EuroWordNet Annual Report 1998

http://www.hum.uva.nl/~ewn

Introduction

Natural language is not only a medium for storage of information, but also the most natural and flexible interface between (especially non-expert, non-skilled) users and information systems. A very important prerequisite for the success of information services is therefore the availability of general, generic language resources, ranging from morpho-syntactic analysers, lexical databases, to complete grammars, and language-understanding and language-generation modules. The multilinguality of Europe makes it even more important to have such modules for processing natural language, not only for each language but also across languages. Only then will it be possible to develop Telematics services and systems for a common market. The need for such language-resources is most directly illustrated by the situation in information retrieval applications. New developments in integrating language modules are mostly limited to English, mainly due to the missing multilingual semantic resources, which are currently not available for other languages. The technology and market is ready for inclusion of this new functionality, but lacks high quality and quantity of lexical semantic data.

EuroWordNet directly addresses this problem by developing a multilingual database with wordnets for a 8 European languages: English, Dutch, Spanish, Italian, German, French, Estonian and Czech. Each of these wordnets is structured along the same lines as the Princeton WordNet around the notion of a synset. Synset are sets of synonymous word meanings, between which basic semantic relations are expressed, such as hyponymy (*cello – string*), meronymy (*string – fingerboard*), role_instrument (*string* – to *bow*). In addition to the relations between synsets, the so-called language-internal relations, each synset in EuroWordNet is also linked to some Inter-Lingual-Index, thus constituting a multi-lingual database. Word meanings across languages linked to the same index item are assumed to be equivalent in meaning. Figure-1 shows a screen-dump of the Periscope program that provides a graphical interface to the database. Here we see a comparison between the Dutch and (American-)English wordnet and the Italian and Spanish wordnet respectively. Rectangular boxes represent synsets (related to *cello*) in the local languages and the next line contains the Inter-Lingual-Index concept to which they are linked. The lines across the wordnets indicate matching pairs of index references.

The EuroWordNet database makes it possible to match the meanings of words within each language but also across languages. There are many ways in which such a database can be exploited but, within the limits of the project, we will demonstrate the use in (cross-language) information retrieval.

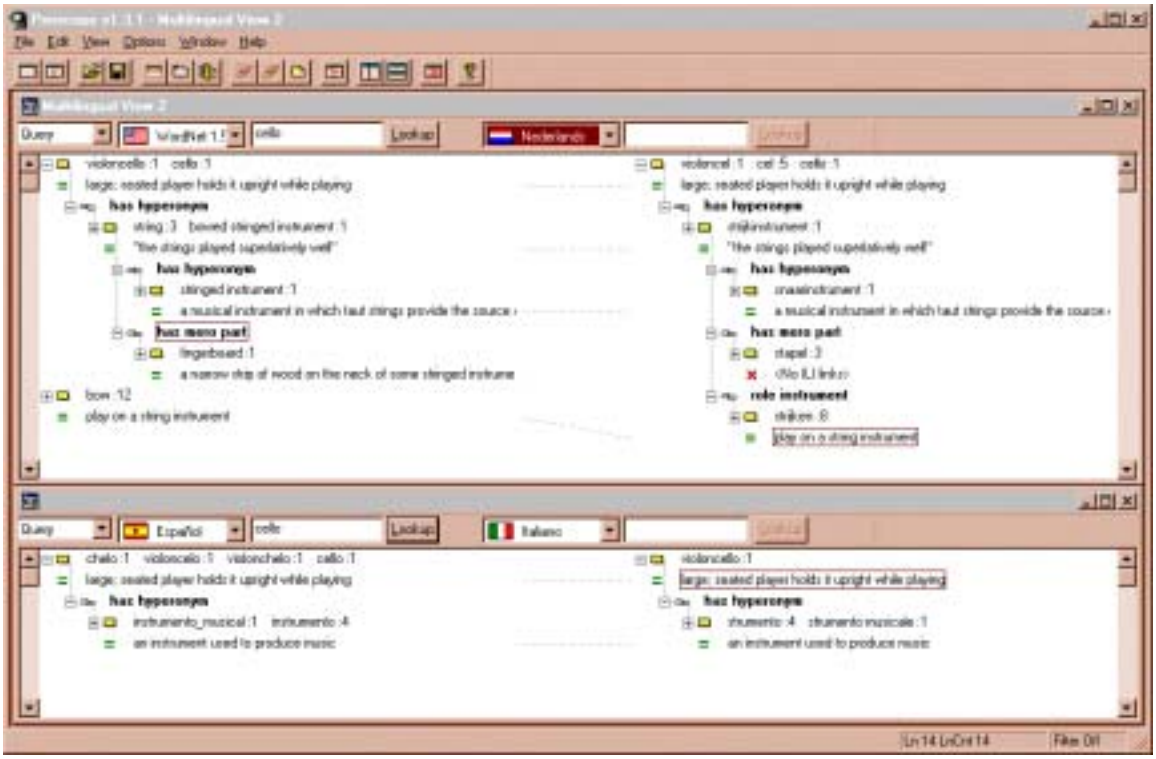

Figure 1: Parallel wordnet structures in EuroWordNet linked to the same ILI-records.

Summary of 1998

The first phase of EuroWordNet concentrated on specification and design. The second phase has resulted in the development of the database, a shared top-ontology with a set of Base Concepts and the core wordnets for Dutch, Spanish and Italian. In 1998, the third phase, the main achievements of the project are:

- development of full wordnets for Dutch, Italian and Spanish.
- extension of the project to develop wordnets for German, French, Czech and Estonian.
- improvement of the Inter-Lingual-Index that connects the meanings across the languages.
- co-operation with ELDA, EAGLES, PAROLE, SIMPLE, and an American-European committee to develop standardised ontologies.
- first experiments using the multilingual database and the EuroWordNet Base Concepts with the Top-Ontology.
- extension of the user-group to 70 members.
- release of a Special Issue on EuroWordNet in the Computer and the Humanities Journal and the release of a CD with free data samples and a viewer.

The wordnets for Spanish, Italian and Dutch have reached their final size and will be completed soon. These results can be licensed early 1999. The other languages have completed their specification phase and have constructed the core wordnets for the most important meanings. Co-operation with other projects and groups has increased this year, which illustrates the impact of the project on many researchers and developers.

The next (final) phase of the project will be devoted to completion of the wordnets for all the languages and the demonstration of the database in cross-language information retrieval applications by the users in the project. Furthermore, we are investigating the possibility to found a European Wordnet Association (EWA), that will provide a longer-term framework for co-operation and for extending to more languages and richer data.

1. COMPLETION OF THE WORDNETS FOR DUTCH, ITALIAN AND SPANISH

In the previous year we developed core wordnets for Dutch, Italian and Spanish[1], around the most important concepts: the Base Concepts. The Base Concepts represent meanings that play a major role in at least 2 other wordnets, and they have been classified by a top-ontology of 63 semantic distinctions, such as Artefact, Living, Physical, Mental. The Base Concepts and the top-ontology provide a common framework and starting point to build the core wordnets. After completion, the core wordnets have been compared to measure overlap in size, coverage, and compatibility of relations. We carried out an in-depth comparison using the EuroWordNet database (Polaris) for 18 semantic fields, and a large-scale comparison for the complete wordnets. For the latter, special tools have been developed. From this comparison we derived measures of overlap (across 2, 3 and 4 wordnets) in concepts, in relations and hyponymy chains. Furthermore, we have been able to extract dubious structures (e.g. extreme long chains which are not shared by other wordnets) due to mistakes in the relations and/or the translations.

Finally, we compared the entries of the wordnets with the corresponding lexicons of the Parole-project with morpho-syntactic information. This showed a high overlap for the most frequent words: words with frequencies above 1000 tokens have 90-97% overlap, 501-1000 tokens 84-90% overlap between the projects. These figures confirm that the most generic vocabulary is also used most frequently.

The result of the comparison have been used to complete the wordnets in Dutch, Italian and Spanish, aiming at full coverage and maximum compatibility among the wordnets and the Parole lexicons. The current size of the wordnets is 50K word senses, which roughly corresponds to 30K synsets (about 1/3 of the size of the Princeton WordNet). For each synset, we provided one or more hyperonyms (its semantic class) and one or more equivalence links to the Inter-Lingual-Index. Where relevant, we tried to encode other types of semantic relations, such as part-whole, causal relations or semantic roles (e.g. agent, patient, location, instrument) but, because of limited resources, this has not been done comprehensively.

[1] For English, we followed another strategy. Since there is already a wordnet for English we focussed on extending the existing wordnet with new relations.

2. EXTENSION WITH WORDNETS FOR GERMAN, FRENCH, CZECH AND ESTONIAN

The database has been extended with wordnets for French, German, Czech and Estonian. The new partners first of all verified the suitability of the EuroWordNet design and database for these languages. This implied redefining the set of Base Concepts on the basis of the most important meanings in the new wordnets and implementing some examples in the database.

The specification of the wordnets for the new languages did not raise major problems. After the set of Base Concepts has been extended, the construction of the core wordnets was started. The current coverage is between 3K-8K synsets per language. For each synset at least a hyperonym and an equivalence link are encoded, and optionally other relations.

In the next phase we will carry out a similar comparison of the wordnets as described above. The results of this comparison will be used for the completion of the wordnets.

3. ADAPTATION OF THE INTER-LINGUAL-INDEX

Separately from the development of the wordnets, we have begun to improve the Inter-Lingual-Index. The sole purpose of the index is to provide an efficient mapping across the languages. In this respect it should be the superset of all the concepts occurring in the different languages but, at the same time, the meaning distinctions should not be too fine-grained.

At the beginning the Inter-Lingual-Index was based on the English WordNet1.5, but it has been extended when concepts occur in two languages but not in wordnet. This is necessary to establish equivalence across these languages.

Furthermore, very close meanings or systematically related meanings in WordNet1.5 have been grouped by more global index-records. When synsets are linked to different but closely related index-records, it is thus still possible to get a matching at a more global level.

Finally, we are linking WordNet1.6 to the Inter-Lingual-Index so that EuroWordNet can be used in combination with either version of the Princeton WordNet.

4. USAGE OF EUROWORDNET IN INFORMATION RETRIEVAL

Even though the demonstration of the database is planned in the final phase of the project, there have already been experiments to use the results. At the University of UNED (Madrid), who is also a partner in the project, experiments have been carried out to use EuroWordNet in (cross-language) information retrieval (Gonzalo, J., F. Verdejo and I. Chugur 1999). These experiments show an increase of 29% retrieval in English, using synsets instead of word forms. If the same task is carried out cross-linguistically using the EuroWordNet database, the results only drop 20%. The current base-line of cross-language retrieval using a bilingual dictionary results in a decrease of 50%. Using EuroWordNet we thus improved the performance with 30%.

Another experiment by Dorr, M.A.Martí & I. Castellón (1988) reports promising results on generating translations between English and Spanish using EuroWordNet extended with Semantic Components.

5. CO-OPERATIONS OUTSIDE THE PROJECT

We have extensively co-operated with other projects to establish agreements and standards on the specification of lexical semantic data. This has resulted in contributions to the EAGLES project, which, among others, aims at developing guidelines for lexical semantic resources for Natural Language Processing, and contributions to the ANSI committee for ontology standards, which aims at developing a standardised Reference Ontology that can be used in Information Systems. Both co-operations show that there is world-wide belief in the feasibility to process the information-content for realistic applications.

The role of EuroWordNet in these co-operation has been to provide specifications and definitions for the design of these standards and guidelines. Especially, the fact that EuroWordNet represents an efficient framework for sharing semantic specifications makes its contribution unique. Any standardised ontology or lexical semantic resource that is properly linked to the Inter-Lingual-Index can be shared by all the languages linked to the index via their wordnet. We therefore believe that the EuroWordNet architecture can be used in the future to develop task-specific resources for many European languages in a cost-effective and consistent way.

Other co-operations are established with the Parole project (see above) but also with Simple and Trevi (ESPRIT-23311). The latter two projects aim at developing more specific databases with semantic information. Both projects have used the Base Concepts and the EuroWordNet top-ontology for their specification. In collaboration with ELDA, users-licenses have been developed. ELDA will also act as the distributor of the resources.

Finally, some of the partners have established national co-operations to further develop their wordnets and wordnets for other languages in their country, following the specification of EuroWordNet.

User Group, Promotion and Awareness

A major dissemination event this year has been the release of a CD with samples from the Dutch, Italian and Spanish wordnet that can be accessed with the graphical viewer Periscope. The CD has been distributed for free at the LREC conference in Granada and the data can also be downloaded from the www-site of EuroWordNet. In the near future we will update the CD with samples from the other languages.

Another important release is the special issue of the Computers and Humanities Journal on EuroWordNet (Vol 32, Nos 2-3, 1998), which is also released as a book. This volume gives a very good introduction in many aspects of the project.

Except for the regular meetings with individual users, we have attended major workshops and conferences, among which: SENSEVAL (Brighton), LREC (Granada), ACL/ALLC, EURALEX (Liege) and ACL/COLING (Montreal).

The user-group has now grown to 70 members (45 Universities and Research Institutes, 25 Companies and end-users), and many request to use the data have been received by the separate partners. Globally, we can distinguish the following parties among the users:

- publishers, either providing the initial resources or interested in the development of similar products.

- research institutes and R&D departments of universities and companies working in the field of knowledge engineering or linguistic databases, interested in building similar resources.

- research institutes and R&D departments of universities and companies working in the field of Language Engineering interested in using or applying similar resources or developing services or products that make use of multilingual semantic resources: especially information processing.

- end-users interested in products helping them to deal with information.

In addition to the use for (cross-language) information retrieval, there are many other applications that can directly benefit from multilingual semantic resources: information-acquisition tools, authoring-tools, language-learning tools, translation-tools, summarizers. Within the user-group there are also many institutes developing wordnets according to the EuroWordnet specification. Wordnets developed in collaboration with EuroWordNet cover the following languages: Basque, Catalan, Portuguese, Lithuan, Swedish, Russian, Greek.

Future Work

The most important work for the future will be completion of the wordnets for the new languages and the integration of the EuroWordNet data in the end-user applications of the users in the project: Bertin (Paris), Xerox (Grenoble) and Lernout and Hauspie (Antwerp). Furthermore, we will continue to investigate how the database can be exploited for other Natural Language Processing Tasks, such as Word Sense Disambiguation, Language-Learning, Summarisers, Machine-Translation.

We expect that the user-licenses will be ready by January 1999 and that the first data can be licensed in April 1999 from ELDA. As suggested above, we are starting a EuroWordNet association (EWA) to ensure future extension, information sharing and standardisation of lexical semantic resources within the EuroWordNet framework.

References

Gonzalo, J., F. Verdejo and I. Chugur 1999. Using EuroWordNet in a Concept-Based. Approach to Cross-Language Text Retrieval, in Applied Artificial Intelligence, Special on Multilinguality in the Software Industry: the AI contribution.In press.

B. Dorr, M.A.Martí & I. Castellón. 1998. Spanish EuroWordNet and LCS-Based Interlingual MT". Proceedings of 1st International Conference. European Resources and Evaluation, LREC'98, Granada, 1998

Sanfilippo A., N. Calzolari, S. Ananiadou, R. Gaizauskas, P. Saint-Dizier, P. Vossen (eds). 1998. EAGLES, Preliminary Recommendations on Semantic Encoding. Interim Report.

Nancy Ide, Daniel Greenstein, Piek Vossen (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998, Kluwer Academic Publishers, Dordrecht.

Vossen, P. (eds) 1998 EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht

Antonio Rubio et al. (eds.) First International Conference on Language Resources & Evaluation Proceedings, Granada 28-30 May 1998.

Project WWW references

EuroWordNet: http://www.hum.uva.nl/~ewn

Simple, Parole: http://www.ilc.pi.cnr.it/parole/parole.html

Eagles: http://www.ilc.pi.cnr.it/EAGLES/home.html

Elra: http://www.icp.inpg.fr/ELRA