TERMIS: A corpus-driven approach to compiling an e-dictionary of terminology

Nataša Logar[1], Iztok Kosem[2]

[1]University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
[2]Trojina, Institute for Applied Slovene Studies, Škofja Loka, Slovenia
E-mail: natasa.logar@fdv.uni-lj.si, iztok.kosem@trojina.si

Abstract

This paper describes the process of compiling an online dictionary of terminology within the TERMIS project. The compilation began from a morphosyntactically tagged synchronous LSP corpus and involved automatic term recognition performed for single- and multi-word terms with the LUIZ term extraction system and the automatic extraction of lexical information from the corpus via the Sketch Engine tool. The information obtained, along with the results of the GDEX system, was imported into the dictionary editing system to the Termania web portal. A free online terminological database of the public relations field comprised of 2000 entries has been publicly available since July 2013.

Keywords: terminology, corpus, database, public relations, Termania

1. Introduction

Due to the continuous growth of scientific research, all disciplines must assure the development of terminology in their own language. In the case of Slovene, terminology development is connected to the importance of native language. Several terminological dictionaries have been published in Slovenia in the last few decades; however, there still remains a need for terminology description in many different disciplines.

New challenges in terminology have arisen as a result of the Bologna Reform and system of internationalization of higher education that, among other things, promote frequent exchange between students, lecturers, and researchers (Kalin Golob & Stabej, 2007; Humar & Žagar Karer, 2010; Kalin Golob, 2012; Kalin Golob et al., 2012). Interpretation of internationalization in its narrow sense, i.e. an increase in the number of university programs taught in English, implies a resulting abandonment of Slovene as a language of instruction in higher education. As a result, there are now more and more warnings of such a practice turning into a situation in which "Slovene would eventually become a language in which some disciplines would no longer have, or would no longer develop, its own terms, and the communication would be conducted in a foreign language only" (Humar & Žagar Karer, 2010: 9).

One of the solutions to this problem is to provide Slovene terminology with contemporary reference materials, namely terminological dictionaries and databases. This paper describes the development of a terminological database within the TERMIS project, which consisted of six key phases: (a) a corpus, (b) automatic extraction of term candidates, (c) automatic extraction of collocations and grammatical relations, (d) extraction of good examples, (e) data editing and (f) final online visualization of entries.

2. TERMIS

An applied research project titled *Terminology data banks as the bodies of knowledge*: *The model for the systematization of terminologies* (TERMIS; http://www.termis.fdv.uni-lj.si/) was conducted between July 2011 and June 2013, funded by the Slovenian Research Agency. The aim of the project was the compilation of an online terminological dictionary of public relations, with two specific objectives:

a) The development of a freely accessible online dictionary-like terminological database for the discipline of public relations. The database contains 2000 terms with definitions, English translations, and typical collocations. Each entry is linked with a specialized corpus of public relations texts called KoRP (http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl; Logar, 2007) and Gigafida, a reference corpus of Slovene (http://www.gigafida.net; Logar Berginc et al., 2012).

b) The development of an online dictionary editing system that is easy to use so that an expert in the field, i.e. a terminologist, can start using it without any prior knowledge. Dictionary writing systems are freely available on the Termania online portal (http://www.termania.net; Romih & Krek, 2012; Kompara & Holozan, 2011: 145).

This paper focuses on the first objective only.

3. Corpus

The basis of the project was KoRP, a corpus of public relations texts. The corpus contains 1.8 million words and is a monolingual and synchronous specialised corpus. The corpus has been freely accessible online since it was completed in July 2007. Recently, the corpus was lemmatized and morphosyntactically tagged with the latest statistical tagger for Slovene, called Obeliks (http://oznacevalnik.slovenscina.eu/Vsebine/Sl/SpletniServis/SpletniServis.aspx; Grčar, Krek & Dobrovoljc, 2012). The texts in the KoRP corpus were selected according to carefully designed criteria (Logar, 2007), which make the corpus representative of a public relations field in Slovenia.

4. Term extraction

There are many approaches to extraction of term candidates from specialized corpora. Almost all of them use a combination of linguistic knowledge of terms, and mathematical statistics on word and word sequence distribution in corpora (Vintar, 2008: 100; Vintar, 2009: 346–347 and literature therein cited). Using the LUIZ term extraction tool (http://lojze.lugos.si/cgitest/extract.cgi; Vintar, 2010) we have extracted from the KoRP corpus:

a) single-word term candidates: nouns, verbs, adjectives, and adverbs;

b) multi-word term candidates: noun phrases and verb phrases.

Both single- and multi-word term candidates have been extracted using morphosyntactic patterns and term weight, calculated by comparing the frequency in the KoRP corpus and the frequency in a general corpus, in our case FidaPLUS, a reference corpus of Slovene (http://www.fidaplus.net; Arhar Holdt & Gorjanc, 2007), and phraseological stability of an extracted terminological unit. We have identified 39 morphosyntactic patterns in total: 30 with a noun as a headword, 9 with verb as a headword. The result of the extraction was lists with 47,007 multi-word units (excluding proper nouns) and 16,190 single-word units (excluding proper nouns).

The lists were carefully analyzed and evaluated in order to determine the successfulness of the extraction method. This highlighted two issues:

> a) When the top part of the list containing extracted term candidates was compared with the top parts of the noun and verb frequency lists in KoRP, we noticed only minor differences, but all in favour of the lists of extracted terms; in other words, the lists with extracted terms offered better results. Our expectations were thus confirmed, so we subsequently decided to use only automatically extracted lists of term candidates for building our headword list.
>
> b) The analysis of the top 100 units on the lists of all 30 multi-word patterns containing a noun headword showed that the terminologically most productive patterns were *Adj N*, *Adj and Adj N* and *Adj Adj N*. Over 50% of the analyzed extracted units in the lists were proper terms and thus relevant for our headword list (see Table 1).

We were able to obtain 2000 terms for the dictionary headword list by analyzing 3000 items on the lists containing single-word noun term candidates, and 4000 items on the list of multi-word term candidates (using all 30 patterns).[1] The analysis

[1] The extraction of adjectives, adverbs, and verb phrases did not yield terminologically relevant results, so they are not discussed in this paper.

was conducted by a terminologist and two experts in the field of public relations. In the next phase of the project, we automatically extracted lexical information for the words and multi-word units (e.g. compounds) on the created headword list.

Pattern	**Number of terms in the top 100 units on the list**	**Example**
Adj N	87	*blagovna znamka*
Adj and Adj N	62	*notranja in zunanja javnost*
Adj Adj N	45	*integrirano marketinško komuniciranje*
Adj N S N	20	*vladni odnosi z javnostmi*
Adj N and N	17	*strateško načrtovanje in upravljanje*
N Adj N	17	*upravljanje žgočih problemov*
R Adj N	11	*cenovno občutljiva informacija*
N S N	7	*odnosi z javnostmi*
N Adj Adj N	6	*model dvosmernega asimetričnega komuniciranja*
N N	6	*vir informacij*

Table 1: The 10 terminologically most productive patterns containing multi-word term candidates with a noun headword.

5. Automatic extraction of lexical information

When reporting on the compilation process of a new Lexical Database for Slovene (http://www.slovenscina.eu/spletni-slovar/leksikalna-baza; Gantar, 2009; Gantar & Krek, 2011), Kosem, Gantar & Krek (2012: 118) said:

> The decision to use automatic extraction of lexical information from the corpus /.../ comes from the need to reduce time and costs connected with the production of dictionaries, by utilizing new possibilities offered by state-of-the-art tools for corpus analysis.

Due to these very reasons, combined with the fact that we collaborated on the TERMIS project, as well as the *Communication in Slovene* project (http://www.slovenscina.eu/projekt), where this lexical description of contemporary Slovene has been produced, we used the method of Kosem, Gantar & Krek (2012) in our TERMIS project for extracting lexical information (syntactic relations, collocations, and examples) for single and multi-word terms from the KoRP corpus. The method uses the Sketch Engine tool and its Word sketch function (http://www.sketchengine.co.uk/; Kilgarriff et al., 2004; Kilgarriff & Kosem, 2012), so we had to prepare and upload the KoRP corpus in our local installation of the Sketch Engine. Due to the different nature of the project, and the corpus, some changes were necessary in the extraction algorithm and its constituent parts. For example, Sketch Grammar was slightly adapted (Krek, 2012), new GDEX (Good Dictionary Examples) configurations for good example extraction were prepared, and minor tweaks to API script (Application Programming Interface) were made (Kosem, Gantar & Krek, 2012; Kilgarriff et al., 2008; Kosem, Husak & McCarthy, 2011). In addition, a new DTD for the Termania dictionary portal was prepared to enable importing of information in the database, as well as its visualization.

After two test automatic extractions, we divided the terms into 10 different groups according to their frequency/salience values for relations for three groups of terms:

a) single-word terms:

- verbs:
    - group 0: frequency: 1–29
    - group 1: frequency: 30–199
    - group 2: frequency: >200
- nouns:
    - group 0: frequency: 1–19
    - group 1: frequency: 20–99
    - group 2: frequency: 100–699
    - group 3: frequency: >700

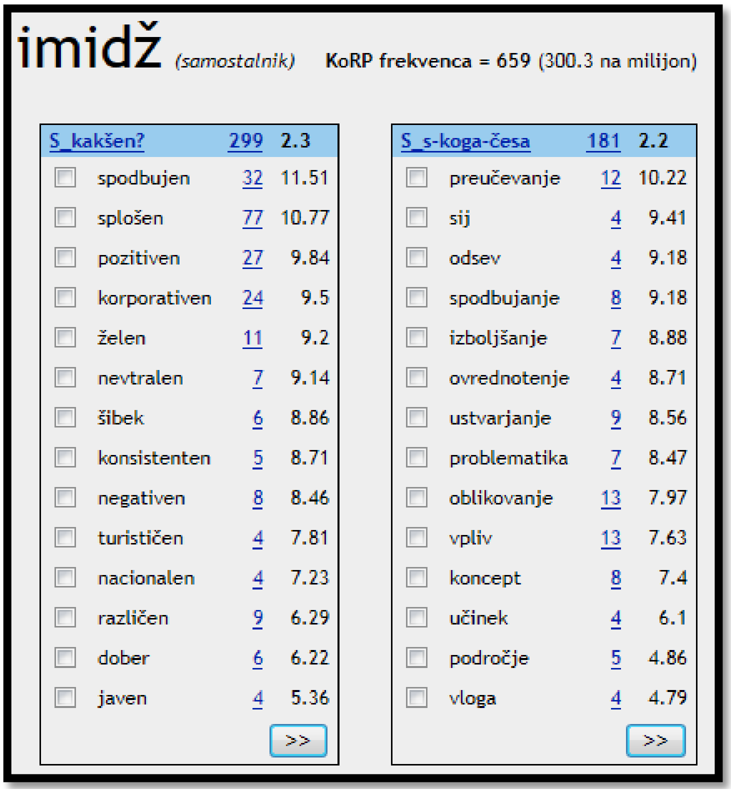
b) multi-word terms (adjective + noun, noun + noun):[2]

- group 0: frequency: 1–9
- group 1: frequency: 10–129
- group 2: frequency: >130

For terms in groups 0, all information available in word sketch was extracted. For other groups, we set four parameters for extraction (minimum collocation frequency, minimum collocation salience, minimum gramrel frequency, minimum gramrel salience) for each grammatical relation (example of settings is shown in Table 2, and an example of information they refer to is shown in Figure 1).[3]

[2] Automatic extraction of lexical information for other patterns, e.g. noun + preposition + noun, was not possible at the time.

[3] Explanation of values in Figure 1: top number in the second column indicates minimum gramrel frequency (e.g. 299 for the relation *S_kakšen?*), top number in the third column indicates minimum gramrel salience (e.g. 2.3), all the numbers in the second column indicate minimum collocation frequency (e.g. 32 for *spodbujen*), and all the numbers in the third column indicate minimum collocation salience (e.g. 11.51 for *spodbujen*).



imidž (samostalnik) KoRP frekvenca = 659 (300.3 na milijon)

	S_kakšen?	299	2.3
□	spodbujen	32	11.51
□	splošen	77	10.77
□	pozitiven	27	9.84
□	korporativen	24	9.5
□	želen	11	9.2
□	nevtralen	7	9.14
□	šibek	6	8.86
□	konsistenten	5	8.71
□	negativen	8	8.46
□	turističen	4	7.81
□	nacionalen	4	7.23
□	različen	9	6.29
□	dober	6	6.22
□	javen	4	5.36

	S_s-koga-česa	181	2.2
□	preučevanje	12	10.22
□	sij	4	9.41
□	odsev	4	9.18
□	spodbujanje	8	9.18
□	izboljšanje	7	8.88
□	ovrednotenje	4	8.71
□	ustvarjanje	9	8.56
□	problematika	7	8.47
□	oblikovanje	13	7.97
□	vpliv	13	7.63
□	koncept	8	7.4
□	učinek	4	6.1
□	področje	5	4.86
□	vloga	4	4.79

Figure 1: Partial word sketch for *imidž* in the KoRP corpus (the Sketch Engine).

It is worth emphasizing that we initially employed the same settings as our colleagues for compiling single-word noun and verb entries in the lexical database; however, the automatic extraction of lexical information for multi-word units (through MWU links in the Sketch Engine) was first tested in the TERMIS project. With the exception of values for minimum collocation salience for nouns and values for minimum gramrel salience for verbs, which remained unchanged, we had to reduce the minimum values for all other parameters of grammatical relations. This was expected, given the fact that the KoRP corpus (1.8 million words) is much smaller than the Gigafida corpus (1.2 billion words), used in extracting the information for the lexical database.

6. GDEX

Part of the method for extracting lexical information involves the GDEX tool. GDEX ranks corpus examples according to their dictionary potential by using criteria such as sentence length, whole-sentence form, sentence complexity, presence/absence of rare words, presence of URLs etc., and is therefore a very useful function for lexicographers (Kilgarriff et al., 2008; Kosem et al., 2011; Kosem, Gantar & Krek, 2012). It has been envisaged from the very beginning that the dictionary of public relations will include collocations as well as examples, so we yet again utilized the knowledge gained during the compilation of the Slovene Lexical Database (Kosem et al., 2011; Kosem, Gantar & Krek, 2012).

	min. coll. freq.	min. coll. sal.	min. gramrel freq.	min. gramrel sal.	gramrel type
O_količina	2	0.5	6	10.0	O
O_nedoločnik_cs	2	0.5	8	0.2	O
O_povratni_se	2	0.5	8	0.2	O
O_povratni_si	2	0.5	8	0.2	O
O_s_števili	2	0.5	6	1.0	O
O_tretja_oseba	2	0.5	8	0.2	O
O_z_lastnim_imenom	2	0.5	6	1.0	O
O_zanikanje	2	0.5	6	10.0	O
S_.*_p2	2	0.5	6	10.0	S
S_.*_p3	2	0.5	6	10.0	S
S_.*_p4	2	0.5	6	10.0	S
S_.*_p5	2	0.5	6	10.0	S
S_.*_p6	2	0.5	6	10.0	S
S_.*_r	2	0.5	8	0.2	S
S_.*_r2	2	0.5	8	0.2	S
S_.*_r3	2	0.5	8	0.2	S
S_.*_r4	2	0.5	8	0.2	S
S_.*_r5	2	0.5	8	0.2	S
S_.*_r6	2	0.5	8	0.2	S
S_.*_s2	2	-20.0	6	2.0	S
S_.*_s3	2	-20.0	6	2.0	S
S_.*_s4	2	-20.0	6	0.2	S
S_.*_s5	2	-20.0	8	0.5	S
S_.*_s6	2	-20.0	6	1.0	S
S_.*_x_g2	2	-20.0	6	0.5	S

Table 2: Part of settings for grammatical relations for nouns, group 2.

We prepared five different GDEX configurations for nouns and two for verbs; the configurations differed in values of certain parameters (e.g. optimum example length: 15–40 words/15–35 words/15–30 words). After several evaluations we selected two final configurations: one for nouns (single-word and multi-word) and one for verbs. The difference in ranking of examples by different configurations was especially noticeable for more frequent nouns, i.e. nouns with frequency over 600, while the comparison of rankings for single-word nouns, multi-word nouns, and verbs with frequency under 250 displayed little or no difference; however, this is to be expected due to a smaller number of examples for each collocate of low frequency words. Nonetheless, even in the cases of collocates with fewer examples, GDEX saved valuable time by ranking, and thus selecting for automatic export, the two best examples.

If compared with the GDEX configuration used for the Slovene Lexical Database, only three changes have been required for the terminology extraction. The changes to frequency settings were expected. Table 3 shows the differences in settings for nouns.

Classifier	Slovene Lexical Database	TERMIS
penalty for examples containing tokens with frequency of less than 3	yes	no
lemma frequency	yes, frequency = 1000	no
additional classifier for second-level collocations	yes, weight 10	yes, weight 10 (min. frequency of a collocate: 2)

Table 3: Part of GDEX configuration settings for Lexical Database for Slovene and TERMIS (single-word and multi-word nouns).

In addition to the three changes in settings used for noun terms, another change was made in the extraction of information for verb terms; namely, we added a classifier for optimum position of the keyword (i.e. term), so that the examples containing the keyword in the last two thirds of the sentence were ranked higher.

After all the configurations had been prepared, we ran the API script and extracted the information in XML format, and after minor conversions (e.g. gramrel names) imported the data into the editing tool of the Termania terminology portal.

Using word sketches, sketch grammar and GDEX, we extracted collocates and good examples (two examples per collocate). Each collocate was automatically listed under the relevant grammatical relation. Using this approach, we avoided manual corpus analysis for nearly 2000 terms, including the consultation of word sketches. Manual corpus analysis was used for only 150 multi-word terms that did not contain the combination *adjective* + *noun* or *noun* + *noun*, i.e. multi-word terms where automatic extraction of lexical information was not possible.

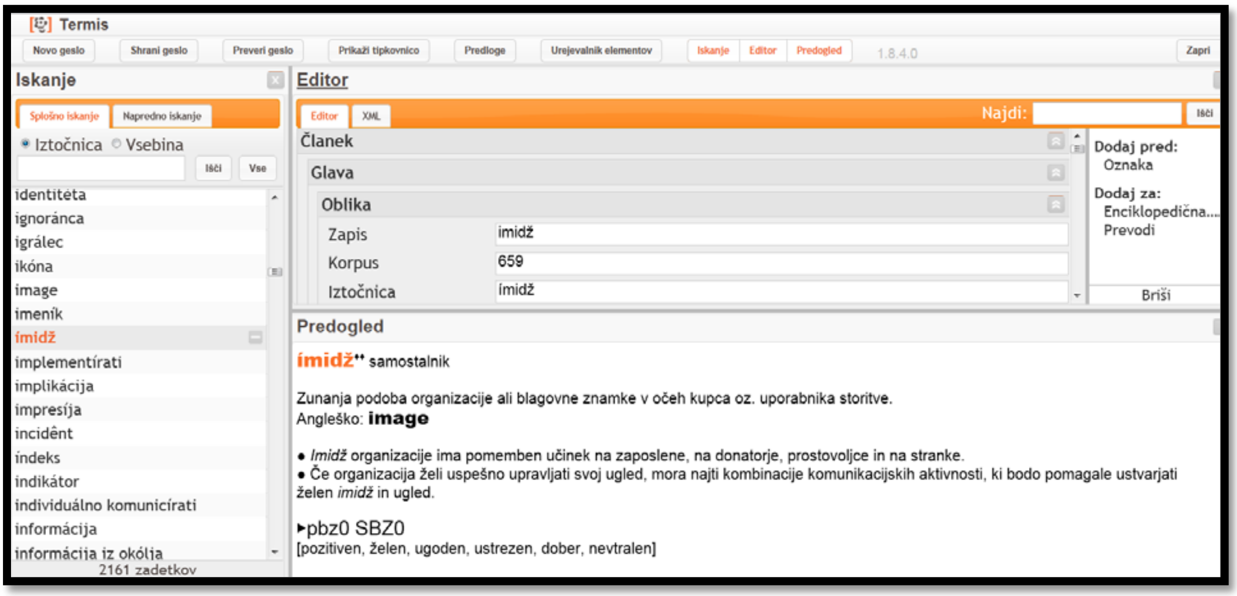
7. Editing the data

“No matter how many features are used to summarize the data, the lexicographer still needs to critically review the summary” (Kilgarriff & Kosem, 2012: 48). One of the differences between the compilation of a terminological dictionary and the compilation of a general dictionary is that there is much less polysemy in a terminological dictionary. Consequently, the work with the dictionary editor on the Termania portal (Figure 2) mainly comprised the redistribution and grouping of semantically related collocates, identification of compounds, and moving and reordering of corpus examples. In rare cases, we were required to re-examine the word sketch of the term, and in 10% of collocates the automatically extracted examples were too similar, so we analyzed the concordances and manually selected another example. This was the case for rare words and phrases where GDEX is of little use; in fact, in many cases the manual analysis revealed that there are no alternative different examples in the corpus, as the authors of corpus texts cited the same source in the same or a very similar manner. Even in cases when all the word sketch information was extracted (e.g. for 479 nouns with corpus frequency of less than 20), the automatic extraction reduced the time required for editing; deleting irrelevant information was quicker than the alternative, i.e. searching for relevant information in concordances and manually exporting each example.

8. Visualization of data for online dictionary

The visualization of the terminological dictionary of public relations is currently in its final stage. Online availability of the dictionary database was included in the original project proposal.

There are some important characteristics of online dictionaries or databases, including a customizable interface, filters, hyperlinks, video content, etc. (e.g. see Corréard, 2002; Schryver, 2003: 152–160; Heid & Gouws, 2006: 981; Caruso 2011). Simultaneously, we are aware of the rather unexpected findings of Müller-Spitzer, Koplenig & Töpel (2011), that multimedia content and other functionalities of online dictionaries are regarded as rather unimportant by users, especially if compared with the importance of reliability, clarity, and the up-to-date nature of information (see also Koplenig, 2011). Thus, we focussed on one particular aspect of visualization: how to present data to the user in a clear and understandable manner, considering that for a terminological dictionary there is likely to be a large amount of data for a single entry.



● Iztočnica ○ Vsebina

identitéta
ignoránca
igrálec
ikóna
image
imeník
ímidž
implementírati
implikácija
impresíja
incidênt
índeks
indikátor
individuálno komunicírati
informácija
informácija iz okólja
2161 zadetkov

Zapis	imidž
Korpus	659
Iztočnica	ímidž

ímidž** samostalnik

Zunanja podoba organizacije ali blagovne znamke v očeh kupca oz. uporabnika storitve.
Angleško: **image**

- *Imidž* organizacije ima pomemben učinek na zaposlene, na donatorje, prostovoljce in na stranke.
- Če organizacija želi uspešno upravljati svoj ugled, mora najti kombinacije komunikacijskih aktivnosti, ki bodo pomagale ustvarjati želen *imidž* in ugled.

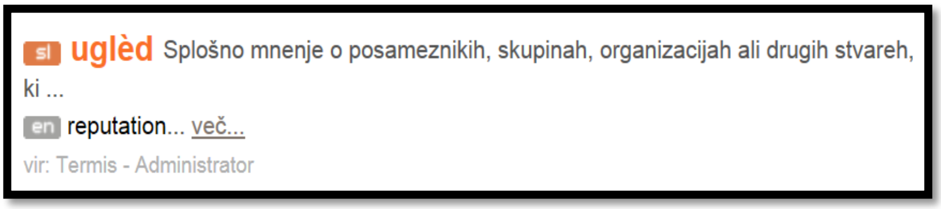
▸pbz0 SBZ0
[pozitiven, želen, ugoden, ustrezen, dober, nevtralen]

Figure 2: Dictionary editor of the Termania portal.

As shown in Figures 3–5, each entry (at the moment) consists of **three levels of display**:

I. The home page of Termania contains a search window that enables searches in all dictionaries included in the portal. At this level, the search results from the dictionary of public relations include the following information: headword, beginning of a (short) definition, and translation of headword into English (Figure 3).

II. By clicking on the headword, we open the second level where additional information is displayed: frequency in the KoRP corpus in the form of diamonds,[4] grammar information, entire short definitions, two corpus examples, collocates grouped by grammatical relation (Figure 4), related entries in the same dictionary (cross references), and, in the last part of the entry, links to concordances in Gigafida and KoRP.

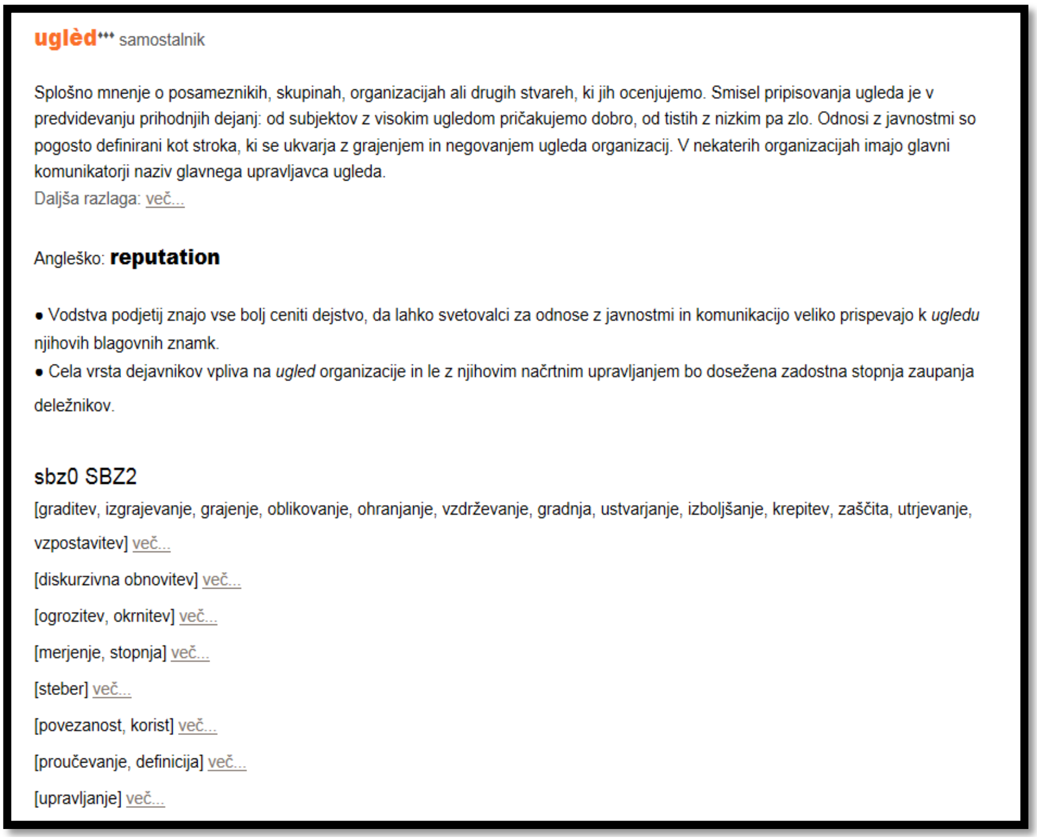


sl **uglèd** Splošno mnenje o posameznikih, skupinah, organizacijah ali drugih stvareh, ki ...

en reputation... več...

vir: Termis - Administrator

Figure 3: Termania portal: first level of entry display.



uglèd♦♦♦ samostalnik

Splošno mnenje o posameznikih, skupinah, organizacijah ali drugih stvareh, ki jih ocenjujemo. Smisel pripisovanja ugleda je v predvidevanju prihodnjih dejanj: od subjektov z visokim ugledom pričakujemo dobro, od tistih z nizkim pa zlo. Odnosi z javnostmi so pogosto definirani kot stroka, ki se ukvarja z grajenjem in negovanjem ugleda organizacij. V nekaterih organizacijah imajo glavni komunikatorji naziv glavnega upravljavca ugleda.

Daljša razlaga: več...

Angleško: **reputation**

- Vodstva podjetij znajo vse bolj ceniti dejstvo, da lahko svetovalci za odnose z javnostmi in komunikacijo veliko prispevajo k *ugledu* njihovih blagovnih znamk.
- Cela vrsta dejavnikov vpliva na *ugled* organizacije in le z njihovim načrtnim upravljanjem bo dosežena zadostna stopnja zaupanja deležnikov.

sbz0 SBZ2

[graditev, izgrajevanje, grajenje, oblikovanje, ohranjanje, vzdrževanje, gradnja, ustvarjanje, izboljšanje, krepitev, zaščita, utrjevanje, vzpostavitev] več...

[diskurzivna obnovitev] več...

[ogrozitev, okrnitev] več...

[merjenje, stopnja] več...

[steber] več...

[povezanost, korist] več...

[proučevanje, definicija] več...

[upravljanje] več...

Figure 4: Termania portal: second level of entry display (partial screenshot).

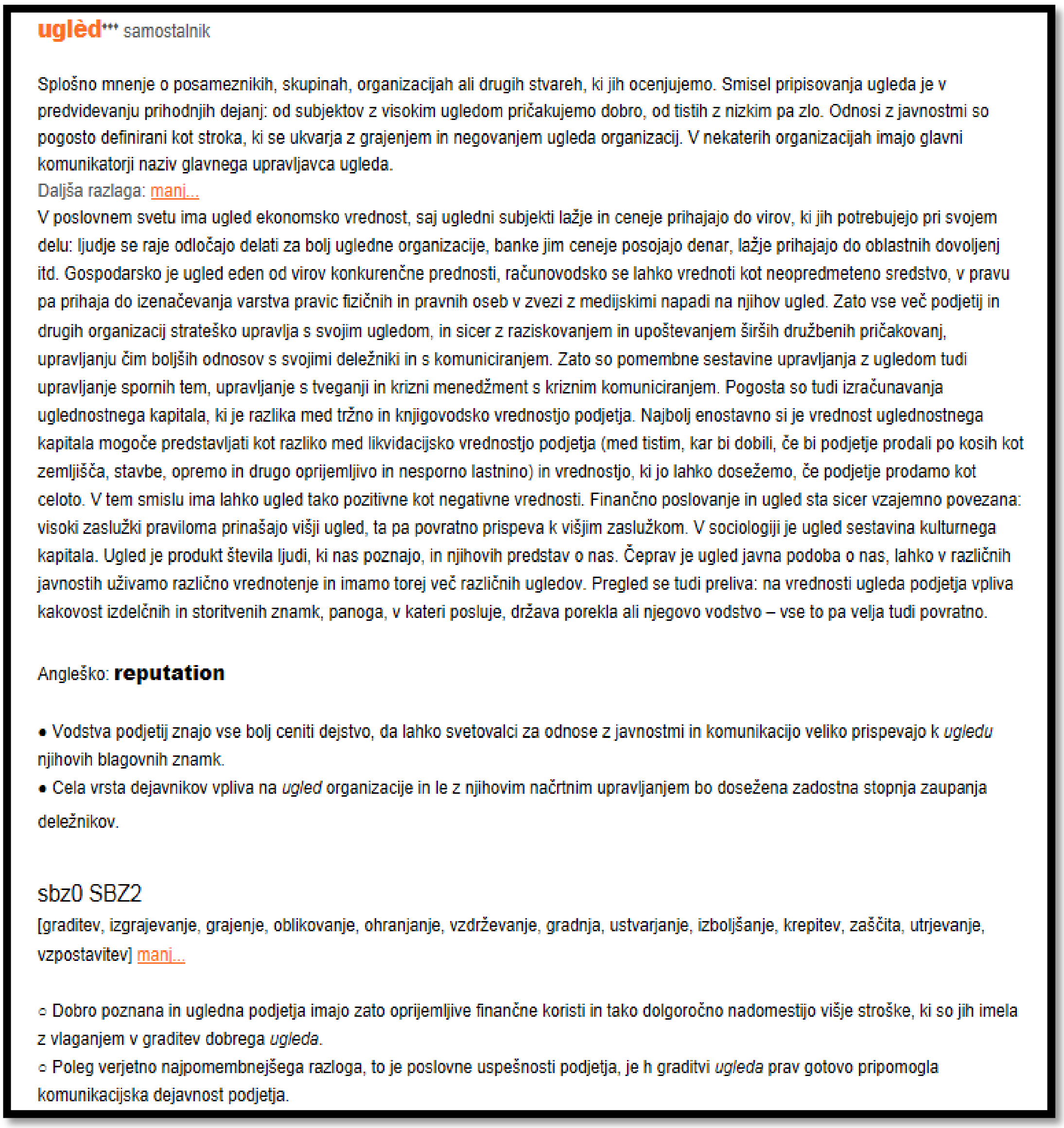
[4] Number of diamonds and related values: one diamond = frequency between 1–99, two diamonds = frequency of 100–699, three diamonds = frequency 700 and above.

III. The user accesses the third level by clicking on *več...* (*more...*), available in two places:

a) at a short definition, where a click on *več...* reveals a longer, encyclopaedic definition (Figure 5, above),

b) and at each group of collocates, where a click on *več...* reveals corpus examples, two per collocate (Figure 5, below).

At the time of writing the paper, we are conducting a survey among public relations experts and translators. The survey will provide information on understandability, clarity, accuracy, readability, amount, usefulness and relevance of the database contents and its structure. The survey findings will be implemented in the design of the dictionary.



uglèd*** samostalnik

Splošno mnenje o posameznikih, skupinah, organizacijah ali drugih stvareh, ki jih ocenjujemo. Smisel pripisovanja ugleda je v predvidevanju prihodnjih dejanj: od subjektov z visokim ugledom pričakujemo dobro, od tistih z nizkim pa zlo. Odnosi z javnostmi so pogosto definirani kot stroka, ki se ukvarja z grajenjem in negovanjem ugleda organizacij. V nekaterih organizacijah imajo glavni komunikatorji naziv glavnega upravljavca ugleda.

Daljša razlaga: manj...

V poslovnem svetu ima ugled ekonomsko vrednost, saj ugledni subjekti lažje in ceneje prihajajo do virov, ki jih potrebujejo pri svojem delu: ljudje se raje odločajo delati za bolj ugledne organizacije, banke jim ceneje posojajo denar, lažje prihajajo do oblastnih dovoljenj itd. Gospodarsko je ugled eden od virov konkurenčne prednosti, računovodsko se lahko vrednoti kot neopredmeteno sredstvo, v pravu pa prihaja do izenačevanja varstva pravic fizičnih in pravnih oseb v zvezi z medijskimi napadi na njihov ugled. Zato vse več podjetij in drugih organizacij strateško upravlja s svojim ugledom, in sicer z raziskovanjem in upoštevanjem širših družbenih pričakovanj, upravljanju čim boljših odnosov s svojimi deležniki in s komuniciranjem. Zato so pomembne sestavine upravljanja z ugledom tudi upravljanje spornih tem, upravljanje s tveganji in krizni menedžment s kriznim komuniciranjem. Pogosta so tudi izračunavanja uglednostnega kapitala, ki je razlika med tržno in knjigovodsko vrednostjo podjetja. Najbolj enostavno si je vrednost uglednostnega kapitala mogoče predstavljati kot razliko med likvidacijsko vrednostjo podjetja (med tistim, kar bi dobili, če bi podjetje prodali po kosih kot zemljišča, stavbe, opremo in drugo oprijemljivo in nesporno lastnino) in vrednostjo, ki jo lahko dosežemo, če podjetje prodamo kot celoto. V tem smislu ima lahko ugled tako pozitivne kot negativne vrednosti. Finančno poslovanje in ugled sta sicer vzajemno povezana: visoki zaslužki praviloma prinašajo višji ugled, ta pa povratno prispeva k višjim zaslužkom. V sociologiji je ugled sestavina kulturnega kapitala. Ugled je produkt števila ljudi, ki nas poznajo, in njihovih predstav o nas. Čeprav je ugled javna podoba o nas, lahko v različnih javnostih uživamo različno vrednotenje in imamo torej več različnih ugledov. Pregled se tudi preliva: na vrednosti ugleda podjetja vpliva kakovost izdelčnih in storitvenih znamk, panoga, v kateri posluje, država porekla ali njegovo vodstvo – vse to pa velja tudi povratno.

Angleško: **reputation**

- Vodstva podjetij znajo vse bolj ceniti dejstvo, da lahko svetovalci za odnose z javnostmi in komunikacijo veliko prispevajo k *ugledu* njihovih blagovnih znamk.
- Cela vrsta dejavnikov vpliva na *ugled* organizacije in le z njihovim načrtnim upravljanjem bo dosežena zadostna stopnja zaupanja deležnikov.

sbz0 SBZ2

[graditev, izgrajevanje, grajenje, oblikovanje, ohranjanje, vzdrževanje, gradnja, ustvarjanje, izboljšanje, krepitev, zaščita, utrjevanje, vzpostavitev] manj...

○ Dobro poznana in ugledna podjetja imajo zato oprijemljive finančne koristi in tako dolgoročno nadomestijo višje stroške, ki so jih imela z vlaganjem v graditev dobrega *ugleda*.

○ Poleg verjetno najpomembnejšega razloga, to je poslovne uspešnosti podjetja, je h graditvi *ugleda* prav gotovo pripomogla komunikacijska dejavnost podjetja.

Figure 5: Termania portal: third level of entry display (partial screenshot).

9. Conclusion

The aim of the TERMIS project was the development of a model for compiling Slovene terminological dictionaries or systematically structured databases in a relatively short amount of time. The method of automatic extraction of term candidates and lexical information (including collocations and examples) from the corpus, used in compiling the terminological database of public relations terms, is in fact language independent; individual parameters of different tools can be adapted to other languages, something that is important for modern lexicography that promotes automating as much of the lexicographic work as possible (Kosem, Gantar & Krek, 2012; Rundell & Kilgarriff, 2011). The use of language technologies and lexicographic tools, as described in this paper, has not only facilitated a quicker building of terminological database for the discipline of public relations, but has also made the analysis more objective.

It appears that user-friendliness and the availability of various multimedia functions, enabled by the online dictionary medium, are yet to be fully developed by dictionary-makers; similarly, dictionary users are still getting accustomed to these functions (Müller-Spitzer, Koplenig & Töpel, 2011). The online format has removed the need for space-saving techniques; but in contrast has raised two questions: how much data is still manageable for the user, and how should dictionary information be effectively organized in this new medium?

> /O/ne of the really distinctive features of dictionaries and other lexicographical tools is that they provide quick and easy access to the specific types of data from which a specific type of users can retrieve the information that may cover their specific types of needs in a specific type of extra-lexicographical situation (Tarp 2010: 40).

We are currently conducting a survey among the users of our dictionary that will provide answers to certain questions related to design and structure of dictionary data, but only the feedback of the wider public can evaluate how successful we have been in achieving the aim of our dictionary project.

The TERMIS project has highlighted how language technologies can speed up the building of terminological databases. In addition, language technologies can be used to identify types of information that can be difficult to obtain by manual analysis.

We have developed a model for building a terminological database that could be adopted by other disciplines in Slovenia for the compilation of respective terminological dictionaries. We believe that in times of internationalization of disciplines and research, an effective way to facilitate the development of terminology is by using the approach demonstrated by TERMIS: by making state-of-the-art electronic terminological resources.

10. Acknowledgements

The findings presented in this paper are the result of collaborative work; the authors would like to thank Polona Gantar, Simon Krek and Špela Vintar.

The research was conducted as part of the project entitled *Terminology data banks as the bodies of knowledge: The model for the systematization of terminologies*, funded by the Slovene Research Agency (contract no. 1000-11-274193), Pristop d.o.o., and Chamber of Commerce and Industry of Slovenia. The project has been supported by the following sponsors: Elektro Ljubljana d.d., Mercator d.d., Pošta Slovenije d.o.o., and Zavarovalnica Maribor d.d.

11. References

Arhar Holdt, Š. (2011). *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Arhar Holdt, Š., Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), pp. 95-110.

Caruso, V. (2011). Online specialised dictionaries: a critical survey. *Proceedings of eLex 2011.* Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 66-75.

Corréard, M. (2002). Are space-saving strategies relevant in electronic dictionaries. *Proceedings of the 10th EURALEX international congress*, Copenhagen: Center for Sproktehnologi, pp. 463–470.

Gantar, P. (2009). Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo*, 54(3/4), pp. 69-94.

Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) *Natural language processing, multilinguality: 6th international conference*. Modra: Tribun EU, pp. 72-80.

Grčar, M., Krek, S. & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 89-94.

Heid, U., Gouws, R. (2006). A model for a multifunctional dictionary of collocations. *Proceedings of the 12th EURALEX international congress*. Torino: Edizioni dell'Orso, pp. 979-988.

Humar, M., Žagar Karer, M., eds. (2010). *Nacionalni jeziki v visokem šolstvu*. Ljubljana: Založba ZRC, ZRC SAZU.

Kalin Golob, M. (2012). Jezik slovenskega visokega šolstva: med zakonodajo, strategijo in vizijo. In V. Gorjanc (ed.) *Slovanski jeziki: iz preteklosti v prihodnost*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 95-109.

Kalin Golob, M., Stabej, M. (2007). Sporazumevanje v znanosti in na univerzi: uboga slovenščina ali uboga jezikovna politika? *Jezik in slovstvo*, 52(5), pp. 87-91.

Kalin Golob, M., Stabej, M., Stritar, M., & Červ, G. (2012). *Primerjalna študija o učnem jeziku v visokem šolstvu v Republiki Sloveniji in izbranih evropskih državah*. Accessed at:

http://www.mizks.gov.si/fileadmin/mizks.gov.si/pageuploads/Slovenski_jezik/FDV_-_ucni_jeziki_v_visokem_solstvu.pdf.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX international congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425-432.

Kilgarriff, A., Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger, M. Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 31-55.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch engine. *Proceedings of the 11th EURALEX international congress*. Lorient: Universite de Bretagne-Sud, pp. 105-116.

Kompara, M., Holozan, P. (2011). What is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using Termania website. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 140-146.

Koplenig, A. (2011). Understanding how users evaluate innovative features of online dictionaries – an experimental approach. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 147-150.

Kosem, I., Gantar, P. & Krek, S. (2012). Avtomatsko luščenje leksikalnih podatkov iz korpusa. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 117-122.

Kosem, I., Husak, M. & McCarthy, D. (2011). GDEX for Slovene. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 150-159.

Krek, S. (2012). New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3*. Brno. Accessed at: http://trac.sketchengine.co.uk/wiki/SKEW-3/Program#.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Logar, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: doktorska disertacija. Ljubljana: Filozofska fakulteta.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What makes a good online dictionary? – Empirical insights from an interdisciplinary research project. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 203-208.

Romih, M., Krek, S. (2012). Termania – prosto dostopni spletni slovarski portal. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 163-166.

Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.) *A taste for corpora: a tribute to professor Sylviane Granger*. Amsterdam: John Benjamins.

Schryver, G. de (2003). Lexicographers' dreams in the electronic-dictionary age. *International journal of lexicography*, 16(1), pp. 143-199.

Tarp, S. (2010). Functions of specialized learners dictionaries. In P. Fuertes-Olivera (ed.) *Specialised dictionaries for learners*. Berlin, New York: De Gruyter, pp. 39-53.

Vintar, Š. (2008). *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Vintar, Š. (2009). Samodejno luščenje terminologije – izkušnje in perspektive. In N. Ledinek, M. Žagar Karer, M. Humar (eds.) *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, pp. 345-356.

Vintar, Š. (2010). Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141-158.