Izvorni znanstveni članak
UDK 811.163.6'373.4:81'42

Primljen 10. 2. 2003.
Prihvaćen 18. 4. 2003.

IDENTIFYING MARKERS OF SEMANTIC RELATIONS IN SLOVENE

Špela Vintar and Vojko Gorjanc *
Faculty of Pilosophy, University of Ljubljana, Slovenia

Semantically related words – synonyms, hyper- and hyponyms, abbreviations – often collocate or appear in similar contexts and it is usually possible to identify the domain of a word either on the basis of its lexical environment or the text it occurs in. Thus a group of related words can easily be assigned to a specific domain, but to identify the relations between them may not be as trivial. Certain phrases or lexemes express such relations in a more or less fixed syntactic pattern, and by their identification it becomes possible to retrieve certain pairs of words together with the information about the type of semantic link between them. If we, for example, identify the pattern "X, also known as Y" to indicate complete or near synonymy, we can extract the noun phrases linked by the pattern as a pair of synonyms.

Such methods are already extensively used in various fields of information retrieval, however so far few attempts have been made for Slovene. The paper gives some methods of identifying such markers and how they can be used – in combination with other evidence – to extract related words or phrases. We describe the problem of variability, as these semantic links often exceed sentence boundaries and appear in various forms. Furthermore, the markers themselves can be ambiguous and express different semantic relations. Nevertheless, we envisage the application of such methods in lexicography, terminography and particularly in thesaurus construction, a task still to be accomplished for Slovene. For all experiments described we used the FIDA reference corpus of Slovene.

Key words: corpus-based terminology, semantic relations, term extraction, information retrieval

INTRODUCTION

The lexicon of a language is a structured network of concepts and relations between them. When the words of a language are combined and woven into a text, the text normally represents an instance of conceptual reality and at the same time a referentially bounded framework.

Within this framework words and the concepts behind them relate to each other in various ways. On the level of the lexicon, a word such as *parent* entails the existence of a *child*, on the level of the conceptual reality an *aeroplane* is related to *flamingo* in that they both fly, and on the textual level words are automatically put in relation with one another by occurring in the same context. Furthermore, we can refer to the same concept with various words, or use pronouns and other deictic devices to point to a previously introduced concept (Saeed, 1998).

However, when we want to explain the relations between concepts within the reality portrayed we often use explicit linguistic structures or phrases, such as *X is defined as Y, X is an instance of Y, There are several types of*

* Vojko Gorjanc, Oddelek za slovenistiko, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, SL-1000 Ljubljana, Slovenia; tel.: 386-1-241 1306; fax: 386-1-425 9337; e-mail: vojko.gorjanc@guest.arnes.si

X, for example A, B, C etc. These function as pathfinders within the conceptual network and map our understanding of the text and the domain.

By identifying such markers it is possible to extract not only the lexicon of a specific domain but also the relations between lexical items, which can be used for several purposes. In information retrieval, this method can be applied for knowledge acquisition and integration into expert systems. For example in medical literature certain lexical patterns are used frequently enough to be formalized for knowledge extraction purposes, e.g. *Drug Y causes Disease X* (Cimino and Barnett, 1993).

Similar methods are proposed by other authors in the field of terminology, either for the purpose of automatic construction of knowledge databases (Bowden et al, 1996), thesaurus construction (Richardson et al, 1998) and conceptual sampling for terminography (Meyer et al, 1999). This paper was particularly inspired by the latter.

Although much work has been done, almost all authors have performed their experiments on English, most of them using domain-specific corpora of scientific or technical texts. The aim of this paper is to explore the possibilities of extracting semantically related words from Slovene using a 4-million-word subcorpus of texts belonging to the category natural science taken from FIDA, the 100-million-word reference corpus of Slovene (Erjavec *et al*, 1998). Apart from the fact that this field has hardly been explored so far for Slovene, the approach itself had to be completely modified since Slovene is a highly inflected language and the phrases we wanted to extract occur in many different variants.

IDENTIFYING AND CLASSIFYING MARKERS OF SEMANTIC RELATIONS

We were interested in the following basic relations between lexical items, as defined also in (WordNet, 2000): synonymy (full or partial), hyperonymy (X *is a kind of* Y), hyponymy (Y *is a kind of* X), meronymy (X *is a part of ...*) and holonymy (*parts of* X *are ...*). Our approach involved the following steps:

- identification of phrases (markers) that link two related concepts in a text, e.g. *X is a kind of Y, X is also known as Y*
- finding the syntactic pattern that allows retrieval of related concepts, e.g. *NPnom, sg belongs in the group of NPpos, pl*
- classification of the markers according to two measures, **yield** and **reliability**.

The first step inevitably involved some guesswork and relying upon intuition. Although previous research on English provided a good basis, some features simply cannot be translated into Slovene. For example, the indefinite article in English quite reliably points to something new being introduced, so that the structure *X is a Y* is typical in definitions and indicates that X is a hyponym of Y. Having no articles, Slovene renders the corresponding structure only with the auxiliary verb X *je* Y, which is too ambiguous to be considered a marker. Another method of finding relevant markers is to take two words known to be related (such as *disease* and *cancer*), retrieve all co-occurrences from the corpus and examine the phrases that link them.

Below we list some markers that were identified and supported by corpus evidence:

- **hypo- and hypernymy**: je, kot je (na primer), kot so npr., je vrsta, je _ vrsta, prištevamo med, sodi* med, med _ sodi*, spada* med, sodi* v družino, uvrščamo med, med _ uvrščamo, uvrščamo v skupino...
- **synonymy**: ali, ali tudi, imenujemo tudi, imenovan* tudi, (sinonim _), je sinonim za, znan* tudi kot, znan* tudi pod imenom, z drugim imenom...
- **mero- and holonymy**: ima, ima _ dele, je iz, je sestavljen iz, vsebuje...

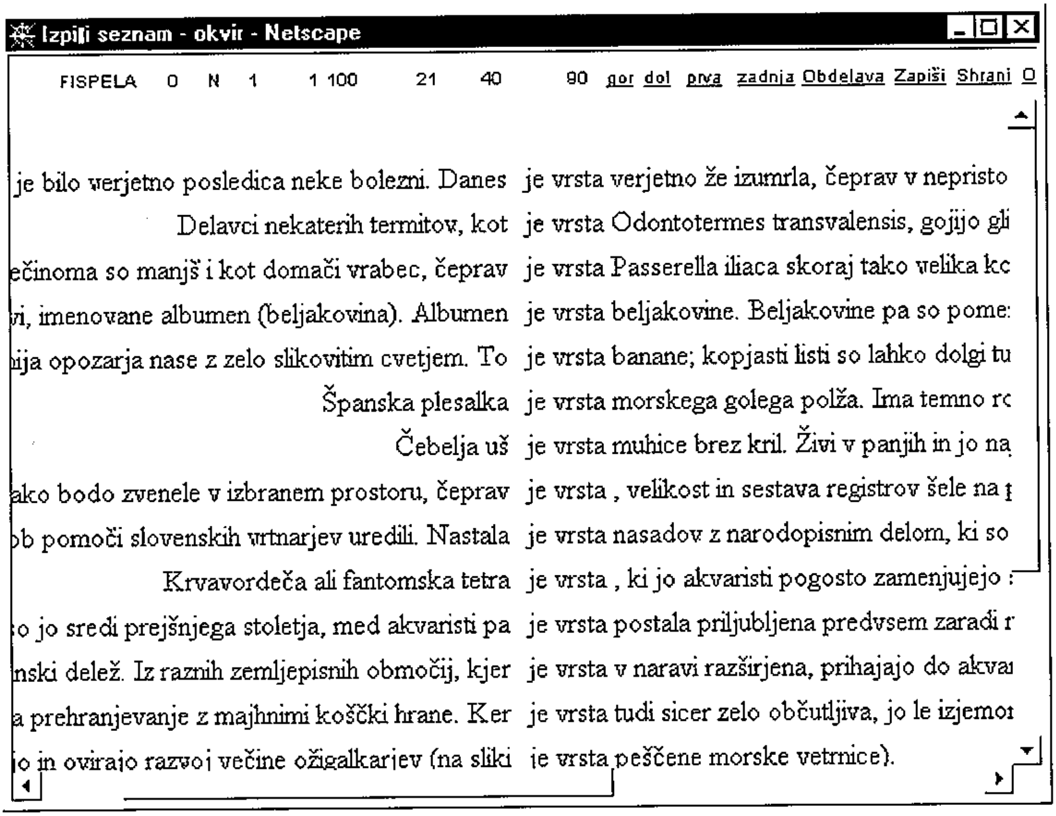
Once the phrase that marks a semantic relation has been identified, we can retrieve sentences containing the marker from the corpus. However, a look at the concordances (see Figure 1) shows that only a portion of the sentences retrieved actually contain two related concepts. To separate useful examples from the rest we need to identify the syntactic pattern that makes the extraction possible. For most semantic relations, the basic pattern is a noun phrase linked to another noun phrase by the marker, for example:

Zato spletne strani imenujemo tudi HTML dokumenti.

[Web pages are thus called also HTML documents.]

However, an important element that facilitates extraction and improves accuracy is case. Slovene has six cases and three grammatical numbers (singular, dual and plural), which provides valuable clues about potential relatedness. If we are for instance looking for a pair of synonyms, both noun phrases have to agree in number.

Some patterns are non-contiguous, meaning that one or more words may appear between the concepts we want to extract, others are variable in that an indefinite number of extractable items may follow, for example:


Izpiši seznam - okvir - Netscape
FISPELA 0 N 1 1 100 21 40 90
gor dol prva zadnja Obdelava Zapiši Shrani O

```
je bilo verjetno posledica neke bolezni. Danes  je vrsta verjetno že izumrla, čeprav v nepristo
           Delavci nekaterih termitov, kot  je vrsta Odontotermes transvalensis, gojijo gli
ečinoma so manjš i kot domači vrabec, čeprav  je vrsta Passerella iliaca skoraj tako velika kc
vi, imenovane albumen (beljakovina). Albumen  je vrsta beljakovine. Beljakovine pa so pome:
hija opozarja nase z zelo slikovitim cvetjem. To  je vrsta banane; kopjasti listi so lahko dolgi tu
                         Španska plesalka  je vrsta morskega golega polža. Ima temno rc
                               Čebelja uš  je vrsta muhice brez kril. Živi v panjih in jo na
ako bodo zvenele v izbranem prostoru, čeprav  je vrsta , velikost in sestava registrov šele na p
ob pomoči slovenskih vrtnarjev uredili. Nastala  je vrsta nasadov z narodopisnim delom, ki so
            Krvavordeča ali fantomska tetra  je vrsta , ki jo akvaristi pogosto zamenjujejo :
o jo sredi prejšnjega stoletja, med akvaristi pa  je vrsta postala priljubljena predvsem zaradi r
nski delež. Iz raznih zemljepisnih območij, kjer  je vrsta v naravi razširjena, prihajajo do akvai
a prehranjevanje z majhnimi koščki hrane. Ker  je vrsta tudi sicer zelo občutljiva, jo le izjemoı
o in ovirajo razvoj večine ožigalkarjev (na sliki  je vrsta peščene morske vetrnice).
```

Figure 1: Concordance lines for je vrsta [is a kind of] from the FIDA corpus

Med enokaličnice *spadajo* palme, trave, bambusi, perunike in lilije.

[Monocotyledons *include* palms, grasses, bamboos, irises and lilies.]

Table 1: Hyponymy

Phrase	Freq	Pattern	Yield	Reliability
je vrsta [is a kind of]	90/19	NPN ~ NPG	0.21	0.79
je _ vrsta [is a ___ kind of]	156/24	NPN ~ NPG	0.15	0.75
prištevamo med [is counted among]	12/9	NPA ~ NPA, NPN, ki jih ~ NPA	0.75	1.00
sodi* med [belongs among]	182/45	NPN ~ NPA	0.24	0.86
spada* med [belongs among]	168/34	NPN ~ NPA	0.20	0.88
sodi* v družino [belongs in the family of]	13/4	NPN ~ NPGpl	0.30	0.50
uvrščamo med [is categorised as]	60/22	NPA ~ NPA	0.36	0.77

Another problem is the fact that some phrases may function as markers of two distinct relations, depending on the context. Thus, the phrase *imenujemo tudi* [named also] generally marks synonymy, however when the initial NP is used with a demonstrative pronoun, or the real synonym has already been mentioned in the previous sentence, the relation is almost exclusively hyponymy.

Knowing the pattern and retrieving only instances that correspond to it nevertheless gives less than perfect results. The words and relations we extract may be terminologically irrelevant or unrelated, or the relation between them is not the one we predicted. The majority of these "errors" are easily explained by looking at the broader context of the sentence and could probably be avoided if extraction patterns would be expanded to work across sentence boundaries.

In order to characterise each marker in terms of its accuracy in predicting two related noun phrases and in terms of its productivity in the corpus, two measures were invented: reliability and yield. Yield is defined as the number of corpus occurrences of the entire pattern divided by the number of phrase occurrences. Thus, if the phrase *imenujemo tudi* [named also] occurs 185 times in the subcorpus, of which only 69 cases correspond to the NP-rich syntactic pattern, the yield is 0.37. Reliability, on the other hand, measures how many of the NP-rich examples are indeed terminologically relevant. If 50 sentences containing *imenujemo tudi* would lead to successful extraction of semantically related concepts, the reliability of the marker is 0.72. The table above shows the results obtained for markers of hyponymy.

The results above were obtained using a non-lemmatized corpus, and indeed marker phrases are relatively fixed in their form. This is to say that if the first person plural form of *uvrščamo med* is found to mark hyponymy, the other forms do not necessarily mark the same relation nor do they need to be equally reliable. The third person plural *uvrščajo med*, for example, usually occurs in contexts where false or ambiguous categorisations are presented or advised against:

Druge dežele so odločitev "poni ali konj" rešile drugače, tako da se lahko zgodi, da konje, ki ponekod veljajo za ponije, v njihovih izvornih domovinah uvrščajo med konje.

[In other countries the distinction between "pony" and "horse" was resolved differently, so that it may happen that horses categorised as ponies elsewhere are counted among horses in their native countries.]

Nevertheless, in some cases an asterisk was used to retrieve several forms of a phrase where this was considered "safe", e.g. *znan* tudi kot, imenovan* tudi, sodi* med* etc.

CONCEPT HIERARCHIES

Knowing the marker and the pattern it occurs in allows us to retrieve concepts and integrate them in a relational network. This can be useful in the creation of a conceptual representation of a certain domain or to support thesaurus construction. For the study we describe here, the subcorpus was too small and too varied to provide enough data to represent an entire domain. On the other hand, some text types in natural science, such as textbooks, convey knowledge in a very clear and compact form and lend themselves well to concept extraction.

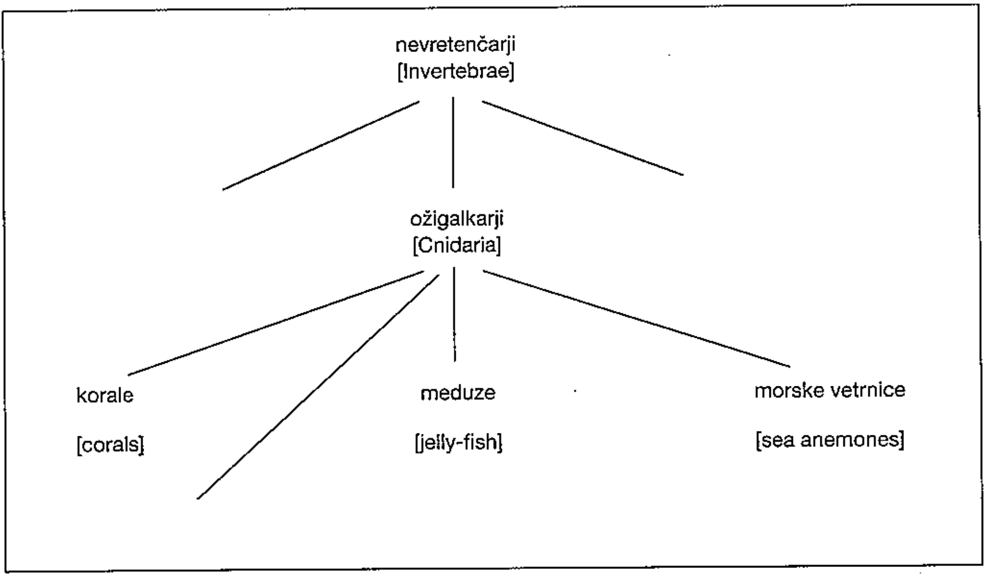


Figure 2: Concept hierarchy extracted from two sentences in the subcorpus

Some markers allow us to retrieve not only pairs of related concepts but an entire group. The sentence below shows how productive the hypernym-marker *med _ spadajo [the group of _ includes]* can be:

Med **gliste** spadajo številni **zajedavci** in **škodljivci**, kot so **talne glistice**, **rudarska glista**, **pljučna glista** in številne **filarije**, ki pri sesalcih povzročajo elefantijazo.

A broader and slightly different approach to the exploration of concept relations is known as *conceptual sampling* or extraction of *knowledge-rich contexts* (Meyer et al, 1999). In corpus-based terminology work we may not be interested merely in extracting related terms but also in developing methods to identify other pieces of information about a certain concept, which may be of help in formulating definitions or illustrating its usage. Thus, for example, certain patterns may single out a concept as having a unique characteristic that helps us distinguish it from its kind. The non-contiguous pattern *je _ vrsta* [is the __ of, is a _ kind of] usually provides such information:

Galapaški kormoran je edina vrsta kormoranov na svetu, ki ne letijo.

[The Galapagos cormorant is the only kind of cormorant that doesn't fly.]

Severni črni kit je največja vrsta kljunatih kitov.

[The northern black whale is the largest of the beaked whales.]

For the sake of objectivity it should be noted that *vrsta* in Slovene means both *kind (type, sort)* and *species*, and since the examples above come from the domain of zoology, the marker *je _ vrsta* should probably be interpreted as *is a/the _ species of*.

PROBLEMS

Although this method may prove successful in extracting some semantically related concept pairs or groups, a large number of relevant terms will remain untedected. This may be due to the fact that certain terms never occur in a detectable pattern, which is not to say that they occur "alone" because this is almost never the case, but rather that their semantical environment is spread over a broader context which cannot be formalised in the manner described above.

In a text dealing with a certain domain, relations between concepts are frequently established on a level that exceeds sentence boundaries. If, for example, one of the concepts forming a relation pair has been previously introduced, the phrasal marker and the pattern would contain only the second concept while the first would be replaced by a superordinate term (see example) or a demonstrative pronoun. A more complex approach to the extraction of related concepts would therefore also include anaphora resolution.

Gobo imenujemo tudi smrček.

[The mushroom is also called the snout.]

Since our approach is sentence-oriented, the identification patterns are formulated for declarative sentence types. In interrogative or imperative sentences the syntactical patterns may be different – often reversed – and affect the results. For instance, the marker *imenujemo tudi [is also called]* generally predicts synonymy or a terminological variant of the same concept, however if used in the form of a question it may mark hypernymy.

Kateri planet imenujemo tudi Zvezda Večernica?

[Which planet is also called the Evening Star?]

Finally, it is the aim of our approach to extract terminologically relevant pieces of information that can be combined into concept relations or used as sources of terminological knowledge about the domain. However, phrasal markers may be used figuratively and link concepts that are only related on the textual or stylistic level, not within the conceptual network of the domain. Therefore, examples like the one below would also be extracted on the basis of their form, although "unnecessary medical toyings" could hardly be considered a hypernym of "measuring the sense of hearing in infants".

Merjenje sluha novorojenčkov na prvi pogled sodi med nepotrebna zdravniška igračkanja.

[Measuring the sense of hearing in infants at first glance belongs to the unnecessary medical toyings.]

IMPROVING THE RESULTS AND FUTURE PLANS

Some of the problems described above could be avoided by refining the syntax patterns and by filtering. The former would probably include the development of a step-by-step analysis, whereby certain patterns would comprise larger contexts. Filtering the results would inevitably involve the treatment of modifyiers which influence or change the contents of the sentence but are not part of the patterns themselves, e.g. words like *sicer, včasih, poredko, komaj* [even if, sometimes, rarely, hardly] etc. If noise occurs in the form of syntactically correct but terminologically irrelevant patterns, additional data (e.g. frequency, keywordness) could be used to filter out only noun phrases that contain terms.

Our 4-million-words subcorpus was composed of texts from various domains belonging to the category natural science in the FIDA corpus, however the results showed that certain markers occur almost exclusively within a certain domain. Furthermore, many markers and patterns show differences in their reliability and yield for different domains as well as across various texts. An extensive comparative study would be needed to determine whether and to what extent markers can be regarded as specific to a domain or text type, or indeed individual author's style.

CONCLUSION

The paper presented an atempt to identify and evaluate phrasal markers of semantically related terms in Slovene texts pertaining to the domain of natural sciences. We focused on four semantic relations (synonymy, hyponymy, hypernymy, meronymy) and successfully identified over 30 phrases that more or less reliably mark semantic relations. Since the approach is sentence-oriented it has several limitations, which could to some extent be improved by using more complex extraction patterns and filtering.

Despite the rudimentariness of this method we envisage its application in corpus-based terminography for the creation of domain-specific conceptual networks or thesauri, as well as in information and knowledge retrieval tasks. In a broader perspective the method could be used to support the creation of a general language thesaurus for Slovene.

REFERENCES

- Bowden, P. R., Halstead, P. and Rose, T. G. (1996). Extracting Conceptual Knowledge from Text Using Explicit Relation Markers. *Proceedings of EKAW-96*. Nottingham: University of Nottingham.
- Cimino J. J., Barnett G. O. (1993). Automatic Knowledge Acquisition from MEDLINE. Methods of Information in Medicine, 32 (2), 120–130. Selected for reprint in van Bemmel J. H, and McCray A. T., eds. (1994): *Yearbook of Medical Informatics, International Medical Informatics Association, Rotterdam*, 384–394.
- Erjavec, T., Gorjanc, V. and Stabej, M. (1998). Korpus FIDA. *International Multi--Conference Information Society – IS'98, 6 – 7 October 1998*. Ljubljana: Institut Jožef Stefan (eds. T. Erjavec and J. Gros), 124–127.
- Meyer, I.; Mackintosh, K.; Barriere, C. and Morgan, T. (1999). Conceptual sampling for terminological corpus analysis. *Proceedings of TKE '99*. Vienna: TermNet (ed. P. Sandrini), 256–267.
- Richardson, S. D., Dolan W. B. and Vanderwende L. (1998). MindNet: acquiring and structuring semantic information from text. *Microsoft Research Technical Publications (MSR-TR-98–23)*. Available from ftp://ftp.research.microsoft.com/pub/tr/tr-98 – 23.doc
- Saeed, J. I. (1998). *Semantics*. Oxford UK, Cambridge MA: Blackwell.
- WordNet, http://www.cogsci.princeton.edu/~wn/online/ [14. 8. 2000.]