Building the Slovene Wordnet: First Steps, First Problems

Tomaž Erjavec
Department of Knowledge Technologies,
Jožef Stefan Institute,
Jamova 39, Ljubljana, Slovenia,
tomaz.erjavec@ijs.si

Darja Fišer
Department of Translation,
Faculty of Arts, University of Ljubljana
Aškerèeva 2, Ljubljana, Slovenia,
darja.fiser1@guest.arnes.si

Abstract

We report on the prototype Slovene wordnet which currently contains about 5,000 top-level concepts. The resource is based on the Serbian wordnet which has been automatically translated with the help of a bilingual dictionary, the literals ranked according to the frequency of corpus occurrence, and results manually corrected. The paper also discusses some problems encountered along the way and points out some possibilities of automated acquisition and refinement of synsets in the future.

Introduction

While several corpus resources exist for Slovene (FIDA, SVEZ-IJS, MULTEXT-East), there is a lack of semantic lexica. Therefore, much of the initial work on the Slovene wordnet had to be based on classical dictionaries and thus required extensive manual intervention.

The expand model (Vossen 1998) was used, which had also been adopted in the EuroWordNet (EWN) and several subsequent wordnet projects (e.g. MultiWordNet[1], BalkaNet[2]). However, its core notion was taken a step further: we assumed that concepts and relations among them overlap across languages better if the languages are closely related. We therefore decided to proceed from the Serbian wordnet as the closest relative of Slovene in the wordnet family. SWN's synsets were manually translated from English and were validated against monolingual and bilingual corpora within the BalkaNet project (Obradovic et al. 2004), which is why it is assumed that both synset equivalence across languages and synset contents are of high quality and representative of language usage.

1 Wordnet Creation

The Jurančič Slovene / Serbo-Croatian dictionary (Jurančič 1991) was used to create bilingual lemma pairs. The lexicon was then used to automatically translate the Serbian literals into Slovene; the literals not found were retained in Serbian and flagged for manual translation. Synset IDs and relations were preserved, while glosses, examples of use and sense numbers were omitted at this stage.

For the start, we only retained the most important concepts, referred to as Base Concept sets 1 and 2 (Vossen 2005: 54-58), which consist of 4,688 (1,219 BCS1 + 3,469 BCS2) synsets. The 153 missing hypernyms that unexpectedly belonged to BCS3 were included, amounting to a total of 4,841 top-level synsets.

The next step was manual translation of untranslated literals and revision of the translated synsets which were characterised by high recall but very low precision. Manual clean-up was carried out in VisDic (Horák & Smrž 2004). Manual revision was speeded up by automated detection of literals which were not found in pre-existing Slovene lexica or the Slovene reference corpus FIDA[3]. Literals were further classified into six bands according to their frequency in the lemmatised FIDA corpus. Band 0 – the lemmas that did not occur in the corpus (2,622 literals) – was examined manually in order to avoid unjustified exclusion of literals from the wordnet.

The top-level Slovene wordnet currently consists of 4,841 synonym sets and contains 19,660 literals. Out of 6,183 synsets in the Serbian wordnet, 73% have been included in the Slovene wordnet. Table 1 shows that synsets from BCS1 and BCS2 are well-represented in the Slovene wordnet[4], while BCS3 is yet to be extended.

2 Problems and Future Plans

Some problems encountered along the way originate in the PWN itself (e.g. connotation inconsistencies, exceedingly fine granularity of senses), others stem from the inherent complexity of translating what is fundamentally an English language resource (e.g. incorrect translations of polysemous literals, problems with lexical discrepancies, cross-PoS problem (see Krstev 2004)), yet third are a consequence of the method used, based on the translation of Serbian synsets into Slovene (e.g. untranslated multi-word literals, problems with lexical gaps and denotation differences (see Bentivogli 2000)). The quality of the Slovene wordnet is thus heavily influenced by the quality and consistency of the resources used: the PWN, the Serbian wordnet and the bilingual dictionary.

In the future, we plan to increasingly use automated means to acquire and refine Slovene synsets by:

[1] http://multiwordnet.itc.it/english/home.php

[2] http://www.ceid.upatras.gr/Balkanet/

[3] http://www.fida.net/slo/index.html

[4] The three mismatched synsets have been identified and will be subsequently added by hand to the Slovene wordnet.

Table 1: Comparison of the number of synsets across POS in the three wordnets

	Slo WN	SWN	PWN
BCS1			
nouns	965	965	964
verbs	254	254	254
adjectives	0	0	0
adverbs	0	0	0
total	1219	1219	1218
BCS2			
nouns	2245	2245	2246
verbs	1188	1188	1188
adjectives	36	36	37
adverbs	0	0	0
total	3469	3469	3471
BCS3			
nouns	94	1187	2686
verbs	59	173	876
adjectives	0	135	265
adverbs	0	0	0
total	153	1495	3827
total synsets	4841	6183	8516

- extracting terms from existing available Slovene terminological lexica and other glossaries, such as EUROVOC;
- using multilingual parallel corpora, such as the EU ACQUIS corpus (Erjavec et al. 2005) to extract bilingual lexica, and use those to find Slovene translation equivalents of English literals.
- extending the corpus-based approach from single-word literals to phrases by extracting appropriate collocations.

Finally, we also need to consider formalised ways of evaluating the progress of the Slovene wordnet, and to identify possible application areas.

References

Bentivogli, L. Pianta, E. Pianesi, F. (2000): Coping with lexical gaps when building aligned multilingual wordnets. In Proc. of LREC 2000, Athens, Greece.

Erjavec, T. Ignat, C. Pouliquen, B. Steinberger, R. (2005): Massive multilingual corpus compilation: Acquis Communautaire and totale. In Proc. of the 2nd Language & Technology Conference, Poznan, Poland.

Horák, A. Smrž, P. (2004): New Features of Wordnet Editor VisDic. Romanian Journal of Information Science and Technology. 7/1–2, pp. 201–213.

Jurančič, J. (1991): Srbskohrvatsko-slovenski in slovensko-hrvatskosrbski slovar (*Serbo-Croatian / Slovene and Slovene / Serbo-Croatian Dictionary)*. Ljubljana: DZS.

Krstev, C. Pavlovic-Lažetic, G. Vitas, D. Obradovic, I. (2004): Using textual resources in developing Serbian wordnet. In Romanian Journal of Information Science and Technology. 7/1–2, pp. 147–161.

Obradovic, I. Krstev, C. Pavlovic-Lažetic, G. Vitas, D. (2004): Corpus Based Validation of WordNet Using Frequency Parameters. In Proc. of Second Intl. WordNet Conf. 2004, Brno, Czech Republic, pp. 181–186.

Vossen, P. (ed.) (1998): EuroWordNet: A multilingual database with lexical semantic networks. Dordrecht: Kluwer Academic Press.

Vossen, P. (ed.) (2005): EuroWordNet. General Document (Ver. 3. Final. Oct 5 2005). `http://www.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip`