Domain Specific WordNet on Customs Law

Zoltán Alexin
Department of Software Engineering
University of Szeged
Árpád tér 2., Szeged, Hungary
alexin@inf.u-szeged.hu

János Csirik
Research Group on Artificial Intelligence
Hungarian Academy of Sciences and
University of Szeged
Aradi vértanúk tere 1. Szeged
csirik@inf.u-szeged.hu

Attila Almási
Department of Informatics
University of Szeged
Árpád tér 2., Szeged, Hungary
vizipal@gmail.com

Veronika Vincze
Department of Informatics
University of Szeged
Árpád tér 2., Szeged, Hungary
vinczev@inf.u-szeged.hu

Abstract

The NLP research group at the University of Szeged took part in the development of the Hungarian WordNet between 2005 and 2007. In 2008, they developed a smaller, domain specific WordNet on customs law. This knowledge base contains about 650 concepts cautiously selected by legal experts from the relevant Hungarian statutory legal texts, above all, from two acts and from other laws and decrees. The resulted hierarchic net of concepts is used in an information retrieval system for quick access to documents that have been previously indexed according to the concepts. In addition to this, the WordNet can be used in the daily routine work or in the training of customs officers as it contains detailed definitions of concepts and precise references to legal places where the given concept is defined. Although the WordNet is not a general legal ontology, it shares common concepts with the LOIS multilingual legal WordNet.

1 Introduction

Hungary became full member of the Schengen Treaty in December of 2007.[1] At the same time, Hungary became the Eastern gate of the European Union. The security of the whole Schengen Zone demands empowering the border guard and tightening passenger control. This includes the improvement of the existing information systems.

A consortium led by the Montana Knowledge Management Ltd.[2] won the support of the National Office for Research and Technology[3] in 2008. During the project named TUDORKA7, the *InfoVadász* document repository and retrieval system developed by the Montana Ltd. was to be tailored to the requirements of the project. The purpose of the planned system was to give the necessary help in the fight against crimes such as drug trafficking, smuggling, excise duty crimes etc. by providing simple and fast access to a large number of documents, legal resources, warrants, protocols, and reports – sometimes written in English, German, or French.

A multilingual knowledge base within this international environment should be a central part of the information system. A knowledge base that contains the most important customs law concepts in several languages in parallel would make search in the document sources easier. In EuroWordNet technology, Inter-Lingual Indices (ILI) connect concepts in different languages and provides an excellent representation method for such a knowledge base (Alonge et al., 1998).

The authors of this paper took part in the development of the Hungarian WordNet (Alexin, Csirik et al, 2006; Miháltz, Hatvani, et al. 2008)

[1] http://abiweb.obh.hu/abi/pdf/Schengen.pdf

[2] http://www.montana.hu/index.php?lang=en

[3] http://www.nkth.gov.hu/english

containing 40 thousand general synsets. They were also involved in a project that created a domain specific extension to this WordNet from 3000 concepts related to economy.

The planned document repository and retrieval system would store not only the documents themselves but relevant metadata like creation date, author, and the language of the document. After providing translations of concepts to different languages, the querying subsystem would apply translated query expressions i.e. list of translated keywords corresponding to the language of the selected document when it computes the fitness or relevance measure to the query. The advantages of a parallel language knowledge base can be exploited in the above manner.

Researchers were unable to create a complete multilingual knowledge base within the given time frame, therefore, they started looking for a multilingual database to connect to, possibly a WordNet knowledge base that would serve as a basis for a Hungarian ontology. The LOIS (Legal Ontologies for Knowledge Sharing) multilingual legal WordNet was that database.

Legal experts found better reasons to create such a knowledge base. The *gloss*/*definition* field can contain not only an informal but the official definition of concepts that usually originates from legal rulings like laws, decrees, or commands. In the *note* field the exact legal reference can be provided.

During the creation of the knowledge base, legal experts had an opportunity to scrutinize the legal texts. This way, they explored conflicts or ambiguities between the definitions. Sometimes a concept is defined in two legal places in a different manner, or the Hungarian rule does not correspond exactly to the EU principles. Several deficiencies like this were found during the work.

In the following sections this customs law WordNet is presented. It consists of about 650 concepts related to customs crimes, excise duty crimes and taxation procedure. It contains the relevant official definition of the concepts whenever it exists, as well as the exact legal reference to the place of definition. If there is more than one relevant definition for a concept it is also marked in the knowledge base, which can be considered as a Hungarian part of the LOIS WordNet although its topic is somewhat different. The Hungarian concepts are connected to their LOIS WordNet counterparts by ILI indices. Whenever customs law concepts have hyponyms in the general Hungarian WordNet it is also marked in the semantic relation field.

2 The LOIS Legal WordNet

The LOIS (Legal Ontologies for Knowledge Sharing) multilingual WordNet was created during an EU funded project EDC 22161 between 2003 and 2006 (Dini, Peters, et al. 2005, Peters, Sagri and Tiscordia 2007). The LOIS consortium was led by the Italian Institute of Legal Information Theory and Techniques in Florence. After a short negotiation a research agreement between the Institute of Informatics at Szeged and the LOIS consortium was signed according to which, Hungarian researchers were granted access to the LOIS multilingual legal WordNet.

The LOIS WordNet originally contained 35000 concepts in five European languages (English, German, Portuguese, Czech and Italian), roughly 7000 concepts in each.

```
<WORD_MEANING ID="1429"
 PART_OF_SPEECH="N" STATUS="FINISHED">
 <SOURCEBASE>LEXDB</SOURCEBASE>
 <NOTE/>
 <GLOSS>a person who has not reached
full legal age</GLOSS>
 <CONCEPTS/>
 <VARIANTS>
   <LITERAL LEMMA="minor" SENSE="1">
    <EXAMPLES>not of legal age; &quot;
minor children&quot;</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="minor" SENSE="1">
    <EXAMPLES>a person who has not
reached full legal age; a child or juve-
nile</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="juvenile" SENSE="1">
    <EXAMPLES>a person who has not
reached the age (usually 18) at which
one should be treated as an adult by the
criminal justice system</EXAMPLES>
   </LITERAL>
 </VARIANTS>
</WORD_MEANING>
```

Fig. 1. The concept of *juvenile* as defined in the LOIS WordNet

The LOIS WordNet uses its own Inter-Lingual Indices to identify the concepts (synsets). The IDs of the semantically identical synsets are the same in each of the five languages. Synsets, mostly nouns, are taken from the general legal science and there are few verbs, adjectives and adverbs. Generally, each synset has a definition which sometimes comes from Celex[4], the legal

[4] http://eur-lex.europa.eu/en/index.htm

document repository of the EU or from legal handbooks. In Figure 1 an example of a LOIS synset is shown.

3 The customs law WordNet

In the framework of the customs law WordNet project, the researchers from Szeged first began to collect a term vocabulary from Hungarian legal texts by automatic methods. The consortium finally decided that two acts should be processed: Act on taxation procedure[5] and Act on excise duty[6]. Legal experts from the Department of Constitutional Law were invited to the project. They manually checked the terminology and advised to augment them with other important terms e.g. from the Penal Code. Unfortunately, they had no other digitized resource to begin with. Later the consortium asked the researchers from Szeged to add further terms from the publicly available commands of the Commissioner. When the list of terms was finalized, legal experts began to collect glosses. The related laws, decrees and legal handbooks were systematically thumbed over. If more than one gloss was found for a term, then all explanations – having made a record of their source – were included in the knowledge base.

```
<SYNSET>
  <ID>HuWN-911671085</ID>
  <SYNONYM>
  <LITERAL>fiatalkorú
     <SENSE>0</SENSE>
  </LITERAL>
  </SYNONYM>
    <DEF>Fiatalkorú az, aki a bűncselek-
mény elkövetésekor tizennegyedik élet
évét betöltötte, de a tizennyolcadikat
még nem.</DEF>
  <SNOTE>1978. évi IV. tv. Btk. 107.§.
(1)</SNOTE>
  <SNOTE>LOIS ID="1429"; a magyar jog-
rendben kis- és fiatalkorú megkülön-
böztetés létezik</SNOTE>
  <SNOTE>jog</SNOTE>
  <POS>n</POS>
  <ILR>HuWN-148541600
     <TYPE>hypernym</TYPE>
  </ILR>
</SYNSET>
```

Fig. 2. The concept of *fiatalkorú* (*juvenile*) as defined in the customs law WordNet

When the term vocabulary was finished, computational linguists together with legal experts ordered the terms in a hierarchy. The originally paper-based notes and Microsoft Excel spreadsheets were compiled into a WordNet by linguists using the VisDic editor program (Horák and Smrž, 2004). Principally, the hypernymy relation was implemented but also holonymy occurred several times.

The <DEF> node (gloss) contains the definition of the synset, which legal experts usually took from an act being in force or from legal handbooks. The part-of-speech of the synset is marked in the <POS> node. Synonyms of a term were collected from legal handbooks. In several cases, synonyms were multiword expressions due to the characteristics of the legal terminology. Linguistic relations like hypernymy or holonymy were coded in <ILR> nodes. The <ID> nodes contain the ILI indices of the synsets.

In Figure 2 an example of a synset from the Hungarian customs law WordNet is shown. It can be seen, that the Hungarian counterpart of the LOIS synset "juvenile" has a Hungarian WordNet <ID> due to the fact that the customs law WordNet was made as an extension to the Hungarian WordNet.

In the first <SNOTE>, one can find the exact reference to the legal place where the gloss is taken from, namely Penal Code (Law IV. of 1978.), section 107. In the second <SNOTE>, the LOIS ILI index and an explanation in Hungarian are included.

3.1 Conflicts between linguistic and legal requirements

When building the WordNet it was often found that the requirements of linguistics and law were contradictory so researchers had to make priorities. It was decided that, first, they meet the requirements of law and, then, take linguistics into consideration where possible.

As a consequence, the customary linguistic rule applied in WordNets that the definition of a synset must contain a hypernym of the concept or its synonym (Miller et al., 1990) has been modified for, in most cases, definitions are mere lists of words.

In the Hungarian WordNet (Alexin et al., 2006; Miháltz et al., 2008), within synsets, notes are units that make short, supplementary comments possible. However, in the customs law WordNet notes have been given a new function. They are used to include information that cannot be entered as a part of the definition but provide substantial, indispensable data e.g. exact place of the definition in the legal texts, numerical data

[5] Hungarian Act no. XCII. of 2003.
[6] Hungarian Act no. CXXVII. of 2003.

(e.g. alcohol concentration, quantity of importable goods, etc.)

When creating the hierarchy, the *bottom-up* method was followed because concepts derived from legal sources proved to be rather specific and they were usually used to create base-level synsets only. This, however, made the work simpler because hypernyms could be selected relying on the hierarchy of Hungarian WordNet.

In the customs law WordNet there are nine *unique beginner* synsets. Due to the decision mentioned above, it may happen that an element identified as an object on the base-level gets linked to a non-object hypernym synset or occurs in the tree of the *unique beginners* e.g. *abstraction* or *state*. This linguistically indefensible state was impossible to eliminate. Due to the phraseology of law these apparent "inconsistencies" have remained.

4 Connections between the Hungarian customs law WordNet and the LOIS Legal WordNet

The last step of the work was to establish connections between the two WordNets. Legal experts examined the English version of the LOIS WordNet and produced a list of synsets that may have connections to the customs law WordNet. A linguist and a legal expert then – taking the definitions into consideration – checked manually the list item by item to figure out whether the relation between the two concepts is valid, It was also checked whether the LOIS synset was more general than the synset in the customs law WordNet. In several cases the LOIS WordNet did not contain glosses for the synsets therefore the decision on identicality could not be made.

When the two synsets proved to be undoubtedly identical, the connection has been marked in the *note* field of the synset in the customs law WordNet as follows: LOIS ID="nnnn", where nnnn is the ILI index of the corresponding synset in the LOIS WordNet. A short explanation was also added. See Figure 2.

	Connected to LOIS	Cannot be connected to LOIS	All
General legal synset	81	116	197
Excise duty synset	113	337	450
Total	194	453	647

Table 1. The number of connections between the customs law WordNet and the LOIS WordNet

In Table 1, statistics on the customs law WordNet is presented. 194 out of the 647 (30%) synsets from the customs law WordNet have a counterpart in the LOIS WordNet. Among them 113 synsets are closely connected to the excise duty terminology (declaration, payment, definitions, crimes etc.), while 81 synsets are general legal terms.

In the whole customs law WordNet, 450 out of the 647 synsets were taken from the excise duty terminology. Their definitions come from legal rulings (laws, decrees, orders, etc.) being in force, e.g. tax warehouse, licensee of the tax warehouse, the onset of tax paying obligation. The remaining 197 synsets are general legal terms with definitions taken from handbooks, e.g. interest, loss, official, representation.

The number of adjectives, nouns and verbs in the two WordNets are shown in Table 2.

	LOIS WordNet (English)	Customs Law WordNet
adjectives	0	0
nouns	6720	647
verbs	51	0

Table 2. The distribution of the adjectives, nouns, and verbs among the synsets of the two WordNets

5 Conclusion

The presented Hungarian customs law WordNet was made with a view of a multilingual information retrieval and query system which would be capable of returning the relevant documents written in any of the supported languages by answering a query expression entered in Hungarian. This task can be accomplished in a narrow semantic domain like customs law. A multilingual knowledge base, a WordNet can provide the semantically correct translations of the most frequent terms.

A multilingual legal knowledge base is essential in international administration, business, insurance, jurisdiction or counseling. This knowledge base should be the kernel of later information systems, and can be used for other purposes as well. For example, it may help to understand the differences between jurisdictions of countries i.e. it may be used as a legal definition (reference) database.

The presented Hungarian customs law WordNet was an experiment on this approach. The developed WordNet has fulfilled our expectations towards such a knowledge base.

Finally, it has a well-founded connection to the LOIS multilingual legal WordNet and to its synsets in several languages. At the same time, it has evoked new questions whether to translate all of the LOIS synsets to Hungarian, or to broaden the current customs law WordNet. Both are possible objectives of forthcoming R&D projects.

Acknowledgments

The authors wish to thank Márton Sulyok, Judit Tóth and other members of the Department of Constitutional Law at the Faculty of Law of the University of Szeged for their contribution to the project.

The research presented in this paper was supported by the TUDORKA7 and MASZEKER projects of the Jedlik Ányos 2007 and 2008 Programs of the National Office for Research and Technology (NKTH, http://www.nkth.gov.hu/) of the Hungarian government.

References

Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy. 2006. Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction, Project report, In: *Proceedings of the Third International WordNet Conference GWC 2006*, South Jeju Island, Korea, 2006, pp. 291–292.

Antonietta Alonge, Laura Bloksma, Nicoletta Calzolari, Irene Castellon, Maria Antonia Marti, Wim Peters, Piek Vossen. 1998. The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities*, 32(2–3):91–115.

Dini, L., Peters, W., Liebwald, D., Schweighofer, E., Mommers, L., and Voermans, W. 2005. Cross-lingual legal information retrieval using a WordNet architecture. In *Proceedings of the 10th international Conference on Artificial intelligence and Law* (Bologna, Italy, June 06–11, 2005). ICAIL '05. ACM, New York, NY, 163–167.

Horák, A., Smrz, P. 2004. VisDic — Wordnet Browsing and Editing Tool, In: *Proceedings of the Second International WordNet Conference GWC 2004*, pp. 136-141.

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T. 2008. Methods and Results of the Hungarian WordNet Project, In: *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008, pp. 311–320.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. 1990. Introduction to WordNet: an Online Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Peters, W., Sagri, M. and Tiscornia D. 2007. The structuring of legal knowledge in LOIS, *Artificial Intelligence and Law*, Volume 15, Issue 2 (June 2007), pp. 117–135. Springer Verlag, ISSN: 0924-8463.