EuroWordNet Annual Report 1997

Summary of the Year

After the completion of the design of the lexical database, EuroWordNet has mainly worked on the implementation of the EuroWordNet database and the encoding of the lexical semantic data. We defined a set of most important concepts, the so-called Common Base Concepts, which is shared by all the wordnets. These Base Concepts have been classified using a top-ontology with basic semantic distinctions, specifically developed for this purpose. The set of Base Concepts and their top-ontology classification has been used as the common framework to develop the separate wordnets. Now the database has been completed and the first wordnets in 4 different languages have been loaded and compared. These fragments include the most important meanings on which most other meanings depend, making up the core of the total wordnets.

The database design has been discussed at various conferences and workshops. This has clarified our position that the wordnets are autonomous language-specific networks, reflecting the unique lexicalization patterns of languages. The discussions showed that our flexible design gives many advantages to address the complexity of a multilingual database and to capture the differences in lexical expressibility of languages.

In addition, we have started cooperation with the American Ansi Group on Ontology Standards and the European EAGLES Special Interest group on Lexical Semantic Resources. Finally, we have extended the project to include other European languages: French, German, Czech and Estonian.

The final year of EuroWordNet will be devoted to completing the wordnets and including the new languages.

The Multilingual Database

The EuroWordNet database has been completed. The basic design is illustrated in Figure 1 below. In this Figure each language-specific wordnet is represented as a separate database with language-internal relations between the word meanings (represented as synset: set of synonyms in a language). In addition to the language-internal relations, each meaning is related to the closest concept in the so-called Inter-Lingual-Index or ILI. This index is an unstructured list of concepts mainly taken from WordNet1.5 and adapted to improve the mapping of word meanings across the wordnets. Via the ILI it is possible to go from one wordnet to another wordnet, but also to get access to the top-ontology and a domain-ontology.

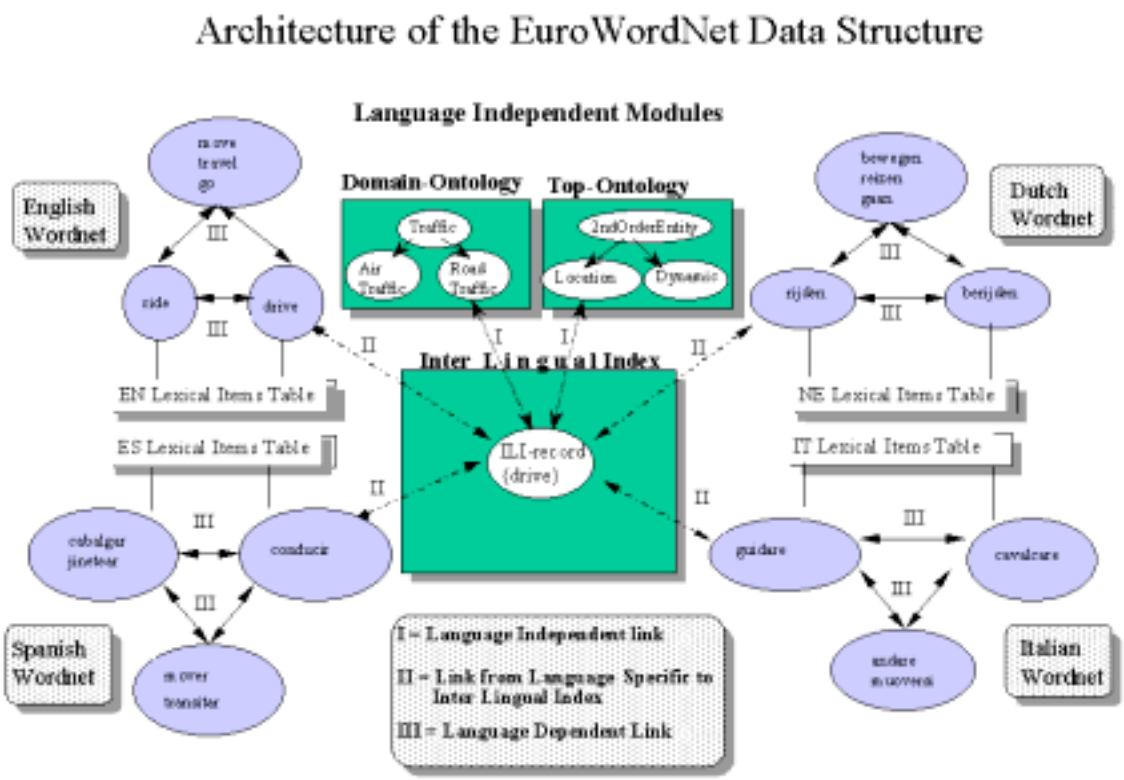

Figure 1: The overall architecture of the EuroWordNet database.

The database itself comes in two packages:

1. Polaris: a tool for building and editing wordnets which can be licensed from Novell Linguistic Development (Antwerp).
2. Periscope: a free graphical viewer which can be obtained from the WWW: http://www.let.uva.nl/~ewn

Both programs can be used to access and compare the wordnets. The next screen dump from Polaris shows how semantic clusters of wordnet meanings can be compared using a projection function which generates the word meanings in another language related to the same ILI-records.

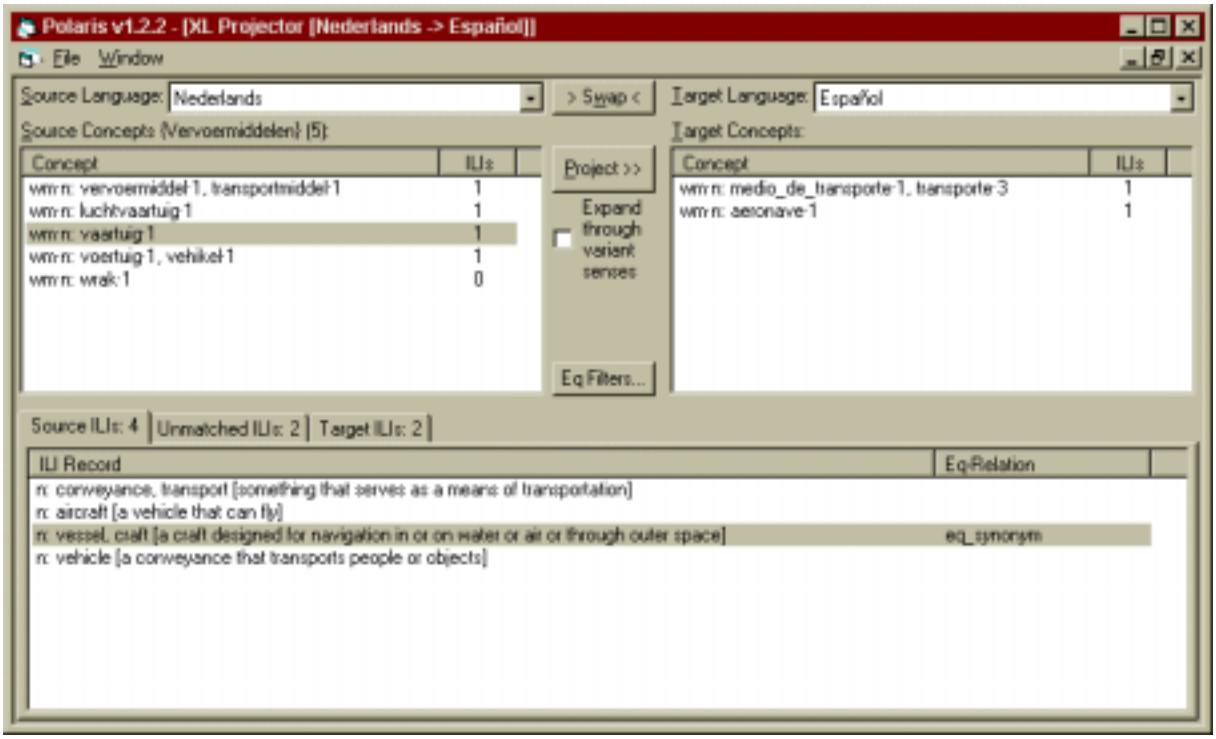

Figure 2. Projection of vehicles in the Dutch wordnet to the Spanish wordnet.

The common Base Concepts and the Top-Ontology

To ensure compatibility of the wordnets we have specified a set of Base Concepts which are shared by all the wordnets. The Base Concepts have been selected on the basis of their importance in the lexical resources that are available for building the wordnets. The main criteria have been the number of relations with other concepts or word meanings and their position in the hierarchies. The total set of Base Concepts is 1059 WordNet1.5 synsets.

To get to grips with these Base Concepts we have developed a top-ontology with 63 basic semantic distinctions, such as Artifact, Natural, Object, Substance, Experience, Mental, Physical, Location, Cause. The first level is divided into:

- 1stOrderEntities: all concrete things, only named by nouns: e.g. *plant*, *vehicle*, *place*, *substance*.
- 2ndOrderEntities: all Dynamic and Static Situations, which can be referred to by nouns, verbs and adjectives: e.g. *live*, *life*, *kill*, *die*, *death*, *dead*, *poor*, *rich*, *gain*, *loose*.
- 3rdOrderEntities: all ideas, concepts and knowledge that exists in mind, only named by nouns: e.g. *idea*, *thought*, *plan*, *information*, *knowledge*.

Some of the distinctions are exclusive: they cannot be combined, while others can be combined into more specific combinations of top-concepts. Figure 3 gives some examples of Base Concepts that are clustered by combinations of top-concepts. The Base Concepts and the Top-Ontology are freely available and can be downloaded from the WWW-site.

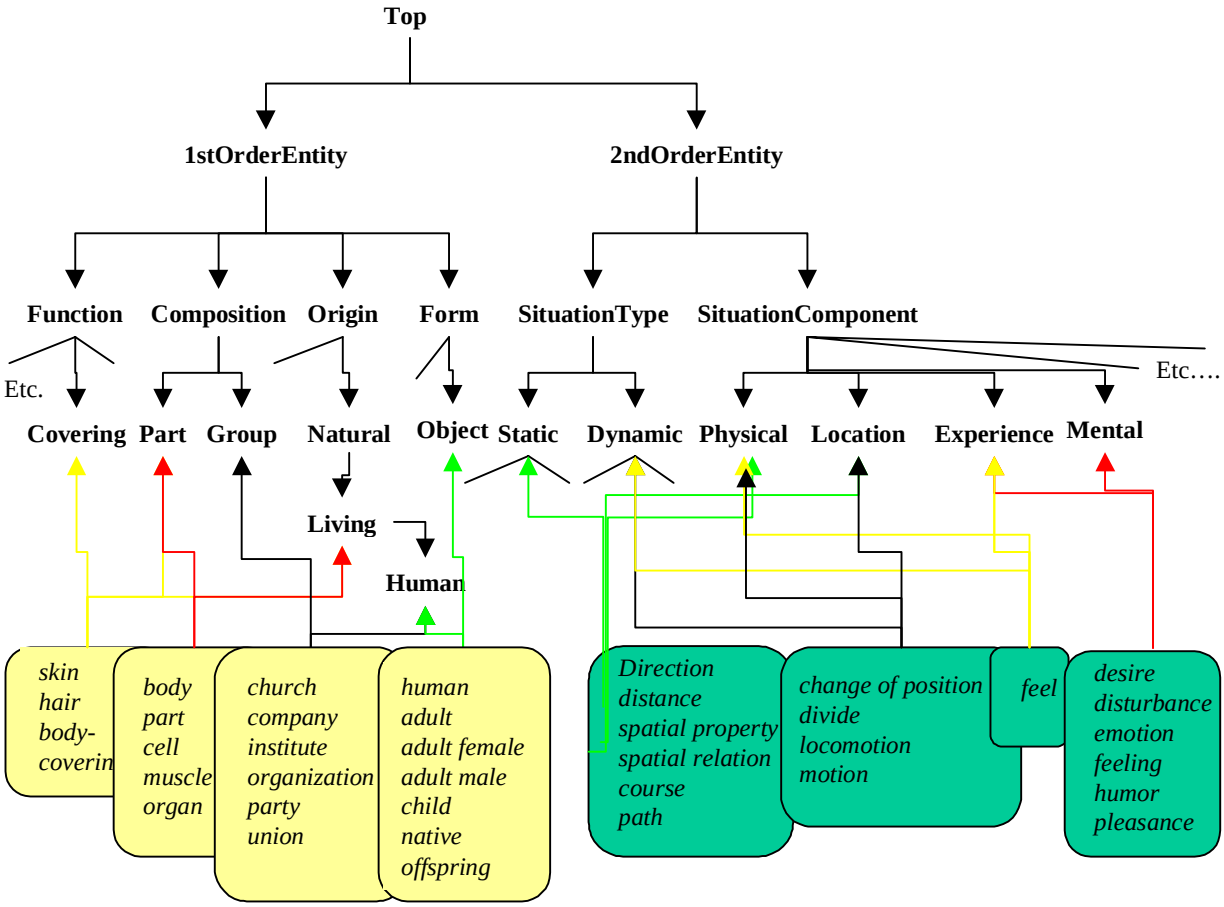

Figure 3: Clustering of Base Concepts by the EuroWordNet top-ontology.

The core wordnets

We have created core wordnets in Dutch, Italian and Spanish. These core wordnets include:

- All word meanings linked to the common set of Base Concepts (1059 synsets).
- All relevant hyperonyms of the word meanings related to the Base Concepts.
- The most important hyponyms of the word meanings related to the Base Concepts.

This selection has been extended in various ways, depending on the local situations and approach, aiming at a final set of at least 10,000 concepts and 20,000 synonyms (where a synset represents a concept). For each of these concepts the following information has been minimally specified:

- Hyperonyms
- Synonyms (synset members)
- Equivalence relations to the ILI (mostly a WordNet1.5 synset)

Optionally, any other relation has been added. The Polaris screen dump on the next page shows two wordnet fragments of the Dutch and Spanish wordnet. The current core wordnets contain between 10,000 and 20,000 concepts with 2 up to 3 language internal relations as average. In the case of the English wordnet a slightly different approach has been followed. Because an English wordnet already exists, we only generated the relations which have been added in EuroWordNet with respect to WordNet1.5. Samples of the core wordnets will be

available on the WWW-site. The full core wordnets can be obtained from ELRA in the near future.

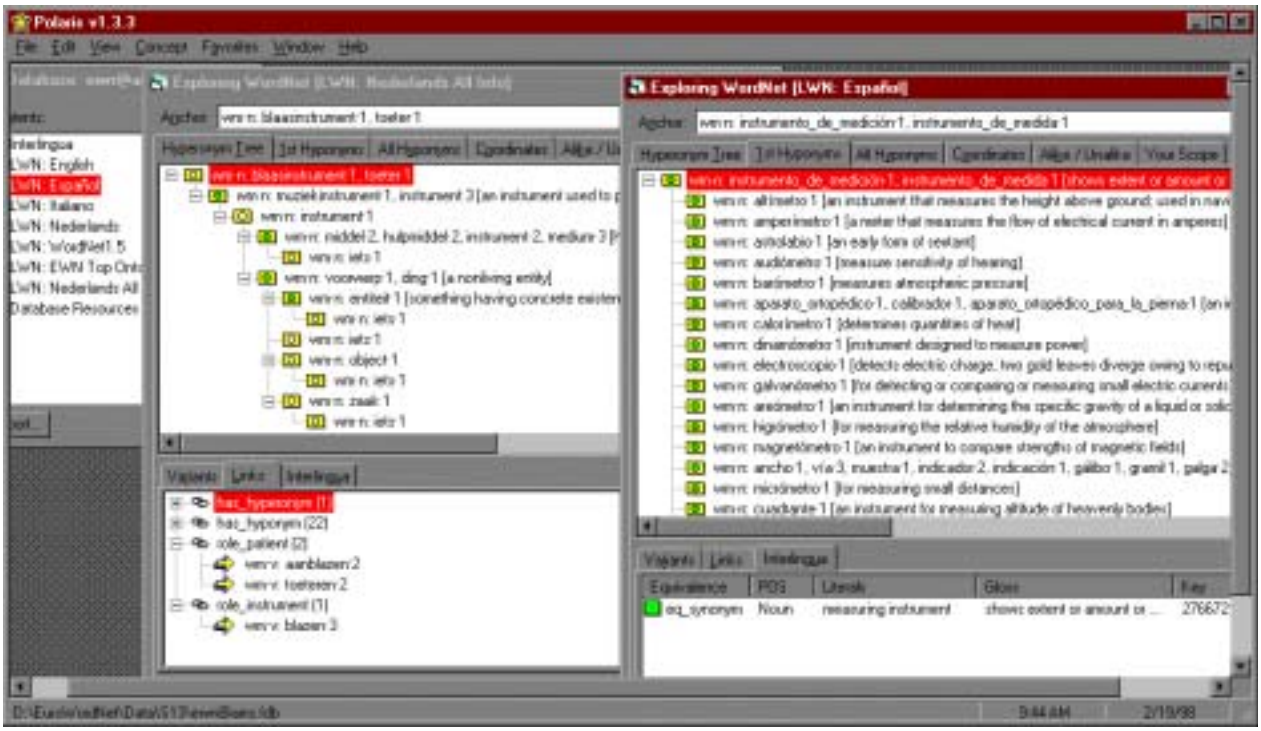

Figure 4: Polaris representation of Spanish and Dutch wordnet fragments.

User Group, Promotion and Awareness

In addition to various conferences and publications, we organized a workshop at the ACL/EACL-97 in collaboration with two LE1 projects: Sparkle (LE1 2111) and Ecran. The title of the workshop was: “Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications”. The workshop aimed at a discussion on the scope and formats of semantic resources and information acquisition tools with scholars in the field and researchers from commercial R&D departments who have experience in developing and using them. The main topics of the workshop are:

1. compatibility and standards of multilingual semantic resources and lexical acquisition tools.
2. the validation of multilingual semantic resources and lexical acquisition tools.
3. performances of semantic resources and lexical acquisition tools in NLP tasks.
4. partial or phrasal parsing of text.
5. linking text with lexical databases: sense-differentiation, sense-tagging and sense-disambiguation tasks, domain-differentiation of text and lexical resources.

At a more global level, we have had various meeting and exchanges of information with the EuroWordNet User-Group. The user-group now comprises almost 50 members from a wide variety of institutes. Some of the users have joined the project extension, others have attended the workshop mentioned above.

EuroWordNet User-Group, d.d. January 1998.

Organisation	*First Name*	*Last Name*	*City*	*Country*
Bertin	J.P.	Marteau	Plaisir Cedex	France
Bibliograf	Juan P.	Acordagoicoechea	Barcelona	Spain
Cambridge Language Services Ltd.	Susan	Allen-Mills	Cambridge	Great Brittain
Cap Volmac	Pim	Van Der Eijk	Utrecht	Netherlands
Centre Of Computational Linguistics	Ruta	Marcinkeviciene	Kaunas	Lithuania
Centro De Investigacions Linguisticas E Literarias "Ramon Pineiro"	Fernando	Magan	Santiago De Compostela	Spain
Coach House South	Sue	Atkins	East Sussex	Great Brittain
Datamat	Gian Franco	Abbrescia	Rome	Italy
Dept Of CS, Carnegie Mellon Univ	Doug	Beeferman	Pittsburgh	USA
EBSCO Subscription Service	Cristina	De La Pena	Madrid	Spain
IBM	Louis	Sopena	Madrid	Spain
IMC Bureau	Frits	Van Latum	Rotterdam	Netherlands
Incyta S.L.	Juan A.	Alonso	Cornella	Spain
Infolab, Tilburg University	Dr. Hans	Weigand	Tilburg	Netherlands
IRIT-CNRS	Patrick	SAINT-DIZIER	Toulouse Cedex France	France
Kenniscentrum Cibit	Cor	Baars	Utrecht	Netherlands
Komet Multilingual Text Generation	Dr. John A.	Bateman	Darmstadt	Germany
Laboratorio De Lingüística Informática Autónoma De Barcelona Edificio Bellaterra, Españ	Carlos	Subirats-Rüggeberg	Bellaterra	Spain
LOGOS-GROUP	Cinzia	Bazzani	Modena	Italy
Office Line Engineering NV	Werner	Ceusters	Zonnegem	Belgium
Oxford Logic, Inc.	Frank	White		Scotland
Oxford University Press	Patrick	Hanks	Oxford	Great Brittain
RKD	J.H.E.	Van Der Starre	Den Haag	Netherlands
SENA	Perikles	Tsahageas	Filothei	Greece
Sharp Laboratories Of Europe Ltd., Oxford	Antonio	Sanfilippo	Oxford	Great Brittain
Sun Microsystems Laboratories	Philip	Resnik	Chelmsford	USA
TNT Express Worldwide	Joy	Herklots	Amsterdam	Netherlands
UNED/Dpto. Ingenieria Electrica	Prof. Felisa	Verdejo	Madrid	Spain

Organisation	*First Name*	*Last Name*	*City*	*Country*
Univ. Do Minho, Dept. De Informatica	Jose Joao	Diaz De Almeida	Braga	Portugal
Universidad Alfonso X El Sabio	Aurora	Martmn De Santa Olalla	Madrid	Spain
Universidad Autonoma De Madrid	Dr. Antonio	Moreno Sandoval	Madrid	Spain
UNIVERSIDAD EUROPEA DE MADRID	Manual	BUENAGA	Madrid	Spain
Universitat Tubingen	H.	Feldweg	Tubingen	Germany
University Of Athens	Stelios	Piperidis	Athens	Greece
University Of Dundee, Microcentre	Stefan	Langer	Dundee	Great Brittain
University Of Euskal Herriko/ Informatika	Arantza	Diaz De Llarraza	Donostia	Basque Spain
University Of Goteborg	J.	Jarborg	Goteborg	Sweden
University Of Heidelberg, Institute For Computer Linguistics	Peter	Hellwig	Heidelberg	Germany
University Of Lund, Department Of Linguistics & Phonetics	Åke	Viberg	Lund	Sweden
University Of Princeton, Cognitive Science	Christiane	Fellbaum	Princeton	USA
University Of Tartu - Dept. Of General Linguistics	Prof. Haldur	Oim	Tartu	Estonia
USC Information Sciences Institute	Eduard	Hovy	Admiralty Way	USA
Van Dale Lexicografie B.V.	Margreet A.	Moerland	Utrecht	Netherlands
VPRO	Bruno	Felix	Hilversum	Netherlands
Vrije Universiteit	Hans	Burg	Amsterdam	Netherlands

Into 98

In 1998 the wordnets will be extended to cover about 30,000 concepts or synsets, correlating with 50,000 word meanings or synonyms. Furthermore, we will include wordnets for German, French, Czech and Estonian. The latter wordnets have been built as monolingual resources and will be linked to the ILI and compared with the other wordnets. All the databases will become available by the end of 1998 or early 1999. Distribution will be handled by ELRA.

Furthermore, we have submitted 6 papers to the Computers and the Humanities Journal for a special issue on EuroWordNet. These paper will represent a good introduction to the project. This issue will be published in 1998.

Reports

All documents, reports and public data can be downloaded from the EuroWordNet Web-site: http://www.let.uva.nl/~ewn:

Deliverables:

D001 User-requirements and functional specification
D005 Definition of the links and subsets for noun
D006 Definition of the links and subsets for verb
D007 High-level Architecture of the EuroWordNet Database
D010D011 Coding Relations for the First Subset of Nouns and Verb
D017D034D036 The EuroWordNet Base Concepts and Top-Ontology

Published EuroWordNet papers

Vossen, P., 1996 Right or Wrong. Combining lexical resources in the EuroWordNet project. Published in: M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, C.R. Papmehl, Proceedings of Euralex-96, Goetheborg, 1996, 715-728
Gilarranz, Julio, Julio Gonzalo, Felisa Verdejo, 1996 Using the EuroWordNet Multilingual Semantic Database. Published in: Proceedings of AAAI-96 Spring Symposium Cross-Language Text and Speech retrieval.
Vossen, P. 1997 EuroWordNet: a multilingual database for information retrieval, In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.
P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.
Vossen, P., P. Diez-Orzas, W. Peters, 1997 The Multilingual Design of EuroWordNet. Published in: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.
Vossen, P., W. Peters, P. Díez-Orzas, 1997 The Multilingual design of the EuroWordNet Database, In: Kavi Mahesh (ed.) Ontologies and multilingual NLP, Proceedings of workshop at IJCAI-97, Nagoya, Japan, August 23-29.