall w \forall t \forall p \ [[[{}^0\text{Big } p]_{wt} {}^0\text{Jumbo}] \supset \neg [[{}^0\text{Small } p]_{wt} {}^0\text{Jumbo}]].$$

This construction is equivalent to

$$\forall w \forall t \neg \exists p \ [[[{}^0\text{Big } p]_{wt} {}^0\text{Jumbo}] \wedge [[{}^0\text{Small } p]_{wt} {}^0\text{Jumbo}]].$$

But the only conclusion we can draw from the above premises is that Jumbo is a small something and a big something else:

$$\lambda w \lambda t [\exists p [[{}^0\text{Small } p]_{wt} {}^0\text{Jumbo}] \wedge \exists q [[{}^0\text{Big } q]_{wt} {}^0\text{Jumbo}]].$$

Hence, no contradiction.

Nobody and nothing is absolutely small or absolutely large, because everybody is made small by something and made large by something else. Similarly, nobody is absolutely good or absolutely bad, everybody has something they do well and something they do poorly. That is, everybody is both good and bad, which here just means being good at something and being bad at something else, without generating paradox.

But nobody can be good at something and bad *at the same thing* simultaneously (*Good, Bad* / $(\pi\pi)$):

$$\forall w \forall t \forall x \neg \exists p \ [[[{}^0\text{Good } p]_{wt} x] \wedge [[{}^0\text{Bad } p]_{wt} x]].$$

Second objection. The use of pseudo-detachment, together with an innocuous-sounding premise, makes the following argument valid.

Jumbo is a small elephant \wedge Mickey is a big mouse

Jumbo is small \wedge Mickey is big.

If x is big and y is small, then x is bigger than y

Mickey is bigger than Jumbo.

Yet it is not so. We can only infer the necessary truth that if x is a small something and y is a big object *of the same kind*, then y is a bigger object of that kind than x :

$$\forall w \forall t \forall x \forall y \forall p \ [[[{}^0\text{Small } p]_{wt} x] \wedge [[{}^0\text{Big } p]_{wt} y]] \supset [{}^0\text{Bigger}_{wt} y x]].$$

Additional type: *Bigger* / $(ou)_{\tau\omega}$. This cannot be used to generate a contradiction from these constructions as premises, because $p \neq q$:

$$\exists p [[{}^0Small\ p]_{wt}\ a]; \exists q [[{}^0Big\ q]_{wt}\ b]$$

Geach, in [4], launches a similar argument to argue against a rule of inference that is in effect identical to *(PD)*. He claims that that rule would license an invalid argument. And indeed, the following argument *is* invalid:

a is a big flea, so a is a flea and a is big; b is a small elephant, so b is an elephant and b is small; so a is a big animal and b is a small animal. (Ibid., p. 33.)

But pseudo-detachment licenses no such argument. Geach's illegitimate move is to steal the property *being an animal* into the conclusion, thereby making *a* and *b* commensurate. Yes, both fleas and elephants are animals, but *a*'s being big and *b*'s being small follow from *a*'s being a flea and *b*'s being an elephant, so pseudo-detachment only licenses the following two inferences, $p \neq q$:

$$\exists p [[{}^0Big\ p]_{wt}\ a]; \exists q [[{}^0Small\ q]_{wt}\ b]$$

And a big *p* may well be smaller than a small *q*, depending on the values assigned to *p*, *q*.

Third objection. If we do not hesitate to use 'small' not only as a modifier but also as a predicate, then it would seem we could not possibly block the following fallacy:

$$\begin{array}{c} \text{Jumbo is small} \\ \text{Jumbo is an elephant} \\ \hline \text{Jumbo is a small elephant.} \\ \\ \lambda\omega\lambda t \exists p [[{}^0Small\ p]_{wt}\ {}^0Jumbo] \\ \lambda\omega\lambda t [{}^0Elephant_{wt}\ {}^0Jumbo] \\ \hline \lambda\omega\lambda t \exists p [[{}^0Small\ {}^0Elephant]_{wt}\ {}^0Jumbo] \end{array}$$

But we can block it, since this argument is obviously not valid. The premises do not guarantee that the property *p* with respect to which Jumbo is small is identical to the property *Elephant*. As was already pointed out, one cannot start out with a *premise* that says that Jumbo is small (is a small something) and conclude that Jumbo is a small *B*.

4 Conclusion

In this paper, we applied TIL as a logic of intensions to deal with property modifiers and properties in terms of intensional essentialism. Employing the essences of properties, we defined the distinction between non-subsective (that

is privative) and subsective modifiers. While the former ones deprive the root property of some but not all of its requisites, the latter enrich the essence of the root property. The main result is the rule of pseudo-detachment together with the proof of its validity for any kind of modifiers.

Acknowledgments. This research has been supported by the Grant Agency of the Czech Republic, project No. GA15-13277S, *Hyperintensional logic for natural language analysis*, and by the internal grant agency of VSB_TU Ostrava, project SGS No. SP2017/133, *Knowledge modeling and its applications in software engineering III*.

References

1. Cmorej, P. (1996): Empirické esenciálne vlastnosti (Empirical essential properties). *Organon F*, vol. 3, pp. 239-261.
2. Duží, M., Jespersen, B., Materna, P. (2010): *Procedural Semantics for Hyperintensional Logic*; Foundations and Applications of Transparent Intensional Logic. Dordrecht: Springer.
3. Gamut, L.T.F. (1991): *Logic, Language and Meaning*, vol. II. Chicago, London: The University of Chicago Press.
4. Geach, P.T. (1956): Good and evil. *Analysis*, vol. 17, pp. 33-42.
5. Iwańska, Ł (1997): Reasoning with intensional negative adjectival: semantics, pragmatics, and context. *Computational Intelligence*, vol. 13, pp. 348-90.
6. Jespersen, B. (2016): Left subsectivity: how to infer that a round peg is round. *Dialectica*, vol. 70, issue 4, pp. 531-547.
7. Tichý, P. (1988): *The Foundations of Frege's Logic*. De Gruyter.

Multilinguality Adaptations of Natural Language Logical Analyzer

Marek Medved', Terézia Šulganová, and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xmedved1, hales}@fi.muni.cz
445246@mail.muni.cz

Abstract. The AST (automated semantic analysis) system serves as a final pipeline component in translating natural language sentences to formulae of higher-order logic formalism, the Transparent Intensional Logic (TIL). TIL was designed as a full expressive tool capable of representing complex meaning relations of natural language expressions. AST was designed as a language independent tool, which was originally developed for the Czech language.

In this paper, we summarize the latest development of AST aiming at easy transfer of the underlying lexicons and rules to other languages. The changes are test with the English language selected as a representative of a different language family than Czech having general multilingual applicability of the process in mind.

Key words: Transparent Intensional Logic, TIL, logical analysis, natural language semantics

1 Introduction

Capturing full logical representation of natural language expressions is not a trivial task. In our work, we lean on a high-ordered logical formalism the Transparent Intensional Logic (TIL) [1,2] that was originally designed to cover all logical phenomena present in natural language.

From the theoretical point of view, a semantic representation of one expression through multiple languages should be (structurally) very similar. In the following text, we are presenting the details of new developments of a system for TIL semantic analysis called AST (automated semantic analysis). AST was designed as language independent tool that from a syntactic tree sentence representation can create its logical representation.

To be able to prepare input to the AST processor from standard plain text sentences, we use the SET [3] parser that is also designed for multilingual processing. SET is based on a grammar of pattern-matching dependency rules that can be adapted for any new language.

<i>English Pen TreeBank tags</i>			<i>Czech attributive tags</i>		
<i>word</i>	<i>lemma</i>	<i>tag</i>	<i>word</i>	<i>lemma</i>	<i>tag</i>
Some	some	DT	Some	some	k3
agents	agent	NNS	agents	agent	k1gInP
are	be	VBP	are	be	k5mInP
mobile	mobile	JJ	mobile	mobile	k2gId1
,	,	,	,	,	kIx,
other	other	JJ	other	other	k2gId1
agents	agent	NNS	agents	agent	k1gInP
are	be	VBP	are	be	k5mInP
static	static	JJ	static	static	k2gId1
.	.	SENT	.	.	kIx.

Fig. 1. Tag translation example

In the following text, we describe the SET and AST modifications that allow flexible multilingual setup of the whole pipeline. The resulting translation of natural language expressions to logical formulae are currently tested with English, the changes are however general enough for a transfer to another languages.

2 Tagset Translation

The original implementation of TIL analysis [4] leans on morphological aspects of phrase agreement rules based on the Czech attributive tagset [5] that carry lot of information about the grammatical case, number, gender, person etc. The AST tool is based on the same principles of grammatical agreement test with the exploitation of similar set of attributes that allow to drive the analysis decisions to provide correct TIL constructions.

For example when the system is building a logical construction of a single clause and no acceptable subject was found among the sentence constituents, the system supplies an inexplicit subject formed with a personal pronoun. The actual pronoun specification then follows the subject-predicate agreement rules and identifies the pronoun number, gender and person from the form of the main verb.

Within the multilingual setup of AST, we have decided to keep the attributive morphological specification as a form of a “general pivot” which allows us to process another language (English) in the same way as Czech in many rules. We have supplemented the system with a tag translation module, in the test case from the English Penn TreeBank tagset [6] to the Czech equivalent that also contains additional information about gender, case etc.

Each original Penn TreeBank tag is translated into its Czech equivalent according to multiple rules (for an example tag translation see Figure 1). The additional information is based on definitions related to selected part-of-speech

```

TMPL: verbfin ... comma $CONJ ... verbfin ... rbound
MARK 2 5 7 <clause> HEAD 3 DEP 0 PROB 50
LABEL vrule_sch()
LABELDEP vrule_sch_add("#1H 2H")
$CONJ(tag): k3.*yR k3.*xQ k8.*xS

```

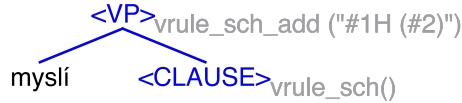


Fig. 2. LABELDEP in a relative clause

categories and in several cases to particular words by using the following list of rules:

- pronouns:
 - masculine gender (gM) for pronouns: *he, his, himself* (similarly for feminine and neuter genders),
 - personal pronoun (xP) for words with tag PP,
 - possessive pronoun (x0) for word with CDZ, PPZ;
- conjunctions:
 - coordinative type for *and, but, for, nor, or, so, yet*,
 - subordinate type otherwise.

3 SET Modifications

AST processes the input syntactic trees based on rule labels specified with a set of grammar rule labels. In case of phrasal trees, the labels can refer to specific constituents within one rule which can relate several right-hand-side terms in one action. In case of dependency parser, we need to introduce several new technical modifications to address complex constructions like inserted clauses and coordinations.

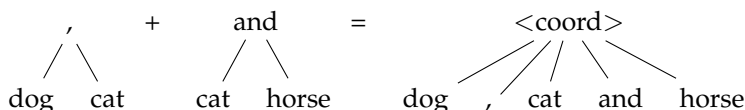
3.1 Action LABELDEP

Within the hybrid tree approach, selected rules in SET can create several dependency edges in one step – e.g. one rule creates a coordination node with 3 children, and at the same time it attaches this coordination node to a governor node. In this situation, AST needs two different label schemata for these two types of attachments to create a correct construction. Therefore, apart from the LABEL action that contains the TIL schemata relevant to the rule [7], we have added a new action called LABELDEP, where the TIL schema for the higher-level dependency is stored. Using this new construction, all nodes in the tree can be assigned a correct schema.

The same solution is applied in case of rules for relative clauses (see Figure 2) – they often recognize a relative sentence and attach it to its governor at the same time. LABELDEP action is used here as well.

3.2 Structured Clauses and Coordinations

The native SET algorithm joins and flattens structures where their inner structure is lacking or debatable – e.g. coordinations connecting 3 and more members (like “dog, cat and horse”) are successively built as 2 or more simple coordinations (e.g. “dog, cat” and “cat and horse”) and the parsing algorithm joins them into one “flat” coordination – all the tokens are appended under a single coordination node:



This is not suitable for the AST algorithm, because each TIL schema needs to know how many parameters are coming and the number is usually not variable. Therefore, we have decided not to flatten the coordinations for AST and rather nest them under each other – in the above example, the necessary form is thus *coord(dog, coord(cat and horse))*. This way, the schemata can process only 2 items at a time, and combine them together at the original parent arching node as depicted in Figure 3.

3.3 Action LABELTOP

SET constructs the top level of the tree automatically, without a reflection among the grammar rules – all the nodes (e.g. *clauses* in a complex sentence) that were not attached by rules anywhere, are attached to the root “sentence” node at the end of parsing. However, since AST needs to define clause coordinations at this level, the top level label has to be specified separately. A new keyword LABELTOP has thus been introduced, which specifies the rule schema of the root node (see Figure 4).

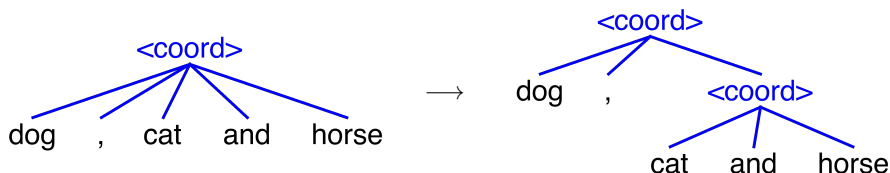


Fig. 3. Coordination nodes splitting in SET for AST processing.

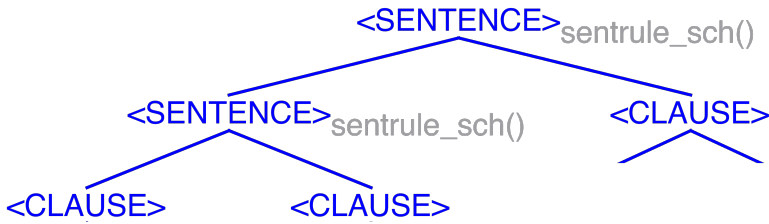


Fig. 4. "LABELTOP sentrule_sch()" label on the sentence level.

3.4 Coordination Nodes Morphology

AST works intensively with morphological tags at the tree nodes. Tree nodes at the SET output are assigned correct morphological tags as they were propagated from the bottom of the tree. However, in case of coordinations of constituents in singular number, the resulting coordination can express both a singular or a plural number according to the context:

[A skier or a climber]_{plural} are risking their lives.

Each club member is [a skier or a climber.]_{singular}

Not taking this into account introduced errors in cases where this information was used to check morphological agreements in AST, e.g. between subject and predicate in Czech.

AST can now handle the dual number situation with a procedure that assigns both the correct tags to the coordination nodes (see Figure 5). In Czech, a correct gender of the plural case is also handled in accordance with standard grammar rules (masculine animate has precedence before other genders). This enables AST to build correct constructions for sentences with heterogeneous coordinations.

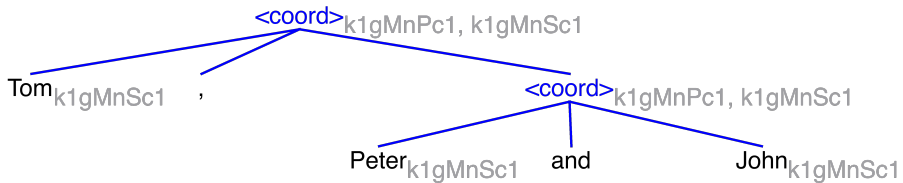


Fig. 5. Multiple grammatical numbers in coordination nodes.

4 AST Modifications

AST as language independent tool can be modified to analyze new language by supplying several language dependent lexicons: lexical item types, verb valency lexicon, prepositional types and sentence schema lexicon. These files were presented in previous publications [7], we are offering only examples of settings for other than the primary language here. Examples of the resulting constructions are presented in Figure 6.

4.1 TIL Types of Lexical Items

TIL lexicon of the bottom level types consists of a list of lexical item specifications with TIL types for each word in the input tree. An example of the types for the verb “*stop*” is the following:

```
stop
/k5/otriv ((o(ooτω)(ooτω))ω)
/k5/otriv (((o(ooτω)(ooτω))ω)ι)
```

This specification states that “*stop*” is an episodic verb [8,9] with no object (when someone just stops) and with one object (when someone stops somebody else).

4.2 Verb Valency Lexicon

For each clause the system identifies the sentence valency frame and triggers a process that according schema matching the extracted valency frame specifies how the sub-constructions are put together. An example for verb “*stop*” is:

```
stop
hPTc4 :exists:V(v):V(v):and:V(v)=[[#0,try(#1)],V(w)]
hPTc2r{at} :exists:V(v):V(v):and:V(v) subset [#0,V(w)] and [#1,V(v)]
```

The two schemata correspond to the situations when *somebody* or *something* stops somebody or something else, or when the subject *stops at something*, in which case the second constituent provides a modifier of the verbal object.

4.3 Prepositional Type Lexicon

If the AST tool decides how a prepositional phrase participates on the analyzed valency frame, the actual preposition is the central distinctive feature. An example of a schema for transformation of a prepositional phrase with the preposition “*in*” is:

```
in
0 hL hW
```

The preposition “*in*” can introduce a locational prepositional phrase hL (*where*), or a temporal *when/until what time* specification hW. The first number can denote a grammatical case of the encompassed noun phrase, however, for English this is left unused.

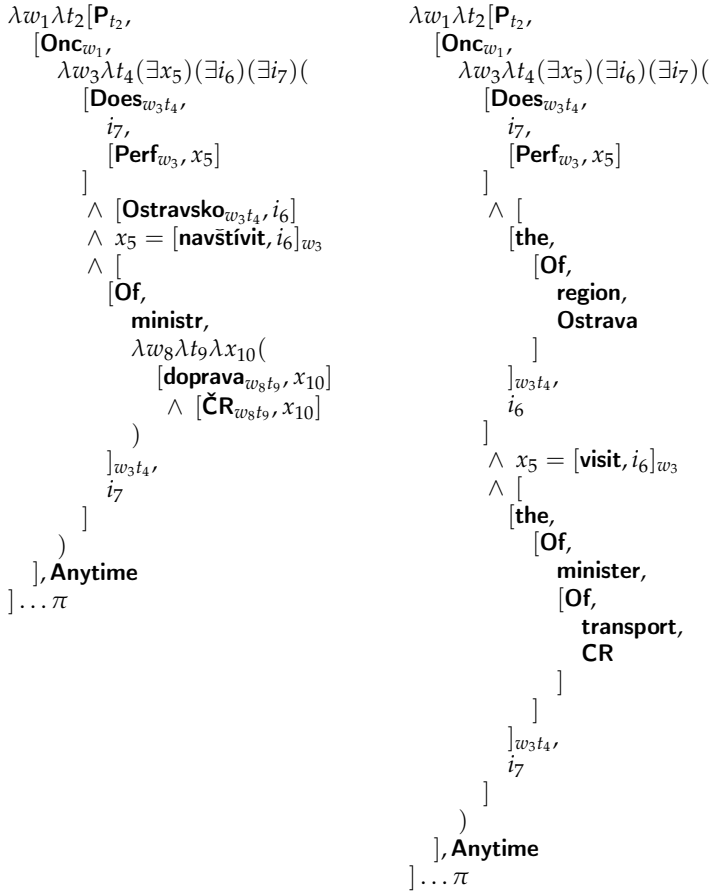


Fig. 6. Example constructions translated from the same sentence in Czech and English: *Ministr dopravy ČR navštívil Ostravsko* and *The minister of transport of CR visited the region of Ostrava*.

4.4 Sentence Schema Lexicon

The sentence schema lexicon drives the creation of the top level sentence logical construction. The schema takes as argument the sub-constructions of the two clauses that are in a subordination or a coordination structure. An example for the subordinate conjunction “when” is:

("when";","): "lwt(tense_temp(awt(#2),awt(#1)))"

This schema specifies that the two sentences have to be combined by applying the subordinate temporal clause as a time interval specification of the main clause.

5 Conclusion

In this paper, we have detailed the latest developments of the pipeline used for logical analysis of natural language sentences by means of the Transparent Intensional Logic. The presented changes aimed at promoting multilingual setup of the system with Czech as a representative of a morphologically rich language providing the attributive basis for phrase agreement specifications and English used as the first tested transfer language.

In the future work, we plan to test the setup with more languages and offer wider scale comparisons of logical structure sharing between different language environments.

Acknowledgements. This work has been partly supported by the Czech Science Foundation under the project GA15-13277S.

References

1. Tichy, P.: The foundations of Frege's logic. Walter de Gruyter (1988)
2. Duží, M., Jespersen, B., Materna, P.: Procedural semantics for hyperintensional logic: Foundations and applications of transparent intensional logic. Volume 17. Springer Science & Business Media (2010)
3. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for czech. In: Human Language Technology. Challenges for Computer Science and Linguistics, Berlin/Heidelberg (2011) 161–171
4. Horák, A.: Computer Processing of Czech Syntax and Semantics. Tribun EU (2008)
5. Jakubíček, M., Kovář, V., Šmerk, P.: Czech Morphological Tagset Revisited. Proceedings of Recent Advances in Slavonic Natural Language Processing 2011 (2011) 29–42
6. Santorini, B.: Part-of-speech tagging guidelines for the penn treebank project (3rd revision). (1990)
7. Medved', M., Horák, A., Kovář, V.: Bilingual logical analysis of natural language sentences. RASLAN 2016, Recent Advances in Slavonic Natural Language Processing (2016) 69–78
8. Tichý, P.: The semantics of episodic verbs. *Theoretical Linguistics* 7(1-3) (1980) 263–296
9. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis (2002)

Idiomatic Expressions in VerbaLex

Zuzana Nevěřilová

NLP Centre, Faculty of Informatics,
Masaryk University, Botanická 68a,
602 00 Brno, Czech Republic
xpopelk@fi.muni.cz

Abstract. Idiomatic expressions are part of everyday language, therefore NLP applications that can “understand” idioms are desirable. The nature of idioms is somewhat heterogeneous — idioms form classes differing in many aspects (e.g. syntactic structure, lexical and syntactic fixedness). Although dictionaries of idioms exist, they usually do not contain information about fixedness or frequency since they are intended to be used by humans, not computer programs.

In this work, we propose how to deal with idioms in the Czech verb valency lexicon VerbaLex using automatically extracted information from the largest dictionary Czech idioms and a web corpus. We propose a three stage process and discuss possible issues.

Key words: idioms, VerbaLex, verb valency, e-lexicography, Czech

1 Introduction

Formulaic sequences are defined as “a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.” [16] Idioms are one type of formulaic sequence characterized by its non-compositionality, i.e. its meaning cannot be deduced from the meaning of its parts.

Unlike proverbs, idioms can have literal meaning as well as figurative meaning. In practice, only some idioms occur in their literal meaning, other idioms are used solely in the figurative meaning. For example, while *kick the bucket* can mean a physical action of somebody’s foot, it usually means *to die*. By contrast, *cry heart out* is used only in the figurative meaning.

The following properties of idioms are often studied:

- their fixedness or degree of prefabrication (continuity, fixedness of the word order),
- syntactic anomalies or lexical constraints (e.g. only words from a given set can be objects of an expression),
- the usage of the literal meaning.

The non-compositionality of idioms is the main cause to build and maintain resources of idioms for the purpose of natural language processing (NLP).

The aim of this work is to re-include idioms in the Czech verb valency lexicon VerbaLex. Currently, verb frames in this lexicon with some fixed part are annotated as idioms but no other information is provided. For example, the fixed part is stored in its inflected form, so the idioms with the same headword are not grouped together. Also, the information about the degree of fixedness (fixed or free word order, usage in different tenses or moods, etc.) is missing.

In addition, the idiomatic frames are sometimes grouped with frames similar in the literal meaning, not the figurative one. For example, the idiomatic frame *uplést si na sebe bič* (make a rod for one's own back) is assigned to the synset *uplést/uplétat* (to knit). On the other hand, the idiom *polykat andělíčky* (meaning to be drowning) is (correctly) assigned to the synset *topit se, polykat* (to drown).

In order to solve the problems described above, we propose a methodology on how to describe idiomatic verb frames with respect to their fixedness, syntactic anomalies, lexical constraints and corpus frequency (without considering the literal or figurative meaning). We also suggest how to assign an idiomatic verb frame to the correct verb synset, either an existing or a new one.

1.1 Practical Outputs

Not only automatic processing of idioms is an interesting topic *per se*, the practical applications of this work can be seen.

Sentiment Analysis Affect is one of the properties of most idioms, so it is clear that dealing with idiomatic expressions is part of sentiment analysis, e.g. [11,15].

Machine Translation Presence of idioms in a sentence may have impact on the quality of statistical machine translation and the system need to transfer idioms properly to target language. For evaluation, see e.g. [13]. There are several methods to overcome errors caused by idioms in statistical machine translation, e.g. [3,2].

1.2 Outline

Section 2 describes the initial language resources. In section 3, we show how are idioms described in other language resources. Section 4 proposes a three phase methodology on how to extract information from a dictionary of Czech idioms and convert it to verb frames. In Section 5, we discuss the possible issues concerning the extraction.

2 Current Resources of Idioms in Czech

Currently, Czech idioms are described in two highly overlapping resources. Their extent and usability from the NLP view differs significantly.

idiom	chovat/hřát (si) hada na prsou/za řadry
grammatical constraints	ot, pas, imp
explanation	věřit někomu nekriticky
translation	cherish a serpent in one's bosom

Fig. 1. Example entry in the Dictionary of Czech phraseology and idioms: the headword is in bold, items are order by headwords alphabetically.

2.1 Dictionary of Czech Phraseology and Idioms

As a source of Czech idioms, we took the *Dictionary of Czech Phraseology and Idioms* (DCPI, [17]). It distinguishes four types of idioms: similes, expressions without verb, verb expressions, and sentences. For our work, we only picked the third part. It contains 19,121 idioms, however the number does not include variants (see below).

An example entry can be seen in Figure 1. The digital version of DCPI contains mostly visual markup and therefore it is not straightforward to extract all variants of the idiom. In the example, all correct variants are:

- chovat hada na prsou
- chovat si hada na prsou
- hřát hada na prsou
- hřát si hada na prsou
- chovat hada za řadry
- ...

2.2 VerbaLex

Verb valency lexicons usually consist of the following units:

- verb synset – a set of synonymic verbs describing an action, event or state
- verb frame – syntactic and semantic description of sentence constituents dependent on the verbs from the synset
- slot – description of each dependent constituent

VerbaLex [7] is the largest verb valency lexicon for Czech. It contains 6,244 verb synsets, 19,158 verb frames, 10,449 unique verbs. An example frame can be seen in Figure 2.

Each verb synset contains information on whether it is used in the passive form. Each slot contains description of some of its syntactic properties: the case of the noun phrase and the preposition if applicable.

Semantic information is available on two levels:

- *semantic role* (also known as thematic role or thematic relation) that a sentence constituent plays with respect to the action or state. VerbaLex contains 33 semantic roles such as agent, patient, location or substance.



Fig. 2. Example frame from VerbaLex. The verbs in the synset sometimes form pairs of perfective/imperfective verbs. This particular frame means *to cradle somebody in one's arms, lap etc.*

- *semantic constraint* on a hypernym (e.g. person). This second level is related to WordNet hypernym [4] (e.g. person:1, where person is a literal and 1 is the sense number).

Currently, other constraints (e.g. a set of words that can fill a slot, information about word order fixedness) are not implemented.

[6, 5.3.5] describes the annotation of idiomatic frames as follows:

- only idioms from DCPI with frequency higher than 10 in the corpus ALL¹ are included in VerbaLex
- some unspecified idioms not present in DCPI are also included in VerbaLex
- the whole fixed part is described as one slot (its “semantic role” is DPHR, meaning *dependent part of phrase*)
- information about the meaning of the idiom is described in the frame definition
- information about syntactic anomalies of lexical constraints is not included

Currently, VerbaLex contains 1,109 frames with DPHR. [10] distinguishes univalent, bivalent, trivalent, and quadrivalent valencies in English and states that trivalent idioms are very rare and quadrivalent ones seem not to exist. Apparently, the situation in the Czech lexicon is very similar: the vast majority of idioms in VerbaLex are univalent.

3 Related Work

In this section, we describe in short other works that deal with idioms in the context of NLP. We do not describe lexicons of idioms in other languages but focus on idioms in language resources usable in NLP. Idioms occur in all the resources mentioned below since they describe stereotypical patterns.

¹ A 600 million corpus created in Natural Language Processing Centre.

VALLEX In [9], the authors mention that VALLEX contains some very frequent idioms but the focus of the work is on verb in their primary meanings.

VerbNet [14] only mentions that the coverage of VerbNet was extended by WordNet [4] and these verb were also part of idiomatic expressions. Moreover, the idiomatic expressions were grouped together according to their meaning, not the surface structure, e.g. *kick the bucket* is grouped with *to die*.

FrameNet [12, 3.2.7] distinguish idioms and support predicates² verb+noun. In FrameNet, idioms are treated as multi-word targets. In both cases, the expression is defined in an appropriate frame according to its meaning.

Pattern Dictionary of English Verbs (PDEV) [5] uses the word *norm* for both literal meaning and conventionalized metaphors and idioms. By contrast, the dynamic metaphors are called *exploitations*. Since our focus is in a dictionary, we deal only with norms. PDEV contains idioms such as *to sing praises of something* but they are not distinguished from other verb patterns. Also, there is no grouping by meaning (i.e. no relationship between *sing praises* and *praise*).

4 Proposed Methods

In order to include idioms into VerbaLex, we propose several steps: extracting the idioms from DCPI, searching the idioms in corpus, and creating verb valency frames for frequently used idioms.

4.1 Extraction from DCPI

During the extraction phase, the most difficult part is dealing with the visual markup and dealing with errors in the markup. We propose the following steps:

1. extract idiom and its explanation from the \LaTeX markup (i.e. delete other content: historical context of the idiom, translations to other languages than English etc.),
2. expand content in parentheses into several variants (e.g. *hřát (si)* means *hřát* or *hřát si*)
3. expand variants delimited by slashes (e.g. *na prsou/za řadry*)
4. remove \LaTeX markup

The most difficult part is expanding variants delimited by slashes since it is not clear how much content is in each variant. We therefore suggest to over-generate the expansion and reduce the unused variants in the next phase. For example from the idiom presented in Figure 1 *chovat/hřát (si) hada na prsou/za řadry*, the algorithm can generate the following variants:

² A support predicate is a governing verb in cases the syntactic and semantic heads differ, i.e. it “does not reliably have the same meaning independently of the frame-evoking element” [12]

1. *chovat/hřát (si) hada na prsou ňadry
2. chovat/hřát (si) hada na prsou
3. *chovat/hřát (si) hada na za ňadry
4. chovat/hřát (si) hada za ňadry

4.2 Corpus Search

In order to check usage of the idiom in current language, we propose to search it in a large web corpus. At first, we decided *not* to process idioms containing auxiliary or modal verbs. The main reason is that auxiliary verbs are not included in VerbaLex at all, modal verbs are included in several frames not in a consistent way.

The proposed steps are as follows:

1. parse idiom in order to identify verb phrase and the rest (objects, adverbials)
2. if the idiom contains auxiliary or modal verb and no other verb, do not process it
3. optionally exclude non-standard Czech
4. recognize variables (e.g. somebody, something, somewhere) in the idiom and in its explanation
5. construct CQL query for the idiom
6. if the idiom is not found in the corpus, do not process it

We do not have information about word order fixedness, so we propose to construct several CQL queries with different word orders. Again, the over-generation is not a problem. We propose to search the verb phrase as a lemma in order to find different word forms. On the other hand, we propose to search exact word forms of the other parts of the idiom since it can contain non-standard words that can easily be lemmatized incorrectly.

After the corpus search, the sum of the frequencies of different word orders should be above a certain threshold. In addition, the distribution of these frequencies can provide information about the word order fixedness.

4.3 Integration to VerbaLex

The last step is the creation of the appropriate verb frames. We are not sure this process can be fully automated. At least verb frames can automatically be proposed and then checked manually. We propose the following steps:

1. create individual slots for different parts of the idiom
2. parse the explanation and identify verb phrase and the rest (objects, adverbials)
3. find an existing verb frame with a similar meaning

We are aware that many idioms do not have a one word equivalent since their meaning is rather complex (see the discussion below). The final decision whether place the idiom among other frames of a verb synset or whether create a new verb synset will be left on lexicographers.

5 Discussion

We expect that the idioms selected by the procedures described above will strongly overlap with those already present in VerbaLex. However, by using our methods the idiom description should be more detailed.

5.1 Standard vs. Non-standard Language

DCPI contains idioms in both standard and non-standard Czech. However, NLP tools are mainly focused on standard language and also only VerbaLex only contains standard Czech. We propose to exclude non-standard Czech at first and to compare frequencies of standard and non-standard idioms in corpora.

5.2 Different Degree of Fixedness

Currently, no information about the degree of fixedness is present in VerbaLex. However, DCPI contains information on syntactic constraints (e.g. the tenses the idiom can be used in). We propose to add this information as a metainformation to each frame and optionally check the constraint in corpus. We also propose to add information about word order fixedness based on the corpus search (described in 4.2).

The most difficult part will be a description of specific syntactic constraints on the argument. For example, in the idiom *aby <PAT> husa kopla* (meaning strong disagreement), we did not find any other arguments in the patient slot than pronouns.

5.3 Assignment of an Idiomatic Frame to Verb Synset

Apparently, not many idioms can be explained by one verb. One such example is *to kick the bucket* which can be “translated” as *to die*. In some cases, the idioms intensify a one verb expression, for example, *to laugh your head off* means to laugh (a lot). In these two cases, connection with the close meaning verb (in our examples, *to die* and *to laugh* respectively) is the desired state. Nevertheless, some idioms convey a complex action and no single verb can replace the idiom. For example, *ocitnout se v křížové palbě* (*find oneself in the cross fire*) means to feel pressure from several sides. The border between intensified meaning and a complex action is probably fuzzy.

5.4 Literal vs. Figurative Meaning

We are aware that some idioms are used in both literal and figurative meaning, e.g. *break the ice*. Many studies measure what meanings are activated in these cases, e.g. [1]. There are also works that detect figurative meaning by computational means, e.g. [8]. However, in processing Czech idioms, we are not that far. So, we propose to postpone the research on literal vs. figurative meanings so far.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. Cacciari, C., Tabossi, P.: *Idioms: Processing, Structure, and Interpretation*. Taylor & Francis (2014), <https://books.google.cz/books?id=RgrsAgAAQBAJ>
2. Durrani, N., Schmid, H., Fraser, A., Koehn, P., Schütze, H.: The operation sequence model—combining n-gram-based and phrase-based statistical machine translation. *Computational Linguistics* (2015)
3. Elloumi, Z., Besacier, L., Kraif, O.: Integrating multi word expressions in statistical machine translation. *MULTI-WORD UNITS IN MACHINE TRANSLATION AND TRANSLATION TECHNOLOGIES MUMTTT2015* p. 83 (2015)
4. Fellbaum, C.: *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press (1998)
5. Hanks, P., Pustejovsky, J.: *Common Sense About Word Meaning: Sense in Context*. In: Sojka, P., Kopeček, I., Pala, K. (eds.) *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 3206, pp. 15–17. Springer Berlin / Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-30120-2_2, 10.1007/978-3-540-30120-2_2
6. Hlaváčková, D.: *Databáze slovesných valenčních rámců VerbaLex*. Ph.D. thesis, Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka (2007)
7. Hlaváčková, D., Horák, A.: *VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech*. In: *Proceedings of the Slovko Conference* (2005)
8. Li, L., Sporleder, C.: Using Gaussian Mixture Models to Detect Figurative Language in Context. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 297–300. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), <http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=1857999.1858038>
9. Lopatková, M., Bojar, O., Semecký, J., Benešová, V., Žabokrtský, Z.: Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In: Matoušek, V., Pavelka, T., Mautner, P. (eds.) *LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue*. vol. 3658, pp. 99–106. Springer Verlag Heidelberg, Karlovy Vary, Czech Rep., Sept. 12-16 (2005)
10. Müller, E.A.: Valence and Phraseology in Stratificational Linguistics. In: Lockwood, D., Fries, P., Copeland, J. (eds.) *Functional Approaches to Language, Culture, and Cognition: Papers in Honor of Sydney M. Lamb*. J. Benjamins (2000), <https://books.google.cz/books?id=quxDfHwth4oC>
11. Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., Wogenstein, F.: A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In: *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. pp. 7:1–7:8. WISDOM '12, ACM, New York, NY, USA (2012), <http://doi.acm.org/10.1145/2346676.2346683>
12. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: *FrameNet II: Extended Theory and Practice*. Tech. rep., ICSI (Aug 2006), <http://framenet.icsi.berkeley.edu/book/book.pdf>
13. Salton, G., Ross, R., Kelleher, J.: An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese (2014)

14. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, Faculty of the University of Pennsylvania (2005), <http://verbs.colorado.edu/kipper/Papers/dissertation.pdf>
15. Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., Spasić, I.: The role of idioms in sentiment analysis. *Expert Systems with Applications* 42(21), 7375 – 7385 (2015), <http://www.sciencedirect.com/science/article/pii/S0957417415003759>
16. Wray, A.: *Formulaic Language and the Lexicon*. Cambridge University Press, New York (2002), <http://linguistlist.org/pubs/books/get-book.cfm?BookID=2109>
17. Čermák, F., et al.: *Slovník české frazeologie a idiomatiky I-IV (Dictionary of Czech Phraseology and Idioms, SČFI)*. Academia, Praha (1983)

Part III

NLP Applications

Recognition of Invoices from Scanned Documents

Hien Thi Ha

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanicka 68a, 602 00, Brno, Czech Republic
xha1@fi.muni.cz

Abstract. In this paper, we describe the work of recognition the first page of an invoice from a set of scanned business documents. This can be applied to document management systems, document analysis systems, pre-processing of information extraction systems. We also present our experiments on Czech and English invoice data set.

Key words: classification, recognition, invoice, OCR, Czech

1 Introduction

The processing of business documents, particularly invoices is playing a vital role in companies, especially in large ones. Usually, invoices are classified and relevant details are extracted manually by staff members and input into database systems for further processing. This manual process is time-consuming, expensive and at risk of errors because of large volumes, various layouts and different delivery formats. For those reasons, automatic analysis systems become essential.

An overview of document analysis and recognition systems can be found in [4]. State-of-the-art of text classification algorithms is in [5], [6], [7]. In these papers, authors list different approaches for text classification tasks. Some argue that Support Vector Machine is more suitable for text data while others state that Naive Bayes is more effective ([6]). In general, bag of word is the most popular method to represent document but features' dimension is huge. Therefore, they pay more attention on feature selection techniques. Fortunately, invoices do not have so many words in common (items are not taken account into these shared keywords). Moreover, scanned invoices have typical layout structures such as blocks and tables. In [2] and [8], they presented rule based approach and case-bases reasoning method for document structure recognition. Furthermore, information extraction from invoices are proposed in [3], [9]. There is a note that, in these systems, they used commercial OCR systems to process invoice images.

In this paper, we focus on classifying invoices of Czech companies which are mainly in Czech and English. The paper is constructed as follows: in section 2, we describe the classification task. Then, we will explain our experiments and discuss the results gained from our data set in section 3 and finally are conclusion and future work in section 4.

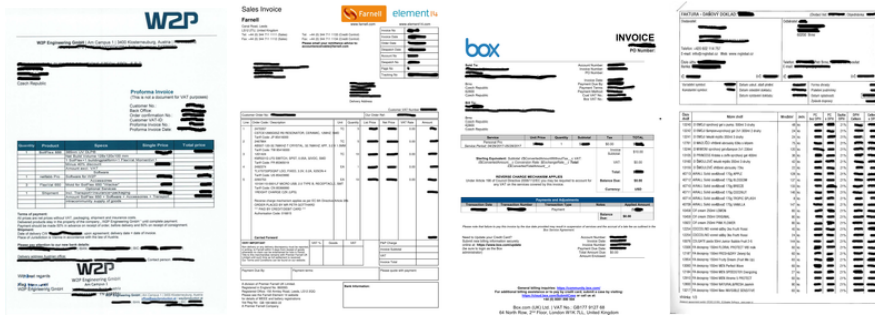


Fig. 1. Various layout formats of invoice examples

2 Classification

Workflow of the system is adapted from standard classification systems (see Figure 2). In the preprocessing step, pdf-to-image converting, image analysis (skew, quality enhancement) and languages detection are done if necessary. Then, an OCR tool is used for converting the document images into layout structures and characters: words, bounding boxes, font features and so on. After that, features are extracted to train models. In the testing phase, documents are preprocessed, features extracted and put throw trained classifier to get predicted output label.

2.1 Preprocessing

Currently, Tesseract Open Source OCR Engine (tesseract-ocr [10])¹ is among the best open source OCR engines. Since the third version, it has been supported more than 100 languages. To run tesseract-ocr from the command window:

```
tesseract imagename outputbase [-l lang] [-psm pagesegmode] [config...]
```

There are configurations for a list of languages, page segmentation mode and output format. If they are not set, then, the default are English, fully automatic page segmentation but no OSD mode, and text output relatively. Users should notice that different orders of language setting (e.g. eng+ces and ces+eng) produce different results. There are 13 selections for page segmentation mode and tesseract-ocr supports output in text, searchable PDF, hocr and tvs.

Here we run tesseract-ocr twice. The former is to detect languages used in the document. In this first run, language setting includes all possible languages of the document. For example, invoices in Czech companies usually have different versions, mostly in Czech, English, sometimes in Polish, German. Then, we base on the language distribution of words in the document to decide which languages and order of languages are used for the latter running tesseract-ocr.

¹ See <https://github.com/tesseract-ocr/tesseract/wiki/Documentation>

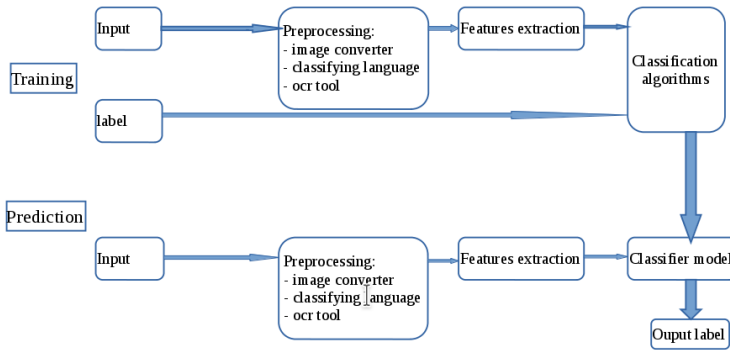


Fig. 2. Workflow of the classification system

For instance, English version invoices in Czech companies sometimes compose both English characters and Czech characters (in names, addresses). So, we set “-l eng+ces” for the second round running tesseract-ocr on these documents and add “hocr” option to get the output in hocr format.

2.2 Invoice Features

A full definition of an invoice by *The businessdictionary.com*² is as follows:

“A nonnegotiable commercial instrument issued by a seller to a buyer. It identifies both the trading parties and lists, describes, and quantifies the items sold, shows the date of shipment and mode of transport, prices and discounts (if any), and delivery and payment terms.

² See <http://www.businessdictionary.com/definition/invoice.html>

Table 1. Data set

fold	Training set		Testing set	
	invoice	not invoice	invoice	not invoice
1	528	466	62	49
2	527	467	63	48
3	527	467	63	48
4	524	470	66	45
5	529	465	61	50
6	530	465	60	50
7	534	461	56	54
8	538	457	52	58
9	542	453	48	62
10	531	464	59	51

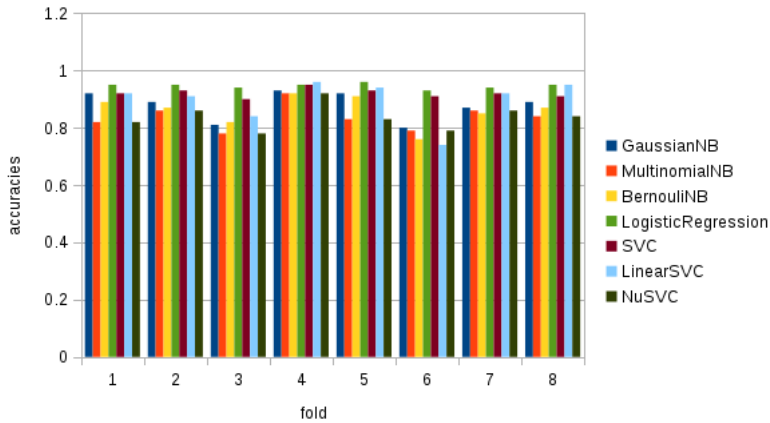


Fig. 3. Accuracies of classifiers

In certain cases (especially when it is signed by the seller or seller’s agent), an invoice serves as a demand for payment and becomes a document of title when paid in full. Types of invoice include commercial invoice, consular invoice, customs invoice, and pro-forma invoice. Also called a bill of sale or contract of sale.”

Based on the definition, we can see that a typical invoice includes title “Invoice” (or “Tax Invoice”, “Sale Invoice”, “Proforma Invoice” and so on), a unique reference number, date of invoice, names and contact details (address, phone number, email) of providers and customers, tax payment (if relevant), description of purchased items, price, total amount, delivery and payment terms. If an invoice spreads into several pages, then there will be a page number, usually at the bottom of the page.

Table 2. Accuracies of classifiers

classifiers	1	2	3	4	5	6	7	8	9	10	mean
GaussianNB	0.92	0.89	0.81	0.93	0.92	0.87	0.93	0.80	0.87	0.89	0.89
MultinomialNB	0.82	0.86	0.78	0.92	0.83	0.76	0.90	0.79	0.86	0.84	0.83
BernouliNB	0.89	0.87	0.82	0.92	0.91	0.85	0.92	0.76	0.85	0.87	0.87
LogisticRegression	0.95	0.95	0.94	0.95	0.96	0.93	0.95	0.93	0.94	0.95	0.95
SVC	0.92	0.93	0.90	0.95	0.93	0.89	0.92	0.91	0.92	0.91	0.92
LinearSVC	0.92	0.91	0.84	0.96	0.94	0.85	0.93	0.74	0.92	0.95	0.92
NuSVC	0.82	0.86	0.78	0.92	0.83	0.76	0.90	0.79	0.86	0.84	0.83

The image shows a scanned invoice document. At the top, there is a header section with various fields and text, including identification numbers and company information. Below the header is a large table with multiple columns, likely representing a list of items or services provided, with columns for quantity, unit, price, and total. The text is somewhat blurry due to the scanning process.

Fig. 4. An example of a misclassified invoice

2.3 Classification

Assume that we have got a set of business documents including orders, invoices, emails, receipts, offer price, attachments, delivery notes and so on. The task is choosing invoices from the collection. The invoices are recognized based on their first page.

From the output of the OCR process, features are extracted. Firstly, lower-case words frequencies are counted. Then, 150 words appear most frequently on the first page of invoices are selected to be keywords. These keywords are then cleaned by removing private nouns (names, locations). After having keywords, a vector of keywords is extracted (equal 1 if the keyword is found in the document, 0 otherwise). On the other hand, the position (top) and size (width, height) of title "invoice" are extracted from the page. Here, feature "left" is not meaningful because title can be on the left, right or at the middle. If there is no such title, zero value is set for each feature. In addition, another important feature is page number. In some invoices, they keep the same general information (title, invoice number, date, supplier and customer information), just different list of items. To separate the first page from other pages, a page number is an essential pivot. If there is not any page number information, we see the page as the first page by default. Here is a piece example of feature vector from a page of an invoice:

```
{ "datum": False, "doklad": False, "faktura": True, "dič": False, "cena": False, "symbol": True, "dodavatel": False, "číslo": True, "spol": True, "variabilní": False, "ks": False, "tel": True, "množství": False, "banka": False, "dodaci": False, "oddíl": True, "splatnosti": True, ..., "top": 123, "width": 321, "height": 46, "page": 1 }
```

We try with probabilistic models such as Naive Bayes, neuron network based classifiers like Logistic Regression and Support Vector Machine.

3 Experiments

3.1 Dataset

For our experiments, 998 documents with 1505 pages are received. Out of 1505 pages, there are 1105 ones in Czech(590 first pages of invoices and 515 are not first pages of invoices), 10 in Polish and 390 in English. Polish files are pruned. Data set for English is small, so we focus on Czech ones. Invoices comes from various vendors, so the layout varies greatly (examples are in Figure 1).

3.2 Results

First of all, PDF files are converted into images. In this experiment, we use Portable Document Format to Portable Pixmap converter (pdftoppm)³. The default resolution of output images is 150×150 (dpi). Users are able to specify it to increase the quality of images. Then, these images are put through tesseract-ocr using language setting '-l eng+ces'. The output file is read and language distribution is counted to define the language of the document. After that phase, tesseract-ocr is used once more time with detected languages to get the words and layout format for feature extraction.

Hocr files are read into dictionaries having following keys: "text", "wordset" (including words and bounding box). Then features which are vector of keywords, top, width, height of the title "invoice" ("faktura" for Czech invoices) and page number are extracted.

To assess classifiers, we use 10-fold cross-validation. Accuracies of classifiers for each fold are listed in table 2 and depicted in Figure 3. Among classifiers, Logistic Regression scores the best with average 95.02% are recognized correctly. Support vector machines (SVC) and Linear SVC get nearly the same median

³ See <https://freedesktop.org/wiki/Software/poppler/>

Table 3. Precision, recall and F-score of Logistic Regression model

fold	TP	FP	TN	FN	precision	recall	F-score
1	60	3	46	2	0.95	0.97	0.96
2	60	3	45	3	0.95	0.95	0.95
3	59	3	45	4	0.95	0.94	0.94
4	63	2	43	3	0.97	0.95	0.96
5	60	3	47	1	0.95	0.98	0.97
6	54	2	48	6	0.96	0.90	0.93
7	54	3	51	2	0.95	0.96	0.96
8	50	6	52	2	0.89	0.96	0.93
9	43	2	60	5	0.96	0.90	0.92
10	58	4	47	1	0.94	0.98	0.96
average	58.5	3	46.5	2.5	0.95	0.96	0.95

Table 4. Result according to feature modification

classifiers	accuracy	precision	recall	F-score
GaussNB				
all features	0.89	0.89	0.90	0.90
keywords only	0.90	0.88	0.94	0.92
keywords+page	0.91	0.89	0.94	0.92
title+page	0.85	0.93	0.78	0.85
Logistic Regression				
all feature	0.95	0.95	0.96	0.95
keywords only	0.94	0.94	0.95	0.94
keywords+page	0.94	0.94	0.95	0.95
title+page	0.84	0.91	0.79	0.84
NuSVC				
all features	0.83	0.89	0.79	0.83
keywords only	0.93	0.91	0.96	0.94
keywords+page	0.93	0.91	0.95	0.94
title+page	0.83	0.89	0.79	0.84

value but the former is more stable through folds than the latter. We should notice that samples in negative class (not a first page of an invoice) are sometimes really similar to positive samples such as page 1 and page 2 in the same invoices, or invoices and order lists.

Having a close look at Logistic Regression model's results, most of files are correctly recognized and there are very few false negative (average 2.5(4%)) and false positive (average 3(6%)). Detail data is in table 3.

Errors are partly because of OCR errors. For example, in the invoice in Figure 4, the title "FAKTURA" (invoice) is wrong recognized as "FAGTURA". Therefore, title position features are set zeros. In this situation, apart from Logistic Regression and Linear SVC, all other classifiers classify it as not a first page of an invoice. Logistic Regression model predicts right 8 out of 10 times whereas Linear SVC scores 10/10.

Surprisingly, when we remove title position features and only keep keywords or keywords and a page number, all models, except Logistic Regression, have improvements in accuracy, recall and F-score, particularly recall because of looser constrains. Examples of changes on average measurements can be seen in table 4.

4 Conclusion and Future Work

In this paper, we have built a classification system to recognize the first page of invoices from scanned documents. Based on used features, it can be adapted for other languages. This work mainly uses words, the smallest unit in document layout, to extract features. In subsequent work, we will construct

blocks of information based on available features (geometric and textual features of word, line), and use Natural Language Processing tools such as named entity recognition to extract semantic meaning of blocks. This will provide important features for classification as well as information extraction from scanned invoices.

5 Acknowledgements

This work has been partly supported by Konica Minolta Business Solution Czech within the OCR Miner project and by the Masaryk University project MUNI/33/55939/2017.

References

1. Ronen Feldman and James Sanger: *The Text Mining Handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press, New York, 2007.
2. Stefan Klink, Andreas Dengel, Thomas Kieninger: *Document Structure Analysis Based on Layout and Textual Features*. In: *Proc. of International Workshop on Document Analysis Systems, DAS2000*, 2000.
3. T.A Bayer and H.U.Mogg-Schneider: *A generic system for processing invoices*. IEEE, 1997.
4. MARINAI, Simone: *Introduction to document analysis and recognition*. Machine learning in document analysis and recognition, 2008, 1–20.
5. Fabrizio Sebastiani. *Machine learning in automated text categorization*. ACM Computing surveys, vol. 34, No. 1, March 2002, pp.1–47.
6. S.L. Ting, W.H. Ip, Albert H.C. Tsang. *Is Naïve Bayes a Good Classifier for Document Classification?* *International Journal of Software Engineering and Its Applications* Vol. 5, No. 3, July, 2011.
7. AGGARWAL, Charu C. and ZHAI, ChengXiang. *A survey of text classification algorithms*. *Mining text data*, 2012, 163–222.
8. HAMZA, Hatem, BELAÏD, Yolande and BELAÏD, Abdel. *Case-based reasoning for invoice analysis and recognition*. In: *ICCBR*. p. 404–418. 2007.
9. SCHULZ, Frederick, et al. *Seizing the treasure: Transferring knowledge in invoice analysis*. In: *Document Analysis and Recognition, ICDAR'09, 10th International Conference on*. IEEE, p. 848–852, 2009.
10. Smith, Ray: *An overview of the Tesseract OCR engine, the Ninth International Conference on Document Analysis and Recognition, ICDAR 2007*, vol. 2, pp. 629–633, IEEE, 2007.

Enlargement of the Czech Question-Answering Dataset to SQAD v2.0

Terézia Šulganová, Marek Medved', and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xsulgan1, xmedved1, hales}@fi.muni.cz

Abstract. In this paper, we present the second version of Czech question-answering dataset called SQAD v2.0 (Simple Question Answering Database). The new version represents a large extension of our original SQAD database. In the current release, the dataset contains nearly 9,000 question-answer pairs completed with manual annotation of question and answer types.

All texts in the dataset (the source documents, the question and the respective answer) are provided with complete morphological annotation in plain textual format. We offer detailed statistics of the SQAD v2.0 dataset based on the new QA annotation.

Key words: question answering, QA dataset, SQAD

1 Introduction

Question Answering (QA) is a rapidly evolving field of Natural Language Processing (NLP) and Informatics. We may regard QA as a basis for next generation search engines. If we set aside natural language interfaces to database queries, each QA system uses a large enough knowledge base that provides information for the answer, often in the form of large textual documents.

In this paper, we introduce a new version of a QA evaluation dataset created from a document collection coming from the Czech Wikipedia. SQAD v2.0 is a largely extended version of the previous SQAD database [1] used in evaluation of a syntax based question-answering system AQA [2]. The key assets of the new dataset version are the $3\times$ increased dataset size with nearly 9,000 question-answer pairs (the previous SQAD contained 3,301 QA pairs). During the development of the dataset, the original short answer contexts have been expanded into full Wikipedia articles.

Besides the exact answer phrases, the new corpus contains the full answer sentence, which can be used in a separated evaluation of the answer selection process. All QA pairs have now been also annotated for the question type and answer type via manual annotation process by two annotators.

From the technical point of view, the database was also modified to avoid document duplicities by applying symbolic links between texts shared across multiple records.

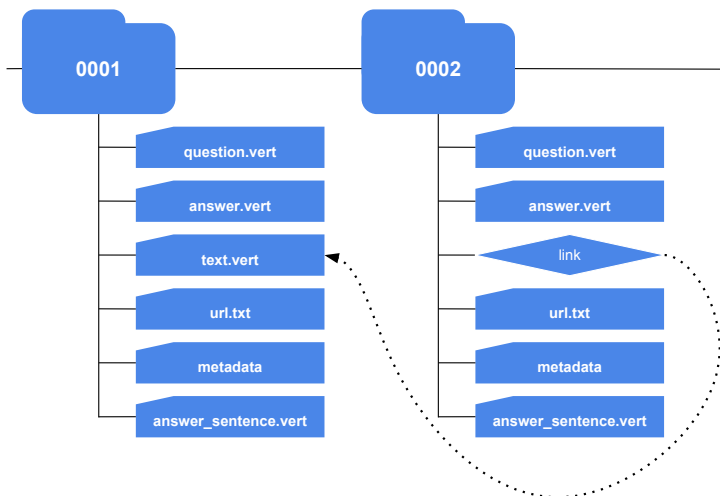


Fig. 1. SQuAD v2.0 components structure

In the following text, we describe the whole process of developing the new SQuAD v2.0. In section 2, we define the database structure and the data content. Section 3 describes the necessary automatic and manual adjustments of the SQuAD v2.0 dataset including new metadata. In section 4, we provide statistics of token and sentence numbers, numbers of question and answer types.

2 Database Structure

Structure of the new SQuAD database generally follows the previous one with some additional files. The previous version [1] consisted of files in plain text form denoting the original text, the question, the answer, Wikipedia URL and the question author. These input files were processed by morphological annotating pipeline formed by the Unitok [3] tokenizer and the Desamb [4] tagger.

The previous SQuAD version did not allow easy sharing of manual modifications in either the underlying texts or the morphological annotation in case when there were multiple questions with the same source text. If the plain text form changed, the morphological annotation had to be rebuilt and in case of manual corrections, all of them had to be reapplied again.

Therefore the new SQuAD version has several changes in the files structure. The new version consists exclusively of vertical files, which represent a textual format of morphologically annotated text. A schema of two question-answer pairs is displayed in Figure 1. The standard plain text version of each component can be generated from this vertical files on request. `question.vert`, `answer.vert`, `answer_sentence.vert` and `text.vert` are sentences in the

Question:			Answer sentence (part of):		
<i>word/ token</i>	<i>lemma</i>	<i>POS tag</i>	<i>word/ token</i>	<i>lemma</i>	<i>POS tag</i>
<s>			<s>		
Z	z	k7c2	Další	další	k2eAgFnSc7d1
jakého	jaký	k3yQgInSc2	paměti-	paměti-	k1gFnSc7
roku	rok	k1gInSc2	hodností	hodnost	k7c6
pochází	pocházet	k5eAaImIp3nS	v	v	k7c6
školní	školní	k2eAgFnSc1d1	Paršovicích	Paršovice	k1gFnPc6
budova	budova	k1gFnSc1	je	být	k5eAaImIp3nS
v	v	k7c6	školní	školní	k2eAgFnSc1d1
obci	obec	k1gFnSc6	budova	budova	k1gFnSc1
Paršovice	Paršovice	k1gFnSc2	z	z	k7c2
<g/>			roku	rok	k1gInSc2
?	?	kIx.	1898	#num#	k4
</s>			<g/>		
			,	,	kIx,
			tehdy	tehdy	k6eAd1
			nazvaná	nazvaný	k2eAgFnSc1d1
			...		

Fig. 2. Vertical format of the question “Z jakého roku pochází školní budova v obci Paršovice? (What year does the school building in Paršovice come from?)” and (a part of) the answer sentence “Další pamětihodností v Paršovicích je školní budova z roku 1898, tehdy nazvaná ... (Another place of interest in Paršovice is the school building of the year 1898 at that time named as ...)” with the expected answer of “1898” marked with bold font.

vertical format (see Figure 2 for an example). The `url.txt` and `metadata` are in plain text form. The `text.vert` file can be also represented by a symbolic link that leads to file with the same content used in a different record. This provides consistency in changes and decrease redundancy in the whole dataset.

3 Database Adjustments

Before the final SQuAD v2.0 release, we have perform multiple automatic and manual changes to correct tagger/tokenizer mistakes and supplement the corpus with additional metadata.

3.1 Automatic Adjustments

As in the previous version, we provide tokenization adjustments, fix out-of-vocabulary mistakes and a few regular morphological errors.

Apart form these changes, we have prepared a semi-automatic process to identify answer sentences. Such file is important for QA system evaluation on

the sentence selection level – whether the QA system is able to pick the correct sentence with the answer from the whole knowledge base.

3.2 Manual Adjustments

After the automatic changes, there were several manual changes performed on the data.

Very valuable information for QA development is represented by the question and answer type annotation of each record. This annotation was provided manually and the following question and answer types have been used.

The collection of annotated *Question types* takes an inspiration from The Stanford Question Answering Dataset [5]. Each question in SQuAD v2.0 is tagged by one of the following question type:

- (i) Date/Time
- (ii) Numeric
- (iii) Person
- (iv) Location
- (v) Other Entity
- (vi) Adjective phrase
- (vii) Verb phrase
- (viii) Clause
- (ix) Other

The annotated *Answer types* were taken from the first layer of Li and Roth's [6] two-layered taxonomy with a few adaptations. Each answer was thus assigned one of following types:

- (i) Date/Time
- (ii) Numeric
- (iii) Person
- (iv) Location
- (v) Entity
- (vi) Organization
- (vii) YES/NO
- (viii) Other

The remaining manual corrections of SQuAD v2.0 are mostly technical and are related to the fields of Wikipedia URL, question and answer files to harmonize them with current Wikipedia content. First, the URL of Wikipedia articles changes quite often which makes problems when the article is moved to a new URL and the previous URL is redirected to another article with different content. Because of the live Wikipedia community, the articles changes frequently and that is why several questions and answers had to be adapted to the current text where the previous information was no longer present in the current data.

Table 1. SQuAD v2.0 knowledge base statistics

Number of tokens	20,272,484
Number of sentences	911,014
Number of sentence selections	6,349
Number of source documents	3,149

4 Dataset Characteristics

The final SQuAD v2.0 dataset consist of 8,566 question-answer pairs related to 3,149 documents obtained from Czech Wikipedia. The documents' texts are included as the underlying knowledge base with the corpus size of 20,272,484 tokens. The answer sentence selection list contains 6,349 sentences – see Table 1 for a summary of the SQuAD knowledge base proportions.

Each question is annotated with of the selected questions types. The characteristics of the expected answers are categorized with the corresponding answer type. Overall statistics of the question and answer type distributions are presented in Tables 2 and 3.

Table 2. Question type statistics in SQuAD v2.0

Date/Time	1,848
Numeric	900
Person	940
Location	1,436
Other Entity	1,440
Adjective phrase	253
Verb phrase	944
Clause	774
Other	31

Table 3. Answer type statistics in SQuAD v2.0

Date/Time	1,847
Numeric	904
Person	943
Location	1,442
Entity	811
Organization	199
YES/NO	940
Other	1,480

5 Conclusions and Future Work

In this paper, we have introduced a new extended and manually annotated version of the SQA dataset for question-answering evaluation. The second version contains nearly nine thousand records with manual annotation of question and answer types with each record.

Current planned steps of the work with the fresh new database concentrate on providing an evaluation of the syntax based question-answering system AQA based on this enhanced and enlarged evaluation dataset.

Acknowledgements. This work has been partly supported by the Czech Science Foundation under the project GA15-13277S and by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. Horák, A., Medved', M.: SQA: Simple question answering database. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2014) 121–128
2. Horák, A., et al.: AQA: Automatic Question Answering System for Czech. In: International Conference on Text, Speech, and Dialogue, Springer (2016) 270–278
3. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU. (2014) 71–75
4. Pavel Šmerk: Towards morphological disambiguation of Czech (2007)
5. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
6. Li, X., Roth, D.: Learning question classifiers. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics (2002) 1–7

Semantic Similarities between Locations based on Ontology

Moiz Khan Sherwani^{*1}, Petr Sojka¹, and Francesco Calimeri²

¹ Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czechia
mksherwani@mail.muni.cz, ORCID: 0000-0001-6061-6753

sojka@fi.muni.cz, ORCID: 0000-0002-5768-4007

² Dept. of Mathematics and Computer Science, University of Calabria, Calabria, Italy
calimeri@mat.unical.it, ORCID: 0000-0002-0866-0834

Abstract. Toponym disambiguation or location names resolution is a critical task in unstructured text, articles or documents. Our research explores how to link ambiguous locations mentioned in documents, news and articles with latitude/longitude coordinates. We designed an evaluation system for toponym disambiguation based on annotated GEO-CLEF data. We implemented a node-based approach taking population into account and a geographic distance-based approach. We have proposed new approach based on edges between the pairs of toponyms in ontology, taking also population attribute into account. Our edge-based approach gave better results than population and distance-based only approaches. The results could be used in any information system dealing with texts containing geographic locations, such as news texts.

Key words: toponym disambiguation, geonames, geographic text retrieval, ontology based geoname relations, toponym similarity

Everything has to do with geography. (Judy Martz)

1 Introduction

Toponym disambiguation or *place name resolution* is a process of assigning location names (toponyms) that appear in article by normalizing them with the help of their respective coordinates and their context of appearance. This process turns out to be quite difficult for locations that are highly ambiguous and in short texts. [14] Gazetteers are the main source for the identification and disambiguation of the location names. Gazetteer serves as the dictionary for the geographical entities, usually with the set of properties (City, Country, Continent, Coordinate, Population, Alternative Names, Administration Division etc.) about every location. GeoNames³ is very well known Gazetteer with all ambiguous location names structured according to the respective classes.

* Reported work has been done during Erasmus+ stay at the NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czechia

³ <http://www.geonames.org>

Toponym disambiguation is regularly performed in two stages. The initial step, *toponym recognition*, discovers all occurrences of names and put them in an archive. There are, for example, over one hundred places in the world with the name Alexandria. The second step, *toponym disambiguation*, assigns latitude and longitude and scope to all names found in the initial step.

Many people prefer to read articles, news and blogs online, so it is important to provide structured, location-based reading sources for the geographic entities. The process of understanding geography from any type of unstructured content, articles or text is called *geoparsing*, *geocoding* or *geotagging* [8,10,9]. Once the geonames are disambiguated, their results could be used for indexing and search, for document similarity computations, for document filtering, infosystems alerting an the like.

The paper is structured as follows: In Section 2 we review related work. In Section 3, our experimental setup to perform the evaluation of the research is discussed. We describe our methodology, evaluation metrics and datasets in Sections 4 and 5. Section 6 serves for reporting the results achieved. We conclude in Section 7 by summarizing our outcomes, and suggest future work.

2 Related Work

The idea of toponym disambiguation is to identify all location names stated in an article and to specify these location names with the coordinates latitude and longitude. In this research, we are not considering the references to some location names, e.g. "1 km south of Brno" or "around the University of Calabria", rather this research focuses on the use of specific location names. Identifying toponyms has been widely studied in *named entity recognition* (NER) research: location names were one of the main classes of named entities to be distinguished in article [12]. Most of the approaches are based on the toponym disambiguation are driven by the physical properties of the toponyms. Some methods rely on external sources [4]. Properties depicting Geo-spatial areas, as well as their relations on the Earth are utilized for disambiguation. This approach is supported by several heuristics as explained by Leidner [7].

One approach utilizes the attributes of the Geo-spatial areas to resolve any uncertainty between toponyms. The significance of an area is frequently computed by having the location with the biggest population. Another approach assumes that an article is likely to refer to places within a constrained geographical zone, so it picks places that are near to each other. For example, if an article contains the ambiguous location names Rome, USA, Texas, this approach will select Rome in USA instead of the Rome in Italy based on the distance between the location names mentioned in the article. [8] Even though great advances in toponym disambiguation have been made this decade, it is still difficult to decide which approach is the best.

A different problem emerges from non-standard practices of the *gazetteers* (Geo-spatial lexicon) when they allocate coordinates to location names. According to [3] Cambridge has just two locations in Wordnet, 38 in Yahoo! Geoplanet,

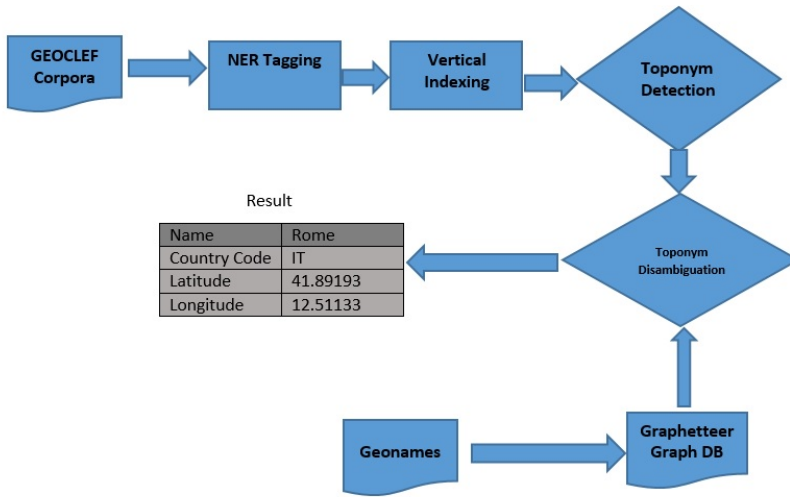


Fig. 1. An overview of our toponym disambiguation work-flow setup

and 40 in GeoNames. The scope may vary extensively from one resource to another, and the latitude and longitude allocated for the same area may fluctuate too, bringing about an unjustified disambiguation when scoring the frameworks.

Another problem arises with the use of different corpora, because the variations in the corpora that are used for the evaluation make it complicated to find the best approach [13]. A new approach that is said to have higher precision considering one corpora will have variations if used on the other corpora.

3 Experiment Setup

Successful disambiguation will be hugely beneficial to systems that utilize place names, as the names themselves cannot always be used for disambiguating them. Let us consider the toponym Springfield, for instance. Around the world, there are no less than 33 settlements with that name – it is also part of college and University names. At the point when the system is given an article about Springfield College, these lines not straightforwardly clear to which Springfield College it indicates. If the article happens to mention other places, this situation changes. For instance, if Springfield, MA (Springfield, Massachusetts) is specified, the message likely alludes to Springfield in Massachusetts state of USA. In this case, toponym was identified in the interesting way based on the State name given with the toponym. We have designed the system in such a manner that it consists of two phases. In the first phase of our experiment, toponyms are extracted from the articles. The second phase reports the disambiguation of each toponym recognized by the first phase, hence all the toponyms are assigned with their respective coordinates based on the gazetteer.

Toponym Extraction and Entity Look-up The first phase for toponym disambiguation requires the labeling of toponyms in the article. To assign the location labels in the article, any named entity recognizer or gazetteer can be connected. The first step involves matching the extracted toponyms with a candidate location name in the location network. We utilize Stanford's Named Entity Recognition system [5] to both coordinate the toponyms in the article with those in the system, and to sort out the toponyms within the articles. As a result of this step, we receive more than one toponym in each article but there are other possible cases: **i)** no match is found, **ii)** one match found (un-ambiguous location), or **iii)** more than one match is found (ambiguous location).

Toponym Disambiguation As a result of the first stage in toponym recognition, we have list of toponyms for each article. This will be used for the disambiguation of toponyms. Resolving the cases mentioned above would be as follows. In the first case, there is no match for the toponym in the area system and therefore it cannot be connected to any area. In the second case, the area specified is unambiguous, and the assignment of connecting it to the network in the system is clear. In the third case, we find toponym with ambiguity, which can be resolved in a variety of ways. Our system takes the result of a previous phase as an input and assigns coordinates in latitude and longitude. To carry out this disambiguation, we have executed different approaches, two of which proved to be effective in studies proposed by [7]. The first is the population based approach and second approach is based on the distance. Moreover we have developed a novel approach that is called the edge-based approach, because we have used the graph database for the computations and would like to introduce a new approach based on graphs to enhance this procedure.

Node-Based Approach This approach is figured in light of the population property of the GeoNames database. It will continuously pick the location which has the highest population for all ambiguous locations. Therefore Rome in Italy will dependably be favored over Rome in any of 15 states of the USA. One limitation of this approach is the incorrect population data of continents that sometimes appears in the GeoNames database, as well as a few different toponyms. We have made changes to the different states and continents population manually where the value of population was mentioned 0. This approach is mostly taken into consideration: if only one location name is mentioned in the article and there is no other reference to be considered for the ambiguous location.

Geographic Distance-Based Approach This approach is figured based on the shortest distance between the toponyms. The distance between the locations is computed by the *haversine formula* that takes latitude and longitude of the locations as an input and results in the distance between the candidates. All ambiguous toponyms present in the article would be used as separate pairs and

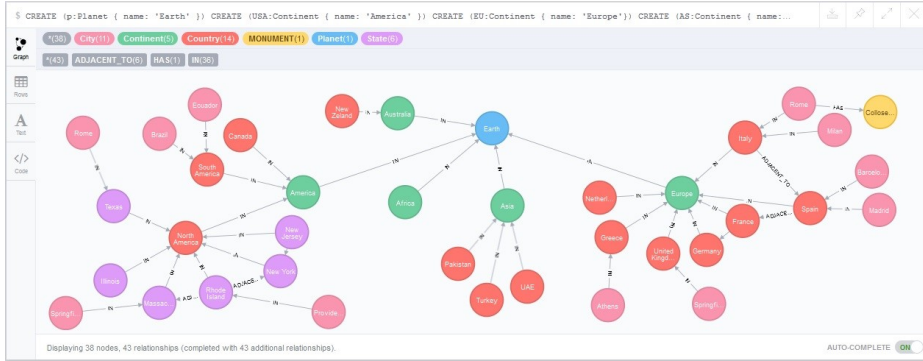


Fig. 2. Sample diagram of graphetteer designed on NEO4

the shortest distance between the toponyms would be selected as the resulting toponym.

Edge-Based Approach We are introducing the edge-based approach for our computation. This approach could only be computed with the graph based databases. It computes the distance based on the number of edges and nodes between the pair of toponyms. But with this approach, we have one constraint. For example, if we have set of toponyms Italy and Rome in the article and we gave Italy and Rome in the Graphetteer as an input to compute the edge-based approach. The resulting output for the edge-based approach might give Italy and Rome in USA based on the smaller number of edges. But our required nodes were to get the Italy and Rome that belongs to Europe. To achieve satisfactory results, we have attached the query of considering the population property of the node to improve the precision.

4 Methodology, Evaluation Metrics

To conduct our experiments, we decided to work on the graph database instead of a traditional relational one.

Graphetteer Previously, researchers have used the gazetteer based on the traditional RDBMS for toponym disambiguation. In this research, the Gazetteer that is used for the location database is taken from GeoNames, US National Geospatial Intelligence Agency and US board on Geographic Names. Graphetteer is the name given to the Gazetteer based on Graph Database, and the conceptual model for Graphetteer was proposed by [2]. Graph database has several new features and algorithms to perform the evaluation. GeoNames database was cleaned based on our research requirements (for example: columns with values of the modification date).

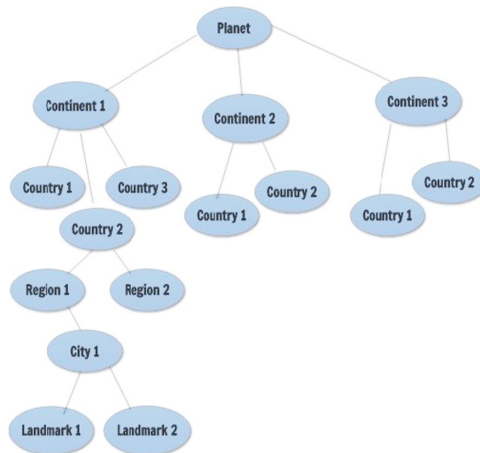


Fig. 3. Sample diagram for our Geographic Ontology

After cleaning the data, it was arranged according to the continents and then NEO4J was used to convert all the database into the Graph database. Cypher Query Language is the main language that is used to work on the NEO4J database. In this database, we have used Continents, Countries, Cities, Regions, Administration Divisions, Counties and Monuments as Nodes. The relationships between them are the edges. Two types of relationship edges were used:

IN edge to show the relationship between towns to cities, cities to countries, countries to continents;

HAS edge to show the famous places, tourist attractions and monuments within a city.

Toponym Recognition An alternative to NLTK's NER classifier is provided by the Stanford NER tagger⁴ [5]. This tagger uses an advanced statistical learning algorithm it's more computationally expensive than the option provided by NLTK. It labels sequence of words in an article based on the 7 classes (Location, Organization, Person, Money, Percent, Date and Time). Our requirement was to extract the location names from the article. For this purpose, we have used 3 class model to label the location names in the articles. Our approach consists of the following steps:

1. Run the Stanford NER tagger on the GEOCLEF database. This would label the article based on three classes (Location, Name and Organization). One drawback of the Stanford NER tagger is that it labels United Kingdom as United /Location Kingdom /Location.
2. Use Python script to merge the multiple named location names that has more than single /Location label into single label.

⁴ <https://nlp.stanford.edu/software/CRF-NER.shtml>

3. Convert the article into “vertical” or “word-per-line (WPL)” format, as defined at the University of Stuttgart in the 1990s. This method allows us to find the distance between the words in the article.
4. Extract every location name that have /Location along with the assigned index.

Toponym Resolution Once we have all the location names listed in the article, We will be using the script to assign the location coordinates, latitude and longitude based on the approaches that we have considered for the evaluation of this research. We created three different scripts to evaluate the approaches based on population, distance and edge-based. All location names files are run through the graphetteer and the resulting locations are achieved with their coordinates for our evaluation.

Metrics We have used typical metrics to evaluate the approaches: Precision, Recall and F_1 -measure F_1 to evaluate the performance of our toponym disambiguation experiments:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is the count of correctly disambiguated toponyms by the system, FP is the count of incorrectly disambiguated toponyms by the system, FN is the count of toponyms not identified by the system. Since our geographical database and the annotated corpus is based on the GeoNames DB, the toponyms effectively distinguished are known by a basic match between the place IDs retrieved by the framework and with the annotated corpus.

5 Dataset, Data Preprocessing

Corpus We used the GEOCLEF corpus to carry out our evaluation. This stands for Geographic Cross Language Evaluation Forum. This corpus consists of articles from English and German sources. For evaluation, we have used the corpus Information Retrieval in English that consist of the 169,477 articles from the Glasgow Herald (British) 1995 and LA Times (American) 1994. These articles are associated with (number of toponyms) tokens. This corpus is widely used to conduct the research based on the Geographic Information Retrieval. Further details can be found at GEOCLEF 2008⁵ link. Toponym statistics are to be found in Table 1.

Gazetteer For toponym disambiguation, we needed a gazetteer to specify locations for each toponym in the articles. To acquire a gazetteer that secured overall data, we used the GeoNames database for locations. It is a completely

⁵ <https://www.uni-hildesheim.de/geoclef/>

Table 1. Numbers of 10 most frequent toponym in the corpus

Toponym	Frequency	Ambiguous
Los Angeles	139,881	Yes
Glasgow	73,402	Yes
United States	65,993	No
Scotland	34,835	Yes
Washington	24,135	Yes
California	23,812	Yes
United Kingdom	16,180	No
New York	12,079	Yes
London	11,790	Yes
England	11,240	Yes

accessible gazetteer containing more than 10 million entries worldwide. Every location entry contains a name, alternative names, administrative level, country codes, latitude/longitude coordinates and elevation. Each location has their respective coordinates and geonames ids that make them unique from every other location entry. We have processed the data according to our needs and all the properties of the data are included in the Graphetteer except for the alternative names with special characters in it.

6 Results

We have compared three approaches: node-based approach, geographic distance-based approach and edge-based approach. Results are summarized in the Table 2. For the node based approach, where we have considered population as the main property for disambiguation. We computed all approaches on 169,450 articles from GEOCLEF and the toponym frequency is also given in the Table 1. There are 1,238,686 toponyms occurrences in all articles together.

All of the data and code is available for download for reproducibility and comparison of approaches. Our evaluation framework is available on the project web page https://nlp.fi.muni.cz/projekty/toponym_disambiguation. Since our graphetteer and the annotated article corpus is based on GeoNames database. Toponyms that are identified as positive candidates are

Table 2. Toponym Disambiguation on GEOCLEF data based on different approaches

Approach	Precision	Recall	F ₁
Node-based	0.70	0.89	0.78
Geographic distance-based	0.39	0.89	0.54
Edge-based	0.74	0.89	0.80

referred by the GeoNames ID resulted by our system and the experts annotated the corpus.

GIS is waking up the world to the power of geography, this science of integration, and has the framework for creating a better future. (Jack Dangermond)

7 Conclusion and Future Work

We have compared three approaches to toponym disambiguation. We have proposed a new approach based on the edges between the pairs of toponyms in an ontology, taking a population attribute into account. According to our comparison between the most commonly used heuristics (population and geographic distance), the best results were achieved using the edge-based approach.

Using a graph database is efficient and as new features could be used to compute like centrality measures, it brings new opportunities for further improvements, e.g. matching nodes based on the relationships and their specific properties.

Several toponym disambiguation approaches could be supported by our framework in the future:

vector representations Taking the context in which a toponym was used is the key for a further increase of precision. Vector space word representations and their similarity computed by word2vec [11,6] or similar system is yet another way to be tested in the future.

weighting Experiments with weighting based on the level of Ontology, e.g.: Continents is on the top level followed by Countries and so on and lowest level is considered as the street or landmark in a City. Starting from the top level higher weights and lower weights for the bottom level ontology.

metadata We can also improve the result by using the metadata of article news, and a knowledge base about the location names.

alternate toponym names One can handle the alternative names for the locations with special characters or letters from other languages than English. To disambiguate toponyms with location names in different languages, corpora based on other languages would also be required.

voting Using different approaches to disambiguation to vote on the right toponym disambiguation. Hybrid approaches are giving excellent results [1].

geonames similarity It is important to quantify similarity of geographical names for the purpose of information retrieval, alerting systems and other uses of disambiguated toponyms.

Different approaches will be compared and evaluated on the same data.

Acknowledgments Funding of the TA ČR Omega grant TD03000295 is gratefully acknowledged.

References

1. Badieh Habib Morgan, M., van Keulen, M.: Named entity extraction and disambiguation: The missing link. In: Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval. pp. 37–40. ESAIR '13, ACM, New York, NY, USA (2013), <https://doi.org/10.1145/2513204.2513217>
2. Bär, M.: Graphetteer – A conceptual model for a graph driven gazetteer (Jan 2016), <http://www.geonet.ch/graphetteer-a-conceptual-model-for-a-graph-driven-gazetteer/>
3. Buscaldi, D.: Approaches to Disambiguating Toponyms. SIGSPATIAL Special 3(2), 16–19 (2011), <https://doi.org/10.1145/2047296.2047300>
4. Buscaldi, D., Rosso, P.: Map-based vs. knowledge-based toponym disambiguation. In: Proc. of the 2nd International workshop on Geographic IR. pp. 19–22. ACM, Napa Valley, CA, USA (2008), <https://doi.org/10.1145/1460007.1460011>
5. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting of ACL. pp. 363–370. ACL '05, ACL, Stroudsburg, PA, USA (2005), <https://doi.org/10.3115/1219840.1219885>
6. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR abs/1405.4053 (2014), <http://arxiv.org/abs/1405.4053>
7. Leidner, J.L.: Toponym Resolution in Text (Annotation, Evaluation and Applications of Spatial Grounding). Dissertation Abstract. ACM SIGIR Forum 41(2), 124–126 (2007), <https://doi.org/10.1145/1328964.1328989>
8. Lieberman, M.D., Samet, H.: Multifaceted Toponym Recognition for Streaming News. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information – SIGIR '11. pp. 843–852 (Jul 2011), <https://doi.org/10.1145/2009916.2010029>
9. Lieberman, M.D., Samet, H.: Adaptive context features for toponym resolution in streaming news. In: Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval – SIGIR '12. pp. 731–740 (Aug 2012), <https://doi.org/10.1145/2348283.2348381>
10. Lieberman, M.D., Samet, H.: Supporting Rapid Processing and Interactive Map-Based Exploration of Streaming News. In: International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2012) (Nov 2012), <https://doi.org/10.1145/2424321.2424345>
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013), <http://arxiv.org/abs/1310.4546>
12. Piskorski, J., Yangarber, R.: Information Extraction: Past, Present and Future. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization. pp. 23–49. Springer (2013), https://doi.org/10.1007/978-3-642-28569-1_2
13. Roberts, K., Bejan, C.A., Harabagiu, S.: Toponym Disambiguation Using Events. In: Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010). pp. 271–276 (2010), <http://www.aaii.org/ocs/index.php/FLAIRS/2010/paper/viewFile/1291/1754>
14. Sagcan, M., Karagoz, P.: Toponym Recognition in Social Media for Estimating the Location of Events. In: Proc. of 15th IEEE International Conference on Data Mining Workshop, ICDMW '15. pp. 33–39 (2016), <https://doi.org/10.1109/ICDMW.2015.167>

Part IV

Text Corpora

Language Code Switching in Web Corpora

Vladimír Benko^{1,2}

¹ Slovak Academy of Sciences, L. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava, Slovakia

² Comenius University in Bratislava, UNESCO Chair in Plurilingual and Multicultural
Communication

Šafárikovo nám. 6, SK-81499 Bratislava, Slovakia

vladimir.benko@juls.savba.sk

<http://www.juls.savba.sk/~vladob>

*I Love Music Party 10. apríla 2009 v New Cage Clube v Lučenci*³
*I love youúúúúúúúú I love Americáááááááá*⁴

Abstract. One of the challenges in building and using web corpora is their rather high content of “noise”, most notably having the form of foreign-language text fragments within otherwise monolingual text. Our paper presents an approach trying to cope with this problem by means of “exhaustive” stop-word lists provided by morphosyntactic taggers. As a side effect of the procedure, a problem of tagging text with missing diacritics is also addressed.

Key words: web-based corpora, language identification, morphosyntactic annotation, Aranea Project

1 Introduction

“Noise” in texts derived from web can be of various nature. In our work, however, we want to address just one particular type of text noise, namely language code switching, by which we also consider “code switching” in discussions, where some of the participants do not use diacritics. In the framework of our *Aranea*⁵ Project [1], we are currently trying to improve the morphosyntactic annotation of the Slovak *Araneum Slovacum* corpus, so that it could be used by lexicographers as a source of lexical evidence for the multi-volume *Dictionary of the Contemporary Slovak Language*⁶. The same problem, albeit to a lesser extent, can be observed in the *Slovak National Corpus*⁷, so a successful solution could be used here as well.

³ <http://www.ilovemusic.sk/clanky/i-love-music-party-10-aprila-2009-v-new-cage-clube-v-lucenci>

⁴ http://dolezite.sk/Raz_v_tom_ma_jasno_253.html

⁵ http://aranea.juls.savba.sk/aranea_about/

⁶ http://www.juls.savba.sk/sss_j_6.html

⁷ <http://korpus.juls.savba.sk/>

pr-web.sk	píská a šumí, mozek se neprokrvuje... Z toho plyne špatná paměť, zapominání, přidá se vysoký	<input checked="" type="checkbox"/>
beo.sk	ti amici jazdit keď nie na ruskej rope a plyne ...na kravske prdy? ¶	<input type="checkbox"/>
pluska.sk	peniaze. Ady má totiž domček v Lozorne, kde na plyne varí, ohrieva vodu i kúri. A každé usporené	<input type="checkbox"/>
pluska.sk	je golfové ihrisko. Takže čo usporim na plyne , vrazím do golfu,“ vysvetľuje. Pritom ale	<input type="checkbox"/>
euroekonom...	od roku 2001. Kupříkladu, z dat UAH MSU plyne ochlazovací trend mezi lednem 2001 a květnem	<input checked="" type="checkbox"/>
euroekonom...	sopkami a ENSO-m) ¶ Co z vývoje Ap indexu plyne ? Nejspíš to, že nás čeká další rok extrémně	<input checked="" type="checkbox"/>
zahori.est...	Francúzsko. K odstráneniu závislosti na ruskom plyne môžu prispieť aj náleziská plynu na južnom	<input checked="" type="checkbox"/>
zahori.est...	troch až piatich rokov bude Slovensko na plyne nezávislé“. ¶ Expremiér Mikuláš Dzurinda	<input type="checkbox"/>
despitebor...	nariadenie nasledovala smernica o zemnom plyne v roku 1998 (98/30/EC). V oboch sa požadovalo	<input type="checkbox"/>
despitebor...	2003/55/EC (5) o elektrickej energii a o plyne predstavujú najväčšie zmeny doterajšieho	<input type="checkbox"/>
despitebor...	znamenalo vyvlastnenie, čo by hlavne pri zemnom plyne viedlo k zvýšeniu cien pre koncových zákazníkov	<input type="checkbox"/>
burjanosko...	střední školy a jeho tři zástupci. ¶ - Co z toho plyne : měli bychom si, jako daňoví poplatníci	<input checked="" type="checkbox"/>
energia.sk	zaostávame. V Čechách mení v elektrine a zemnom plyne dodávateľa 15 až 20 percent domácností.	<input type="checkbox"/>
energia.sk	plyn ho dokáže nahradit, dopyt po zemnom plyne bude rásť a cena pôjde tiež hore. ¶ Ak sa	<input type="checkbox"/>
diskusie.s...	připravili na vojnu, a před začátkem klamstva o plyne boli v Izraeli hlavní americkí generáli	<input type="checkbox"/>
sixpack.cz	telekomu atd.. Prakticky aj pri elektrike, plyne a vode. Tam neexistuje alternativa, ale	<input type="checkbox"/>

Fig. 1. Czech sentences in the Slovak text (“plyne”)

2 The Task

From the lexicographers’ perspective, the problem of foreign language text fragments in otherwise monolingual text is mostly associated with so-called interlingual (“false”) homographs, i.e., lexical items present in two languages, yet having (usually) a different meaning. This phenomenon is especially pronounced between close languages, such as Slovak and Czech. If the Slovak lexical item is rather rare and, on the other hand, the same word form in Czech is frequent, the resulting concordances may contain comparable number of occurrences in both languages. For example, the Slovak word form “plyne” (noun, locative case singular of “plyn”, English “gas”) is a form of “plynout” in Czech (verb, 3rd person singular, English “flow”). As seen in Figure 1, more than one third of the occurrences come in fact from Czech sentences (ticked in the screenshot).

Frequent foreign lexical items can cause problems even if they are not homographs, as they “spoil” the frequencies of out-of-vocabulary word forms that have to be manually checked in search for potential neologisms. If we succeed in reliable detection of foreign-language text fragments, we will be able to remove or, at least, mark them so that they would not be shown as results of corpus queries.

3 The Scope of the Problem

Our long-term experience with Slovak corpora shows that the most frequent foreign-language fragments in Slovak texts come from English and Czech. To estimate the rough proportion of English lexical items in the Slovak corpus we used a simple method: from a list of the most frequent word forms of our *Araneum Anglicum* English corpus, we deleted words that do exist in Slovak as

word	Frequency	Items: 50 Total frequency: 4,986,801
P N the	1,062,246	
P N of	685,764	
P N and	500,643	
P N in	432,580	
P N is	202,642	
P N for	196,944	
P N it	182,233	
P N with	117,479	
P N as	104,812	
P N you	99,701	
P N one	77,755	
P N from	69,304	
P N this	68,374	
P N all	68,082	
P N at	67,137	
P N that	65,695	

Fig. 2. English words in the Slovak corpus

well (such as “to”, “a”, “on”) and used the top 50 from the resulting list to create a corpus query:

```
the|of|and|in|is|that|for|with|it|you|are
|as|be|was|have|this|...
```

The top 16 lines of the respective frequency list are shown in Figure 2.

The total normalized frequency is 1,682.80 i.p.m., which is really quite a lot – it means, that these very 50 English words represent almost 0.17% of all tokens of the corpus.

Using the same method, we can also estimate the proportion of the Czech lexical items. Here, however, the number of inter-lingual homographs is much larger. The situation may also be complicated by the Slovak (and also Czech) text fragments without diacritics that increase the number of the respective inter-lingual homographs. We have, however, decided not to take this phenomenon into account here. The resulting frequency distribution (first 16 lines) is shown in Figure 3.

The total normalized frequency is 1,571.0 i.p.m here, which looks pretty similar to that of English. In reality, however, the number will be much higher due to the already mentioned inter-lingual homographs.

4 The Method

Language identification belongs to traditional tasks in the area of Natural Language Processing, as well as in that of Information Retrieval, with state-of-the-art methods exhibiting precision well over 95% (cf. [5]). These methods can be basically divided into two groups: (1) methods counting frequent words that are based on lists of “stop words”; (2) methods counting individual characters

word	Frequency	Items: 50 Total frequency: 4,655,548
P N ze	1,530,617	
P N co	837,991	
P N se	482,948	
P N velmi	267,340	
P N den	193,408	
P N pro	178,093	
P N jsem	109,782	
P N jako	91,904	
P N ve	77,974	
P N jsou	56,038	
P N když	45,822	
P N ke	43,205	
P N které	36,798	
P N není	34,947	
P N který	32,108	
P N byl	31,500	

Fig. 3. Czech words in the Slovak corpus

or character n-grams – “statistical methods”. The main advantage of statistical methods is that they are typically able to identify language from short strings containing just several hundreds of characters, as well as that they are usually computationally “cheap”. The performance of the stop-list methods usually depends on the size of the respective list. The disadvantage of both is rather low performance in distinguishing very similar languages and practical inability to cope with the texts containing code switching.

In our case, we do not require fast computation as the corpus annotation is a time-consuming process anyway. We also would like to possibly make use of existing tools – the morphological analyzers and taggers. Conceptually, our method can be put into the stop-list category, assuming that the size of the list is limited just by the size of the morphological lexicon used by the respective analyzer.

The main idea is as follows: (1) Besides the basic morphosyntactic annotation by the standard tagger (using the Slovak language model), the corpus is processed by alternative taggers (using language models for languages that we want to identify); (2) information about the result of the morphological lexicon lookup performed by the respective taggers is gathered; (3) by combing information from different taggers language of each token is estimated; (4) using summary information regarding the individual tokens of the sentence its language is stated.

The actual annotation has been carried out by *TreeTagger* [7] using our own language models for Slovak and Slovak without diacritics [2], and by the morphological component of *MorphoDiTa*⁸ [9,8] using the newest Czech language model. The processing used our standard *Aranea pipeline* [3] for each

⁸ As we only need information on the morphological lexicon lookup, the disambiguation phase provided by the tagger is not necessary here.

	ztag	ztag1	ztag2	ztag3	Frequency	Items: 59 Total frequency: 10,000,000
P N	1	1	0	0	3,974,506	
P N	1	1	1	0	2,792,725	
P N	1	1	1	1	2,083,398	
P N	1	1	0	1	442,205	
P N	0	0	0	0	254,507	
P N	0	0	1	0	111,336	
P N	1	2	0	0	61,511	
P N	0	1	0	0	54,026	
P N	1	2	1	0	47,379	
P N	0	0	1	1	43,124	
P N	0	0	0	1	39,878	
P N	1	2	1	1	29,941	
P N	0	1	1	0	18,721	
P N	2	2	0	0	9,995	
P N	1	2	0	1	9,186	
P N	0	1	1	1	8,096	

Fig. 4. Morphological lexicon lookup results (alphabetical tokens considered only)

language, and the resulting partial verticals have been combined by the *cut* and *paste* utilities, so that the resulting vertical would contain 17 columns for corpus attributes as follows: *word*, *lemma*, *atag*, *tag*, *ztag*, *lemma1*, ... The first five attributes belonged to the original Slovak annotation, and the indexed attributes contained alternative annotations – index 1 indicated Slovak without diacritics, 2 indicated Czech, and 3 indicated English, respectively.

For our purposes, the most important in all parallel annotations is usually the respective *ztag* attribute having a non-zero value if the morphological lexicon lookup has been successful.

5 The Initial Experiment

The development of the language identification algorithm based on the respective parallel annotation has been started by producing a *ztag* value distribution for a random 10 million alphabetic tokens from the 200-million token test corpus. Figure 4 shows its first 16 lines.

The table is slightly complicated to read because of values in the first two columns: besides the “0” (wordform not found in lexicon) and “1” (found) values, both Slovak language models produce also numbers larger than 1 in situations where the tagger was not able to disambiguate the respective lemma – in such cases, the number of variant lemmas is shown. It is also clear that the values in the second column cannot be smaller than those in the first one, as the “diacritics-less” language model must always yield at least the same result as the “full” model.⁹ The first two rows of the table are quite expected – most wordforms have been recognized by the Slovak models, followed by both Slovak and Czech models. A surprise is the contents of the third line. What are those words that are present in Slovak, Czech and English? See Figure 5.

⁹ All diacritics have been stripped from the source vertical before the diacritics-less annotation has been applied.

word	lemma3	tag3	Frequency	Items: 10,200 Total frequency: 2,083,398
P N a	a	DT	336,754	
P N v	v	NN	204,724	
P N na	na	TO	202,217	
P N je	Je	NP	134,035	
P N to	to	TO	68,862	
P N o	o	NN	59,365	
P N si	si	NP	56,561	
P N do	do	VVP	55,336	
P N z	z	SYM	49,940	
P N ale	ale	NN	33,046	
P N V	V	NN	32,020	
P N k	k	NN	31,113	
P N by	by	IN	28,362	
P N od	od	NN	27,709	
P N tak	Tak	NP	27,506	
P N po	po	NN	26,848	

Fig. 5. Words recognized by all language models

ztag	ztag1	ztag2	ztag3	Frequency	Items: 58 Total frequency: 10,000,000
P N 1	1	0	0	5,091,887	
P N 1	1	1	0	3,159,066	
P N 1	1	1	1	621,345	
P N 0	0	0	0	327,822	
P N 1	1	0	1	264,436	
P N 0	0	1	0	142,349	
P N 1	2	0	0	74,219	
P N 0	1	0	0	69,113	

Fig. 6. Morphological lexicon lookup results (3 and more letters)

The beginning of the list contains lots of strange short “English” words tagged as “proper noun” (“NP”), even if they are written in lowercase letters. We therefore decided to refine our query and restrict it to words longer than 2 letters. Figure 6 shows the result.

Now the statistics looks much more “probable”, and we can utilize the information that the language identification should be based on longer words.

6 The Algorithm

Besides the experience gathered from the introductory experiment, we also made use of information from external source: we knew that the size of the Czech morphological lexicon is almost by order of magnitude larger than that of Slovak [4], better covering not only large amounts of loanwords and rare lexical items, but also numerous proper and geographical names. This often means the Czech morphological analyzer can recognize many proper names in Slovak text, whereas the Slovak one mostly fails here.

The algorithm design has been performed in an iterative way, using the smaller (2 million token) test corpus that has been compiled by NoSketch Engine¹⁰ [6] after each iteration to analyze the results. The values of internal

¹⁰ <https://nlp.fi.muni.cz/trac/noske>

1	eopen.sk	En.1 en1:0	Cloud		f	
2	quovadis-o...	En.1 sk1:0 cs2:1 en3:2	WAY TO EDUCATION - konferencia		f	
3	hbreavis.c...	En.1 en1:0	Retail		f	
4	blog.fouzo...	En.1 sk1:0 cs2:0 en6:4	1) What 's the first taste you remember ?		f
5	blog.fouzo...	En.1 cs2:0 en7:6	2) An anecdote about your work with food ?		f
6	blog.fouzo...	En.1 sko:0 3:0 cs2:0 enu0:9	3) The fine dining place where you would take someone special ?		f
7	blog.fouzo...	En.1 sko:0 1:0 cs3:1 en9:8	4) Could you tell us something about your future projects ?		f
8	blog.fouzo...	En.1 cs3:2 en5:4	5) The dish to die for ?		f
9	blog.fouzo...	En.1 cs1:0 en4:3	6) Your biggest culinary anxiety ?		f
10	wildcats.s...	En.1 sk2:1 cs2:1 en3:1	Posts Tagged 'vtipné video'		f	
11	kamericana...	En.1 sk1:0 cs1:0 en2:1	Read More		f	
12	divadelnet...	En.1 sk1:0 en2:1	Alebo lets dance .			

Fig. 7. Sentences recognized as English with the respective metadata

variables used to decide the language of the respective sentences could be conveniently encoded and displayed as attributes of the <s> structure. To display sentences according to identified language we used a CQL query like this:

```
<s lang="En.*"> []
```

Figure 7 shows several sentences resulting from that query.

The string preceding each sentence denotes the recognized language and (for each language) two values separated by a colon representing the number of “decidable” tokens and bigrams. For example, sentence 10 has been tagged as English, as it contained three decidable words and one bigram identified as English, as opposed to Slovak and Czech (two words and one bigram).

The current version of the algorithm can be described as follows:

- (1) Only alphabetic tokens of a minimal length (currently 2), optionally followed by a full stop, are considered. Such tokens we call “decidable”.
- (2) Words marked by Slovak language model as foreign (tag “#”) are not considered as Slovak.
- (3) Words marked by English language model as foreign (tag “FW”) or proper noun (tag “NP”) are not considered as English.
- (4) Words starting with a capital letter are considered in all languages only if they have been recognized by the Slovak language model.
- (5) For Slovak, counts from non-diacritics language model are used.
- (6) If no word in the sentence has been recognized by any language model, the sentence is marked as “undecidable” (“Xx”).
- (7) The language with the largest proportion of recognized words becomes the language of the sentence.
- (8) If recognized words counts are equal, the greatest number of recognized bigrams is used for decision.
- (9) If all bigram counts are equal, the sentence is marked as Slovak.

7 The Results

Although high recall has been the priority in this work, it is something that is rather difficult to evaluate in large corpora. We therefore offer data on the

Table 1.

Identified as	Slovak	Czech	English	Undecidable	Total
Sentences (counts)	10,220,517	97,893	23,423	177,006	10,518,927
Sentences (%)	97.16	0.93	0.22	1.68	100.00
Tokens (counts)	191,915,723	1,226,762	303,681	621,421	199,660,383
Tokens (%)	96.12	0.61	0.15	3.11	100.00

precision only. Table 1 shows some results received by the beta version of our algorithm that has been used to process the larger test corpus containing approx. 200 million tokens.

It is apparent that the procedure has identified almost 3 % of foreign-language or undecidable sentences containing approx. 4 % of all corpus tokens. We can also see that the sentences recognized as Czech and English are shorter than average, while the undecidable sentences are longer than average. A short check in the corpus reveals the cause: the long undecidables contain various lists consisting of proper names that have been excluded from our algorithm.

The precision data in Table 2 has been obtained by manual checking samples of 100 sentences from each group.

Even this rudimentary evaluation shows that identification of English text fragments within Slovak text is a relatively easy task, while distinguishing between Czech and Slovak is really a “tough” one. For evaluation of identification of Slovak text this tiny sample is naturally not sufficient. The algorithm as such, however, is already usable – the 4 % loss of data in corpus is more than acceptable. The program has been implemented in *lex* programming language and its current (9th) version is just 330 lines long, inheriting some portions of the code from other programs. We expect, however, that by using a programming language with a native utf-8 support, the program might be even simpler.

8 What Have We Learned

Our experiment has proved that by using existing tools this problem could be solved with minimal additional programming necessary. We have acquired

Table 2.

Identified as (manually)	Identified as (by algorithm)				
	Slovak	Czech	English	Undec.	Cs+En+Xx
Slovak	98	25	0	26	51
Czech	0	43	0	2	45
English	0	3	89	6	98
Undecidable (Xx)	2	21	11	63	105
Other language	0	8	11	3	22

certain insight into the way how the respective taggers and language models work, which can be used to improve the process of morphosyntactic annotation of Slovak corpora. We have also discovered some peculiarities in the respective lexicons, such as English prepositions tagged as (“Czech”) prepositions and, similarly, English articles tagged as (“Czech”) adjectives in the Czech *MorphoDiTa* lexicon.

9 Further Work

The described method can be used for other languages as well. As English text fragments are present virtually in all corpora of the *Aranea* family (including the Chinese and Arabic ones), we would like to improve annotation of all corpora in the foreseeable future.

Acknowledgements. This work has been, in part, financially supported by the Slovak VEGA and KEGA Grant Agencies, Project Nos. 2/0017/17, and K-16-022-00, respectively.

References

1. Benko, V.: Aranea: Yet another Family of (Comparable) Web Corpora. In: Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8–12. Proceedings. Eds. P. Sojka et al. Cham – Heidelberg – New York – Dordrecht – London: Springer, pp. 21–29. (2014)
2. Benko, V.: Tvorba webových korpusov a ich využitie v lexikografii. Dizertačná práca. Bratislava: Filozofická fakulta Univerzity Komenského. (2016)
3. Benko, V.: Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). Portorož: European Language Resources Association (2016) pp. 4245–4248. (2016)
4. Petkevič, V.: Personal communication. Praha (2016)
5. Řehůřek, R. and Kolkus, M.: Language Identification on the Web: Extending the Dictionary Method In: A. Gelbukh (Ed.): CICLing 2009, LNCS 5449. Berlin – Heidelberg: Springer-Verlag, pp. 357–368. (2009)
6. Rychlý, P.: Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, pp. 65–70. (2007)
7. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester (1994)
8. Spoustová, D. “johanka”, Hajič, J., Raab, J. and Spousta, M.: Semi-Supervised Training for the Averaged Perceptron POS Tagger. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens: ACL, pp. 763–771. (2014).
9. Straková, J., Straka, M. and Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore: ACL, pp. 13–18. (2014)

KernelTagger – a PoS Tagger for Very Small Amount of Training Data

Pavel Rychlý

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pary@fi.muni.cz

Abstract. The paper describes a new Part of speech (PoS) tagger which can learn a PoS tagging language model from very short annotated text with the use of much bigger non-annotated text. Only several sentences could be used for training to achieve much better accuracy than a baseline. The results cannot be compared to the results of state-of-the-art taggers but it could be used during the annotation process for a pre-annotation.

Key words: PoS tagging, morphological tagging, language model, Czech

1 Introduction

Part of speech (PoS) tagging is one of the most important tasks in corpus linguistics. PoS taggers assign a PoS tag for each word from an input. They usually learn a language model (or a set of rules) using manually annotated texts. Some taggers could also exploit an optional lexicon to help annotation of words which are not found in the manually annotated text.

One of the main feature of a text corpus in the field of natural language processing is its big size. Text corpora contains from millions to billions tokens. It is not a problem to create a corpus with tens of million tokens even for small languages [1]. On the other hand, manual annotation of such corpora is a big problem, it is a time consuming and expensive process. As a result, manually annotated corpora are rather small, most of them have less than one million tokens. For example, there are only five larger corpora in the Universal Dependencies¹ – the most comprehensive database of annotated corpora. Many smaller corpora have only a few hundred sentences annotated.

It is very hard to learn anything automatically from such small corpora because they contain only a few thousands of different words and most words have only one hit in the whole corpus. The performance (accuracy) of a PoS tagger trained on such corpus is close to the baseline. A bit better performance is achieved by taggers which use a lexicon or a morphological database containing all possible PoS tags for large amount of words. That could be helpful for languages where such lexicon or database is available. Otherwise, construction of them is more expensive than PoS annotation of a small corpus.

¹ <http://universaldependencies.org/>

2 KernelTagger

KernelTagger is a new PoS tagger. It is optimized to exploit as much knowledge as possible from a large non-annotated corpus with the help of possibly very small PoS annotated corpus.

The main idea behind KernelTagger is computation of word similarity from the non-annotated corpus and using the kernel trick to derive a PoS tag for a given word from similarity to words with tag known from the small annotated corpus. There is no learning of any features from the annotated corpus, the tagger remembers PoS tags for each word from the annotated corpus. If the given word occurred in the training corpus the tagger outputs the most probable (most frequent) PoS tag for such word. The tag for unseen words is computed using a modification of a kernel perceptron on known words. The modification consists of using only top 5 most similar known words instead of all known words.

We use this modification because similarity of most (almost all) word pairs is near zero (they are not similar) and the exact number is mostly a noise. On the other hand, the similarity of similar words is quite reliable and could be used for computation.

2.1 Word Similarity Computation

In the early stages of development, we have used several different settings of a word embedding system [2] but the final version use very simple distributional similarity computed from small contexts. We use only one preceding and one following word for each keyword, we compute the logDice [3] salience score and assign the similarity of two words w_a and w_b using the following formula:

$$\text{sim}(w_a, w_b) = \frac{\sum_c \min(D(w_a, c), D(w_b, c))}{\sum_c D(w_a, c) + \sum_c D(w_b, c)}$$

where $D(w_a, c)$ is the logDice score of word w_a and context c . We use only contexts with positive logDice. The left and right contexts are handled separately: the same word before and after a keyword are treated as two different contexts.

2.2 PoS of Unseen Words

The PoS tag for a word which occurs in the training annotated corpus is the most frequent tag for that word (one is chosen randomly for several tags with same frequency).

The PoS tag for an unseen word is computed from PoS tags of most similar known words. First, we set a list of up to 5 most similar words. Then the tag for a word w is defined by the following formula:

$$\text{argmax}_t \sum_x \text{sim}(w, x) P(x, t)$$

where $P(x, t)$ is the probability of the tag t for the word w .

3 Evaluation

We have evaluated the tagger on the DESAM corpus [4], the Czech corpus of about one million manually annotated tokens. It contains lemma and morphological tag for each word. As Czech has quite complex morphology the DESAM tag-set is huge, it consists of 13 attributes with up to 7 different values. We have used only the main part of speech, that means only 12 different values, 11 for words (including one for numbers) and one for punctuation. The most frequent PoS tag is NOUN, it represents 30% of tokens.

We have used two setups for a non-annotated corpus:

1. The whole DESAM corpus. We choose this setup to demonstrate that even small (1 mil.) corpus could be useful for computing word similarities.
2. Part of czTenTen[1] corpus. Only a small part of the whole corpus was used to simulate low resource language. The used part contains 33 million tokens. There was a limit of 10 million word pairs during word similarity computation. That means only 10 million most similar word pairs are stored and used for evaluation. This limit caused that the size of the temporary data files was less than a half of the size from the DESAM setup.

The results are listed in Table 1. There are four test cases for each setup. They differ in the number of annotated tokens used for training (the first column). We can see that even 1000 tokens (representing several dozens of sentences) provides interesting accuracy.

4 Future Work

We would like to make more evaluation on the Czech corpus to measure the influence of the size of the non-annotated corpus. There are also questions on influence of several KernelTagger parameters which we have set according to just a few tests:

1. N for the top N most similar words (now: 5),
2. the size of context in computing similarity of words (now: 1),
3. the threshold of minimal logDice and minimal count of a context to be included in similarity (now: 0 and 2).

Table 1. Evaluation results: The accuracy of KernelTagger for different number of training tokens with annotation and different corpora for computing word similarities.

train tokens	DESAM (1 mil.)	czTenTen (33 mil.)
1,000	70.7	72.9
10,000	78.8	81.7
100,000	87.7	88.5
980,000	92.9	92.8

We are also going to add two modules to handle the most common errors in the *KernelTagger* annotation. First one for handling unknown words according to sub-strings. Second one for handling ambiguous words depending on context. That could increase the usability of *KernelTagger* for languages with weak morphology and high ambiguity of PoS tags for individual word forms.

5 Conclusion

In this paper, we have presented the new PoS tagger *KernelTagger*. It trains a PoS model from (small) annotated text and (big) non-annotated text. The main advantage of the tagger is its ability to provide competitive results for very small annotated texts, as small as several sentences. The tagger could be used especially for low-resource languages and for pre-annotation during manual annotation of texts. It works well for morphologically rich languages.

Acknowledgments. This work has been partly supported by the Academy of Sciences of Czech Republic under the project 15-13277S and by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL. (2013) 125–127
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Rychlý, P.: A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008 (2008) 6
4. Pala, K., Rychlý, P., Smrž, P.: Desam—annotated corpus for czech. In: SOFSEM'97: Theory and Practice of Informatics, Springer Berlin/Heidelberg (1997) 523–530

Manipulative Propaganda Techniques

Technical Report

Vít Baisa, Ondřej Herman, and Aleš Horák

Natural Language Processing Centre
Masaryk University, Faculty of Informatics
Botanická 68a, Brno

Abstract. Influencing the public attitude towards certain topics had become one of the strongest weapons in today's information warfare. The ability to recognize a presence of propaganda in newspaper texts is thus a treasured phenomenon, which is not directly transferable to algorithmic analysis.

In the current paper, we present the first steps of the project aiming at detection and recognition of selected propaganda elements in common texts. We introduce a developed tool used for annotating designated manipulative elements in newspaper texts and providing an overview of complex text characteristics of the underlying corpus data.

The presented Propaganda corpus consists of more than 5,000 newspaper articles by 4 publishing websites.

Key words: propaganda, manipulative techniques in text, propaganda corpus, propaganda detection

1 Motivation and Introduction

The freedom on the internet allows people with malicious intents to spread hatred, fake news, alternative truths, misinterpretations which might in turn instigate extreme nationalism, xenophobia, homophobia, class discrimination etc.

The work presented in this article is a part of an interdisciplinary project funded by Masaryk University. The idea of the project is a reaction to the current issue of politic propaganda of foreign entities via new media and social networks in Czech Republic. This phenomenon brings new opportunities for methodology, safety and law research and offers challenges for interdisciplinary research. On the example of pro-Russian information warfare the project develops methods of discerning, detecting and analysing of manipulative propaganda techniques in newspaper texts. Investigations also partly aim at users sharing manipulative content from the point of their motivation and evaluates security risks for the Czech Republic.

The collaborating parties include: political scientists from the Faculty of Social Studies, legal scientists from the Faculty of Law and computer scientists from

the Natural Language Processing Centre (NLP Centre), Faculty of Informatics (the authors of this paper).

The task of the NLP Centre group within the project is to develop a system capable of:

- 1) regular acquisition of web documents from a list of propaganda websites,
- 2) providing an annotation tools of the web documents for propaganda experts,
- 3) advanced search and data statistics acquisition based on the propaganda corpus data, and
- 4) developing automatic methods for discerning and classifying unseen web documents, stylometry for anonymous authorship recognition.

This paper describes the progress in the first two points.

2 Related Works

The information channels based on the World Wide Web environment approach the public via several basic access points such as the social networks (Facebook, Twitter, Instagram, ...), navigation from preselected news server website (on-line news sources) or the web search. The polarization of the first two channels is their inherent property and people expect them to be predisposed, but the web search is accepted as an *objective* tool in this respect. However, since the huge amounts of full text query results have surpassed human processable limits, the ordering and filtering of the web retrieval results can play a crucial role in polarizing them. Recent studies [1,2] show, how the topic of propaganda, fake news and manipulative text influence the current search engine techniques.

Social networks allow the speed-of-light dissemination of any viral kind of information, making it a perfect Swiss-army knife of possible propaganda and manipulation. [3] studies the tweeting behavior of Twitter propagandists and identify four features which can be used for distinguishing ideological tweeters from neutral users. The social network companies try to fight this situation using complex AI tools, user grouping rights, or fake news collaborative marking [4]. After the recent cases of massive propaganda during elections in several countries (USA, Germany, or France), the social networks employ community fact-checkers allowing thus a distributed way of manual fake news fight [5].

The actual propaganda devices and the propaganda model [6] have been theoretically studied for decades or centuries [7]. Formalizing these human techniques for computer processing is, however, a complex task possibly consisting on many subtasks such as the topic change identification [8], rumour identification and verification [9], or hoaxes and fake news detection [10].

In the following text, we concentrate on the ways of identification of possible manipulative techniques purely from the underlying text and style characteristics, without any factual verification.

3 The Annotation Scheme

Here we do not describe the annotation rules as it is out of the scope of this paper. We give a brief overview of attributes which are annotated and need to be stored in corpus data or in a database.

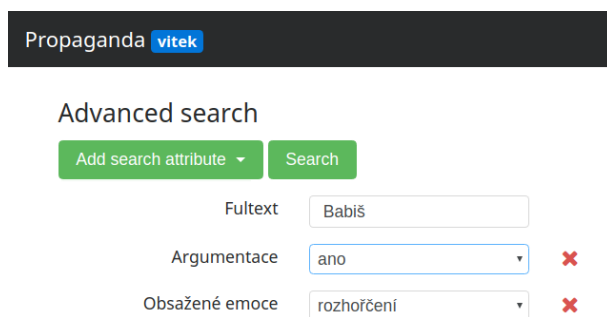
The following attributes are manually annotated for each document. We give also examples of values. The first part consists of attributes which can be assigned to a sequence of words from a document (we call them *range attributes*):

- location (a town, a country: *EU, Česká republika, USA, other country, Rusko, NATO, Rusko + USA*),
- blame (is anyone blamed for something?, boolean),
- labelling (is somebody or something labelled?, boolean),
- argumentation (does the article include arguments in favour or against something?, boolean),
- emotions (*fear, outrage, sympathy, hatred, other, missing*),
- demonizing (is something demonized?, boolean),
- relativizing (is something relativized?, boolean),
- fear (boolean),
- fabrication (boolean),
- Russia (how is the Russian Federation depicted?, *positive, neutral, negative, victim, hero, missing*),
- genre (*news, commentary, interview*),
- expert (is an expert mentioned or cited?, boolean),
- politician I–III (the name of a mentioned politician),
- sentiment I–III (*negative, neutral, positive, acclaim, hateful, missing*),

These attributes are bound to a document as a whole, but the current aim is to bind them also with the particular words and phrases from the documents to be able to capture significant correspondences which will be used in the subsequent machine learning techniques for automatic annotation (see Section 6).

The second group of attributes consists of document-level attributes, which generally cannot be reflected in concrete phrases of the document:

- topic (*migration crisis, domestic policy, international relations/diplomacy, society, other, energetics, Ukraine conflict, culture, Syria conflict, warfare policy, economics, conspiracy*),
- focus (*foreign, domestic, both, indeterminate*),
- opinion (does the article states an opinion?, boolean),
- overall sentiment (*positive, neutral, negative*),
- picture (does the article contain a picture/photo?, boolean),
- video (does the article contain a video, boolean), and
- source (does the article mention or cite a source?, boolean).



Propaganda vitek

Advanced search

Add search attribute ▾ Search

Fultext

Argumentace ✖

Obsažené emoce ✖

Fig. 1. Advanced full-text search entry form with two selected attributes of “*Argumentace*” (*argumentation*) and “*Obsažené emoce*” (*present emotions*) with the values of “*ano*” (*yes*) and “*rozhořčení*” (*outrage*).

4 The Propaganda Corpus

In the first phase, we were provided with a list of processed web document annotations prepared by experts. We have downloaded all the referenced documents and extracted, using templates specific for each of the four servers, the texts of the articles and additional metadata – the document title, text of the leading article, date of publication, and, where possible, the author of the document.

Contrary to our expectations, the format of the pages is not stable in time and the templates needed to be updated for every new batch of articles. The changes consisted of different nesting of the elements and their identifiers, some of the text could not be discovered using the old template and conversely, some unwanted boilerplate would be included.

The extracted text was then tokenized using *unitok* [11] and morphologically annotated using *majka* [12] and *desamb* [13].

5 The Annotation Editor

The editor is designed as a server-client application ready-to-use without installation via standard web browser. The server part is written using a Python web library `bottle.py`¹, the client part uses the `jQuery`² and `Bootstrap`³ libraries.

¹ `bottlepy.org`

² `jquery.org`

³ `bootstrap.org`

The screenshot shows a search engine interface with a dark header containing the text 'Propaganda' and a 'test' button. Below the header, the search results are displayed under the heading 'Výsledek vyhledávání'. The search query is 'Dotaz Putin' and it indicates that 100 of the first 528 results from 5322 documents are shown. The results are presented in a list of seven items, each with a blue link and a snippet of text:

- 'Jak může garantovat bezpečnost režim, kde mizí novináři?', 'Pane ministře, nechte toho!' Stropnický a Filip se v ČT tak hádali, že je musel roztrhávat Moravec
- Naštvaná Ukrajina a zvěsti o novém Majdanu. Moldavsko žije volbami
- 'Sedí prase na Hradě, libuje si v úřadě.' Co chcete od společnosti, která tomu tleská! Čechokanaďan Jírovec ukazuje v zásadní věci na Schwarzenberga
- 'Systém EET zkolabuje a přivodí Babišovi porážku!' uvedl v rádiu senátor Valenta. Promluvil také o ParlamentníchListech.cz
- Válka na Krymu, pravda o Savčenkové, tajnosti kolem malajsijského letadla. Pozorovatel přináší žhavé novinky z Ukrajiny
- Víme, co marketingoví odborníci říkají o ruském dokumentu Prezident v hlavní roli s Vladimírem Putinem
- Volby v USA: Co jste nevěděli o pánovi od Hillary Clintonové. Tu prý platí velké korporace, Trumpa drobní střádalové. No a Putin...

Fig. 2. Example search result

5.1 The Server Side

The documents are stored in so-called vertical format⁴ where the main structure “<doc>” contains the structure attributes corresponding to the annotation scheme. Only a few extra attributes are processed automatically and added to the metadata, namely the second level web domain, top level domain, document word count, original URL and the date of publication (if available on the webpage).

The data is indexed by the modular corpus manager *manatee* [14] which allows fast full-text search within the documents.

The annotation data which comes from the client during manual annotation is stored in a separate SQLite database. The document and range attribute annotations are stored separately.

The HTTP API is maintained by a Python HTTP server provided by the *bottle* library. It can serve either JSON outputs or HTML responses generated by a template engine which is a part of the library.

⁴ <https://www.sketchengine.co.uk/documentation/preparing-corpus-text/>

ak ať drží hubu. A další výroky z drsné

u. A další výroky z drsné politické besedy

granty a příležitost pro čečenskou vládu, ukrajinské setkání příznivců a členů Úsvitu – národní koalice a Bloku / desítky lidí, aby si mimo jiné poslechly i bezpečnostního

Idát na hejtmana Plzeňského kraje **Roman Bakalá**, v tebný, protože jsou tam dvě firmy, které dělají halal čné názvy ulic. Pokud menšina přesáhne deset procent v mě týká Slezska, kde je velká polská komunita, ale když si dočkáme toho, že některá náměstí a ulice budou možná oval místní přítomné, za což si zasloužil hlasitý šum v sále.

: „Zažil jsem islamizaci v Iráku a Kuvajtu. V Kuvajtu chodily ních. Dnes už to není možné. Prakticky je situace taková, trstva, jejich salaafistické Ideologie a výsledek je takový, že echny příkazy Koránu, i když některé příkazy nejsou ani ladu v Kuvajtu došlo k tomu, že jeden profesor islámského

Atributy s rozsahem	
Místo	Ceska republika ▼
Vina	ano ▼
Nálepkování	ano ▼
Argumentace	ano ▼
Obsažené emoce	strach ▼
Démonizace	ano ▼
Relativizace	ne ▼
Strach	ano ▼
Fabulace	ano ▼
Rusko	pozitivni priklad ▼
Žánr	zpravodajstvi ▼
Odborník	ano ▼
Politik 1	<input type="text"/>
Vyznění 1	pozitivni ▼
Politik 2	<input type="text"/>
Vyznění 2	negativni ▼
Politik 3	<input type="text"/>
Vyznění 3	negativni ▼
Atributy dokumentu	
Téma	migracni krize ▼
Zaměření	domaci ▼
Názor	ne ▼

Fig. 3. Annotation of range attributes, the highlighting

5.2 The Client Side

The client serves as the central point of all operations related to the processing of the corpus data:

- 1) basic and advanced full-text search,
- 2) viewing of documents with their metadata, and
- 3) annotation of document and range attributes.

The simple search operation accessible on every page (in the top bar) can be used for searching a word or a phrase. What is important is that not just the exact word or phrase is searched for but also their other word forms. E.g. for query “*Putin*” the system will retrieve documents containing the word in other cases (useful especially for morphologically-rich languages as Czech).

The advanced search allows to combine this full-text search with querying the metadata annotated in the documents. Each attribute from the annotation schema can be used together with a particular value which narrows the results of such a search. Figure 1 shows an example advanced query specification with two attributes “*Argumentace*”—*argumentation* and “*Obsažené emoce*”—*present emotions* with two particular values “*ano*”—*yes* and “*rozhořčení*”—*outrage*, respectively. An example of search results is shown in Figure 2.

Any document from the corpus (accessible through the search results or from a preassigned list denoted by the user name) can be viewed together with its annotations. When viewing the document, it is possible to directly annotate both range and document attributes. Once logged in (the editor does not allow anonymous editors to save any annotations), the document attributes can be

Neméně šokující bylo další Kerryho prohlášení ^{✖ politik2: John Kerry} Kerry v němž uvedl, že do voleb v Sýrii by měl být zapojen současný ^{✖ politik1: Bašár Asad} prezident Bašár Asad, a to navzdory skutečnosti, že oficiální ^{✖ politik3: Barack Obama} postoj Obamy s něčím takovým nepočítá – Bílý dům trvá na odstoupení syrského vůdce.

Fig. 4. Visualization of annotated ranges in documents.

annotated by selecting them in the right hand table containing all attributes and their values (together with the immutable attributes like URL, date etc.).

The range attributes can be annotated once a phrase selection (range) is highlighted: by clicking on first word of a sequence and then on the last word of the sequence, the range is highlighted and a value can be set to a particular attribute in the highlighted part of the metadata table. An example of this operation can be seen in Figure 3.

When viewing documents, badges are shown next to the annotated attributes and a user can open the appropriate ranges with values by clicking the badges. The annotated ranges can be easily removed (see Figure 4).

6 Future Development and Conclusions

With the first version of the editor developed, the new (and possibly also the previous annotations) can express the manipulative features with precise references on word level which will provide valuable data for a) exact specification of the particular subtasks, such as the identification of the propagandistic phenomena in texts, and b) training automatic methods: features from attributes, their values and words within annotated ranges will be extracted and used within selected machine learning techniques. This will allow the next version of the system to be able to pre-annotate the document-level and word-level attributes so annotators have them at the disposal to verify and/or amend them.

Acknowledgements. This project was partially supported by the Grant Agency of Masaryk University within the project MUNI/G/0872/2016.

References

1. Castillo, C., Davison, B.D., et al.: Adversarial web search. *Foundations and Trends in Information Retrieval* 4(5) (2011) 377–486
2. Metaxas, P.T.: Web spam, social propaganda and the evolution of search engine rankings. In: *International Conference on Web Information Systems and Technologies*, Springer (2009) 170–182

3. Lumezanu, C., Feamster, N., Klein, H.: # bias: Measuring the tweeting behavior of propagandists. In: Sixth International AAAI Conference on Weblogs and Social Media. (2012)
4. Weedon, J., Nuland, W., Stamos, A.: Information operations and Facebook. Facebook on-line report (April 27, 2017)
5. Wakabayashi, D., Isaac, M.: In Race Against Fake News, Google and Facebook Stroll to the Starting Line. *The New York Times* 4 (2017)
6. Herman, E.S., Chomsky, N.: A propaganda model. *Manufacturing Consent: The Political Economy of the Mass Media* (1988)
7. Lee, A., Lee, E.B.: *The fine art of propaganda*. (1939)
8. Holz, F., Teresniak, S.: Towards automatic detection and tracking of topic change. *Computational linguistics and intelligent text processing* (2010) 327–339
9. Vosoughi, S.: Automatic detection and verification of rumors on Twitter. PhD thesis, Massachusetts Institute of Technology (2015)
10. Janze, C., Risius, M.: Automatic detection of fake news on social media platforms. (2017)
11. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In Horák, A., Rychlý, P., eds.: RASLAN 2014, Brno, Czech Republic, Tribun EU (2014) 71–75
12. Šmerk, P.: Fast Morphological Analysis of Czech. In Sojka, P., Horák, A., eds.: Third Workshop on Recent Advances in Slavonic Natural Language Processing, Masarykova univerzita, Masaryk University (2009) 13–16
13. Šmerk, P.: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). PhD thesis, Faculty of Informatics, Masaryk University (2010)
14. Rychlý, P.: Manatee/Bonito – a modular corpus manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. (2007) 65–70

Subject Index

- Aranea Project 97
- classification 71
- corpora 111
 - web-based 97
- Czech 51,59,71, 97,107
- dictionary 11
- e-lexicography 11,59
 - English 51
- geographic text retrieval 85
 - geonames 85
- idioms 59
- intensional essentialism 37
- intersective 37
- invoice 71
- Japanese 11
- language identification 97
- language model 107
- lemmatization 21
- Lemon model 11
- Linked Data 11
- logical analysis 51
- manipulative techniques 111
- morphological tagging 107
- morphosyntactic annotation 97
- natural language semantics 51
- OCR 71
- ontology-based geoname relations 85
- ordered-triple theory 29
- PoS tagging 107
- privative 37
- procedural grammar 29
- propaganda 111
- property modifier 37
- QA dataset 79
- question answering 79
- recognition 71
- rule of pseudo-detachment 37
- Slovak 97
- SQAD 79
- subsective 37
- syntax 29
- text mining 21
- toponym disambiguation 85
- toponym resolution 85
- toponym similarity 85
- transparent intensional logic 29, 37,51
- verb valency 59
- VerbaLex 59
- Wikipedia 21

Author Index

- Baisa, V. 111
Benko, V. 97
- Calimeri, F. 85
- Duží, M. 37
- Fait, M. 37
- Ha, H. T. 71
Herman, O. 111
Horák, A. 3, 29, 51, 79, 111
- Klement, D. 3
Kletečka, J. 3
- Lecailliez, L. 11
- Medved', M. 51, 79
Mrkývka, V. 21
- Nevěřilová, Z. 59
- Pala, K. 29
- Rambousek, A. 3
Rychlý, P. 107
- Sherwani, M. K. 85
Sojka, P. 85
- Šulganová, T. 51, 79

RASLAN 2017

Eleventh Workshop on Recent Advances in Slavonic Natural Language Processing

Editors: Aleš Horák, Pavel Rychlý, Adam Rambousek

Typesetting: Adam Rambousek

Cover design: Petr Sojka

Published by Tribun EU

Cejl 32, 602 00 Brno, Czech Republic

First edition at Tribun EU

Brno 2017

ISBN 978-80-263-1340-3

ISSN 2336-4289