What is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using Termania website

Mojca Kompara*, Peter Holozan‡
*University of Primorska, Faculty of Humanities
Titov trg 5, 6000 Koper, Slovenia
‡Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik, Slovenia
E-mail: mokopt@yahoo.com, peter.holozan@amebis.si

Abstract

The paper presents what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site. At a first step an algorithm for the automatic recognition of abbreviation-expansion pairs in newspaper texts has been used. After manual cleaning genuine pairs obtained with the algorithm were automatically included in *Termania* editing software. The editing process was automatised, non-nominal expansions were converted to nominal and language qualifiers were added automatically. The precision of the algorithm for recognition of abbreviation-expansion pairs is 96%, the algorithm for lemmatisation is successful in 70.9% of cases, the algorithm for adding language qualifiers in 95.9% of cases. Out of the obtained results we can say that simple entries are produced automatically, whereas complex "semi" automatically, because translations or explanations of abbreviation-expansion pairs can be added, at the present stage, just manually. The same goes for encyclopaedic data. For that reason we can say that the first Slovene online dictionary of abbreviations is produced "semi" automatically but further attempts to automatize translations and encyclopaedic data will be done in the future.

Keywords: abbreviations; dictionary; automatic production of dictionary entries

1. Introduction

Abbreviations are difficult to deal with (Gabrovšek, 1994) and represent a growing phenomenon present in all languages. The scope of this article is to present what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site.

2. The beginning

Kazem Taghva (1998) is the pioneer in automatic recognition of abbreviations and abbreviation's expansion. Automatic recognition of abbreviations was dealt also by Yeast (1999), Larkey *et al.* (2000), Pustejovsky *et al.* (2001), Schwartz and Hearst (2003), Park and Byrd (2001), Chang *et al.* (2002) and Zahariev (2004), his approach is considered special due to the fact that he is not limiting just to one language recognitions. In the early stage of the research words up to 5 capital letters written in brackets were used as abbreviations' candidates, e.g. *(NATO)*. Words of five letters with the first letter capitalised, e.g. *(Mig)* were also used. Symbols and abbreviations such as *itd., npr., ipd., itn.* etc. were not included. In the fist stage the reference was the Slovene online newspaper *Delo* from 2007. *Delo* had 25,588 such candidates and some occurred more than once. In order to come across a genuine amount of candidates for abbreviations words that are not abbreviations, such as proper names, names of places etc. were excluded, using the *Slovene monolingual dictionary*, after the exclusions the database had over 2,500 candidates. The second step covered candidates for expansions. In order to obtain the abbreviations' expansions from newspaper *Delo,* left context was observed, because expansions are usually placed before the abbreviations, e.g. *European central bank (ECB),* but still not excluding the possibility of right context. A genuine abbreviation is determined by the expansion/s and is divided into official and/or non-official. An abbreviation can have several expansions and the algorithm took into consideration all the expansions an abbreviation could have. To recognize expansions 4 types of abbreviations were used. The first type are the so called *covered abbreviations* where letters match the words in left context, e.g. *FF* with the expansion *Filozofska fakulteta*, *Mig* with the expansion *Mesna industrija Goriške*. The second type are *abbreviations with expansions containing prepositions and conjunctions*, e.g. *FDV Fakulteta za družbene vede*. The algorithm takes into consideration also expansions with one additional word e.g. *za*. The third type concerns *abbreviations composed of the first two letters*, e.g. *NAMA Narodni magazin*. The fourth type covers *abbreviations with prepositions*, e.g. *DZU Družba za upravljanje* where prepositions appear in the abbreviation and also in the expansion. In the final list containing abbreviation-expansion pairs several problems were observed, such as the occurrence of cases, the multiple occurrences of the same expansion and abbreviations without expansions. Abbreviations without the matching expansion in the text were automatically deleted. In the manual revision that followed the most neutral case was preserved and all identical pairs appearing more than once were deleted. Considering the above mentioned criteria 1,800 expansions matched the abbreviations and formed genuine abbreviation-expansion pairs.

3. The development

In the second stage of the development the number of letters in the abbreviation was extended to 10 left and right contexts were observed. All four types of pattern: *(abbreviation) expansion, (expansion) abbreviation, abbreviation (expansion), expansion (abbreviation)* were used. Abbreviations with more than 10 words were not included. Abbreviations identical to legal words (lexicalized abbreviations), e.g. *Nama, Kad, Sod* etc. are very common in Slovene. *Nama* can be both an acronym for *Narodni magazin* or a personal pronoun at the beginning of a sentence. Such abbreviations are usually well-known but problematic and misleading for the algorithm. They were included in the algorithm rules via the dictionary of abbreviations *Slovarček krajšav*. After the newly established rules for recognition a demo version of the algorithm was produced. The system called *MKstrings* is composed of two windows, in the first one we add text rich in abbreviations, after clicking *Click here to process data* in the second window abbreviations and expansions occur as seen from Figure 1.

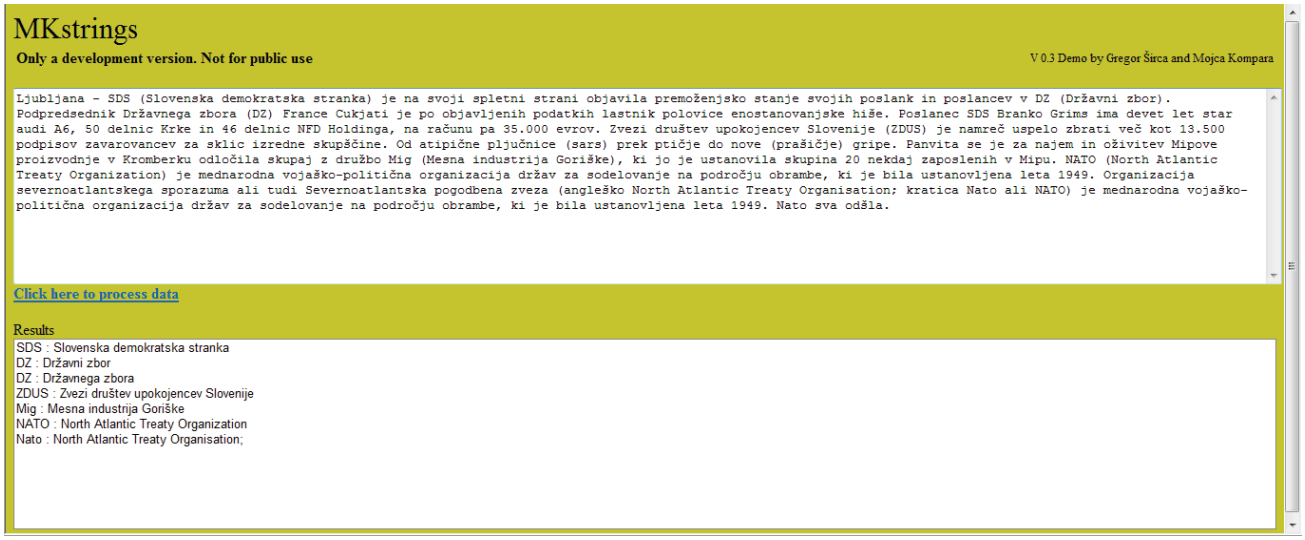

Figure 1: Demo version of the algorithm

As seen from Figure 1, the algorithm is not taking into consideration abbreviations such as *sars, Mipu, NFD, A6*, which is expected according to the rules stated above. Although at first sight the obtained results look really well, the algorithm was improved. Randomly selected texts rich in abbreviations (from the website *24ur.com*) were used in order to observe how the algorithm behaves. Problems occurred mainly in examples containing the abbreviation *e.g. RS* in the expansion. But also after taking into consideration this step the problem was still not solved. Prepositions *za* and *v*, represented a problem too, because at the present stage the algorithm was able to consider just one preposition or additional word in the expansion. Problems occurred also in some copy-pasted examples e.g. *Urada za varstvo konkurence (UVK)*, recognized when retyped. An interesting issue are also patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)*. Such patterns were not observed in the present article and will be recognized in the future. After applying modifications and improvements the software was enlarged in order to be able to filter larger amounts of data. A larger corpus composed of 60 million words (newspaper *Delo* from 2005 to 2009) was used. The algorithm filtered the corpus in 30 minutes and gave 5,820 abbreviation-expansion pairs. The obtained pairs were manually revised and verified using *Google*. The precision of the algorithm is 96%. Recall was not retrieved because the corpus was the whole newspaper and not just a corpus made of just texts with abbreviations. Manual check up of the whole corpus would take too much time and a smaller sample would not show the real situation in the text. Among the good expansions many occurred more than once and/or with tiny modifications, e.g. usage of different cases or spelling as seen in Table 1. In Table 1 just 3 expansions out of 6 are genuine but also the genuine ones are not lemmatised (provided in nominal form) and can not be included in a dictionary as such.

MNZ	MNZ
1 ministrstva za notranje zadeve 2 medobčinskih nogometnih zvez 3 ministrstvom za notranje zadeve 4 Medobčinske nogometne zveze 5 Muzeja novejše zgodovine 6 Muzej novejše zgodovine	1 ministrstva za notranje zadeve 2 medobčinskih nogometnih zvez 3 Muzej novejše zgodovine

Table 1: Good expansions for MNZ entry

After the exclusion of false pairs (4%), verification and revision of good pairs, 2,665 genuine abbreviations-expansion pairs (in different cases) occurred. Among the good pairs there were also some foreign pairs although the recognition focused only on Slovene texts. Among the foreign some problems occurred in misrecognition of parts of expansions, e.g. *FEE for Environmental Education*, where *Foundation* is missing.

4. Termania editing software

After the genuine abbreviation-expansion pairs were obtained from the corpus using the algorithm, the next step was to include the pairs automatically into a dictionary editing software and edit the entries automatically or semi automatically. *Termania*[1] was used as exiting software. It is a free on-line dictionary portal with integrated dictionary browsing and editing tools developed by *Amebis* software company (Kamnik, Slovenia) in cooperation with *Trojina*, the Institute for Applied Slovene Studies. It provides an interface for dictionary browsing and a simple but reasonably versatile on-line dictionary editing tool. The portal is intended for general public users with no specialized computer or lexicographic knowledge, but with an interest in sharing terminological or general language knowledge, either by offering translations in a bilingual or multilingual environment or providing definitions in a monolingual context. The portal is intended to serve as the central terminology data and opinion exchange node for Slovene terminology. Access is free of charge, but does require registration (Krek, 2011). After the automatic inclusion of genuine pairs in *Termania* the editing phase started. The aim of the project was to provide an automatic editing process in simple and complex entries of the first Slovene online dictionary of abbreviations.

5. Conversion of expansions into nominal form

The main problem in automatic production of simple and complex dictionary entries are abbreviations' expansions that appear in non-nominative cases. Slovene has six cases: nominative, genitive, dative, accusative, locative and instrumental, but for dictionary purpose only nominative case is used. Another problem is the number, which can be singular, plural or dual, as well as the occurrence of other languages.

5.1 Presis Interlingua

Presis[2] is a machine translation program developed by Amebis. It supports Slovene to English, English to Slovene and German to Slovene translations and it is part of iTranslate4.eu project[3].

It is a rule-based system consisting of analyzers and generators. Analyzers translate text in Slovene, English or German to Presis Interlingua, while generators translate Presis Interlingua to Slovene or English (German to English translation is also possible, but it is used just for testing and it is not commercially available).

5.1.1 Interlingua sample

Presis Interlingua for the expansion *Plesne zveze Slovenije* is as follows:

(-POV:(-STC:(-PR2:(-SFR:(-DSF:(-PFR:(-DPF:(-PRVo :{4b0199;21f42f1}[0]<2044>))),(-JED:(-SAMe:{8d42;3 09123d,309128d,dfccf8,2c9be,1ba3a0d,195d56f,2dc88fe ,30912e1}[1]<111c>)),(-SFR:(-DSF:(-JED:(-SAMe:{91 47;28ec956}[2]<c508>)))))))))

Each element is in parentheses. It starts with status character (in this case always '-' for sentences, these are used to connect parts of a sentence to a verb template), then there is a three-letter name of element followed by colon (there can be some parameters before the colon, e.g. element *SAM* (noun) has information about number (e – *ednina* (singular)).

For complex elements there is a list of included elements, simple elements (elements connected to actual words) have the following information: inside {} is first the ID of lemma followed by semicolon which is followed by a list of IDs of possible meanings (senses) separated by commas. Inside [] is an index of a word in the original text and inside <> is the ID of morphosyntactic descriptor. All IDs are in hexadecimal.

A detailed description of Presis Interlingua and all its elements can be found in Holozan (2011). Table 2 lists just elements from the sample.

[1] http://www.termania.net/

[2] http://presis.amebis.si

[3] http://itranslate4.eu

POV	sentence
STC	single sentence element
OSB	subject
PR2	object in genitive
PRD	direct object (in English)
SFR	noun phrase
DSF	part of noun phrase
PFR	adjectival phrase
DPF	part of adjectival phrase
PRV	adjective
JED	centre of noun phrase
SAM	noun

Table 2: Some elements of Presis Interlingua.

5.2 The procedure

The idea is to use Presis Slovene analyzer to translate the expansion in question into Presis Interlingua. A special version of the analyzer, which allows only subjects and objects in various cases, is used.

If the result is a subject, it is already in nominative case and no further work is needed. If the result is an object, the translation Interlingua is changed and the object becomes the subject (e.g. *(-PR2* is changed to *(-OSB*) and then fed to the Slovene Presis generator). The resulting Slovene "translation" is the same noun phrase in nominative case.

An important part of this procedure is the meaning information that has to be removed from the translation Interlingua before sending it to the generator otherwise some words in generated nominative form can become replaced by synonyms.

5.3 Results

The expansions were processed using the above mentioned procedure and then the results were manually corrected.

A program was made to count the number of differences between automatic results and manually corrected ones. The first test showed 2531 correct expansions and 655 mistakes. Without the conversion to nominal form, the number of mistakes was 1486, in this way conversion solved 55.9% of cases (and not all mistakes were due to wrong cases, so the result is even better).

5.4 Problems

Most of the problems found can be assigned to one of the following categories.

5.4.1 Number

There are quite a few problems in the cases where nominative plural form is chosen instead of some non-nominative singular form. However, if the program changed all plurals into singulars, new problems appeared where the required expansions really should be in plural e.g. *konvencionalne sile v Evropi*, *cestno prometni predpisi*, so the final result is worse than it was without this change.

5.4.2 Definite forms of adjectives

In some cases an indefinite form of adjectives was generated instead of the definite e.g. *nacionalen energetski program*. The problem is that in Slovene definiteness is only present in nominative and accusative of singular masculine forms, for all other cases the definite and indefinite form is the same. However, for abbreviation's expansions included in a dictionary entry it is essential to use only definite forms. For that reason definiteness was added to all adjectives in Interlingua (in the form of definite article, which doesn't exist in Slovene, but forces definite forms in Slovenian generator). This update solved 27 cases.

5.4.3 Capitalisation

Capitalisation represents a problematic issue. On one hand, there are many expansions, which should be capitalized, on the other hand, even more expansions are not capitalized e.g. *socialno varstveni center*, *obnovljiv vir elektrike*, *indeks telesne mase*, and for that reason the test for automatic capitalization of all the expansions has worsened the results.

In the second run, 16 typical beginnings to be capitalized were added to the program (e.g. *zakon* (act), *fakulteta* (faculty), *slovenski* (Slovene)) and it solved 117 cases.

5.4.4 Wrong adjective/noun disambiguation

In some cases e.g. *Slovenskega ljudskega gledališča*, *Slovenskega svetovnega kongresa*, *Primorskega poletnega festivala* the analyzer made a mistake with disambiguation in considering the first word as a (proper) noun instead of an adjective. The analyzer was updated and now adds penalty for disambiguation for noun phrases where a genitive noun phrase follows proper noun. This update solved 24 cases.

5.4.5 Doubling of results

Conversion of expansions into nominal forms caused doubled expansions in some cases. For the abbreviation *RS*, the following expansions (alongside 5 others) were found and then converted to nominal form:

Računskim sodiščem	Računsko sodišče
Računsko sodišče	Računsko sodišče
računskega sodišča	računsko sodišče

Table 3: Nominal forms for the abbreviation RS.

For the abbreviation *RP*, the expansions were (alongside 3 others):

razdelilnih postaj	razdelilne postaje
razdelilno postajo	razdelilna postaja

Table 4: Nominal forms for the abbreviation RP.

In this case, *razdelilne postaje* is the plural form of *razdelilna postaja*.

5.5 Improved results

After the applied changes for conversion into nominal form and some corrections in manually corrected results (the mistakes were noticed during testing), the result is: 2661 correct expansions, 433 mistakes. Conversion solved 70,9% of the cases (compared to tests without conversion). The majority of the remaining mistakes concerns capitalization and number, but there are also some problems with unknown words.

6. Language identification

Some abbreviations have expansions in other languages, not just in Slovene. Statistical methods are often used for determining the language of a text (Dunning, 1994). However, these methods do not work well on very short text (less than 20 characters) and expansions are generally very short. We can assume quite safe, that if Presis analyser can produce the noun phrase analysis for the given text in some language, this text is in that language (assuming the number of unknown and guessed words do not exceed some percentage of text). Presis has analysers for English and German besides Slovene and this covers most of the expansions.

6.1 The procedure

In the first stage the expansion is sent to the Slovene analyzer. If the analyzer is successful, the language code "sl" is assigned; if not the language code "sl-x" is assigned (Slovene – to be manually checked).

In cases where there is still no code the English analyzer is applied and if it is successful, the language code "en" is assigned, if not the language code "en-x" is assigned (English – to be manually checked).

In the final step, if there is still no code, the German analyzer is applied. If it is successful, the language code "de" is assigned; if not the language code "de-x" is assigned (German – to be manually checked).

If no analyzer was successful and all found unknown words, the language code "xx" (unknown – to be checked and assigned manually) is assigned.

6.2 Results

Table 5 shows the required manual corrections of language tags with some examples.

xx→en	28	Codex Alimentarius Commission Bharatiya Janata Party Eco Management and Audit Sheme
xx→sl	24	Državnem inštitutu za fizkulturo Adriatica Slovenice Slovenske narodne podporne jednote
xx→fr	12	Federation Internationale des Echecs
xx→it	10	Associazione Sportiva
xx→de	6	Sport Events Steinforth Financial Times Deutschland
xx→sr	3	Crveni zvezdi
xx→hr	3	Istarsko narodno kazalište
sl→en	5	Super Proton Synchrotron New York Times John Fitzgerald Kennedy
sl→hr	3	Hrvatska stranka prava
sl→sr	3	Duvanska industrija Niš
sl→it	1	Parti socialiste
en→de	5	Deutsche Bank
en→fr	5	Electricité de France

Table 5: Corrections of language tags.

Out of 3094 language tags, 126 were assigned wrong. 95.9% of selected tags were correct. In Table 5, the final numbers of expansions for each language are shown.

Slovene and English made up most of the cases; French might be interesting to be added to the language identification since there already exists alpha version of analyzer for French for Presis.

For Croatian, Serbian, Bosnian and Montenegrin it will be difficult to make automatic language identification because of their similarity.

Slovene (sl)	2335
English (en)	680
German (de)	23
French (fr)	18
Italian (it)	11
Croatian (hr)	7
Serbian (sr)	7
others (es, ro, bs, eu, hu, ne, pl, mno)	13

Table 6: Number of expansions by language.

One problem that should be solved in future versions concerns foreign names included in Slovene dictionary e.g. *John Fitzgerald Kennedy*. One possibility is to try both Slovene and English analyzer and if they both produce result, the program should decide which is more probable.

7. Entries in Termania

After applying above-mentioned changes the first Slovene online dictionary of abbreviations was available online free of charge on *Termania's* web site. The entries in *Termania* can be divided into simple and complex. Simple are those composed of a Slovene pair *e.g. FF, Filozofska fakulteta* where language is provided and the expansion is in nominative case. At present the algorithm provides such entries entirely automatically, as seen from Figure 2. Such entries work perfectly well in a Slovene dictionary of abbreviations, the only embellishment we could add to such entry is some encyclopaedic data. Such data give some additional information to the user but are not essential in simple dictionary entries and can easily be omitted.

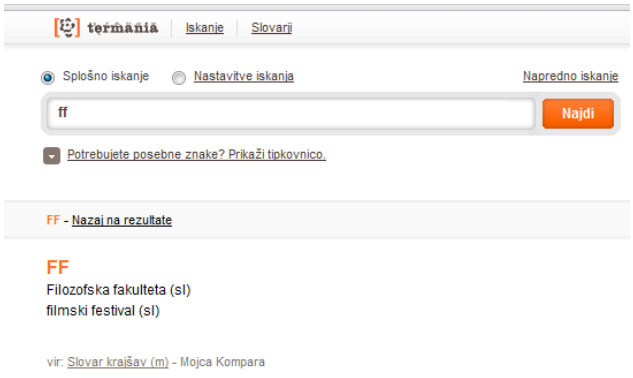

Figure 2: Simple dictionary entry

Complex entries are those containing foreign abbreviations where language is provided and the expansion is checked, but the Slovene translation is missing, as seen from Figure 3. A user friendly dictionary of abbreviations should include translations of foreign abbreviations and also some encyclopaedic data or a description, if there is no official translation. For now translations, descriptions and encyclopaedic data can be included only manually and for that reason such complex entries are not produced automatically, as are the simple ones. In the future we will make an attempt to provide automatic translations, descriptions and encyclopaedic data in complex entries.

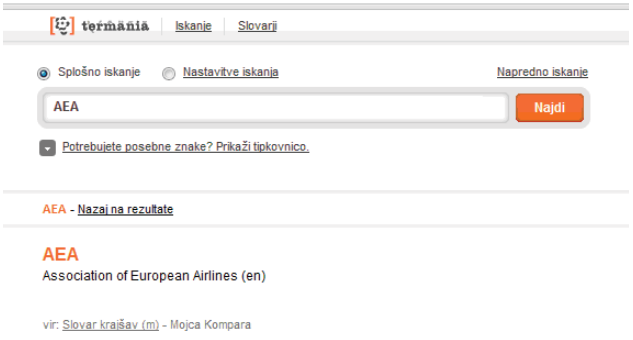

Figure 3: Complex dictionary entry

8. The future

In the future we will try to cope with new challenges, the first one being automatic recognition of translations or automatic translation of foreign expansions into Slovene. We will also focus on patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)* and try to recognise them out of the text. In such cases lexical recognition will not work and will be switched with statistical recognition. Such patterns are frequent in newspaper texts and were not observed in the present article but will be dealt with in the near future. The last automatised procedure will be done on encyclopaedic data. Not all abbreviations will be provided with encyclopaedic data, but a selection will be made. The input of lexicographers will be essential in all phases of our dictionary development and the automatised process is used just to help the lexicographer and could never replace them.

9. Conclusion

The answer to the question what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site is: algorithms, a good knowledge of the topic and fresh ideas. In the paper we present how we automatically extract abbreviation-expansion pairs out of newspaper texts, how we clean manually and obtain genuine pairs, how we cope with the automatic editing phase and add language qualifiers to expansions and transform non-nominative expansions into nominative. We placed our work online, free of charge, on the web site of *Termania* editing software in order to share it with the users. It is the first dictionary that is produced "semi"[4] automatically from newspaper articles with the help of algorithms. Algorithms give the possibility to create a semi automatic dictionary of abbreviations and such dictionary represents the future of electronic lexicography. Algorithms for automatic recognition of abbreviations, lemmatization and adding language qualifiers present a link between the text and the semi automatic production of a dictionary of abbreviations. That is why the production and further development of the algorithm is essential and useful for lexicographers.

10. References

24ur.com. Accessed at: http://24ur.com.

Amebis Termania. Accessed at: http://www.termania.net.

Chang, J.T., Schütze, H. & Altman, R.B. (2008). Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of American Medical Informatics Association (JAMIA)*, IX(VI), pp. 612-620.

Dunning, T. (1994). Statistical Identification of Language. Technical Report MCCS 94-273. New Mexico State University, USA.

Gabrovšek, D. (1994). Kodifikacija angleškega jezika v

[4] Some procedures were done manually, such as selection of genuine pairs and cannot be automatised for now.

specializiranih enojezičnih slovarjih: Too much of everything?. (Codification of Englih in specialised monolingual dictionaries). *Vestnik*, XXVII/I-II, pp. 150-180.

Google.si. Accessed at: http://www.google.si.

Holozan, P. (2011). Samodejno izdelovanje besedilnih logičnih nalog v slovenščini (Automatic generation of textual logic puzzles in Slovenian). Master of Science thesis. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana, Slovenia.

Krek, S. (2011). Termania – Free On-Line Dictionary Portal. Retrieved from: http://www.simonkrek.si/Objave/5Krek_SD011.pdf.

Larkey, L., Ogilvie, P., Price, M.A. & Tamilio, B. (2000). Acrophile: An Automated Acronym Extractor and Server. In *Proceedings of the fifth ACM conference on Digital libraries 2000, San Antonio, 2-7 June 2000*. San Antonio, Texas, USA.

Park, Y., Byrd, R.J. (2001). Hybrid Text Mining for Finding Abbreviations and Their Definitions. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing.* pp. 126-133.

Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M. & Morrell, M. (2001). Extraction and Disambiguation of Acronym-Meaning Pairs in MEDLINE. *Medinfo*, 2001, 10, pp. 371-375.

Schwartz, A.S., Hearst. M.A (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing, 3-7 January 2003.* Lihue, Hawaii, USA.

Slovarček krajšav – Dictionary of abbreviations. Accessed at: http://bos.zrc-sazu.si/kratice.html.

Taghva, K., Gilbreth, J. (1998). Recognizing acronyms and their definitions. *IJDAR*, I(IV), pp. 191-198.

Yeates, S. (1999). Automatic extraction of acronyms from text. In *Proceedings of the Third New Zealand Computer Science Research Students' Conference, 6 April - Friday 9 April 1999.* University of Waikato, New Zealand.

Zahariev, M. (2004). A (Acronyms). PhD thesis. School of Computing Science, Simon Fraser University, Ottawa, Canada.