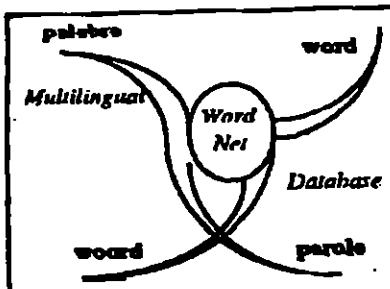


EuroWordNet

Building a Multilingual Wordnet Database with Semantic Relations between Words

LE-2 4003

Project Synopsis

Application Area	Language Resources, Language Engineering
Start Date	March 1996
Duration	36 Months
Total Effort	149 Person Months

Consortium	**Organisation**	**Short Name**	**Role**	**Nat. Code**	**Task**
	• University of Amsterdam	AMS	C	NL	provider
	• Istituto Di Linguistica Computazionale Pisa	ILC	P	IT	provider
	• Fundacion Universidad Empresa	FUE	P	ES	provider
	• Novell Belgium NV	NOV	P	BE	user
	• University of Sheffield	SHE	P	GB	provider

Contact person

Dr Piek Vossen
Project Manager
Computer Centrum Letteren
University of Amsterdam
Spuistraat 134
1012 VB Amsterdam
The Netherlands

Tel: +31 20 525 4624
Fax: +31 20 525 4429
Email: Piek.Vossen@let.uva.nl

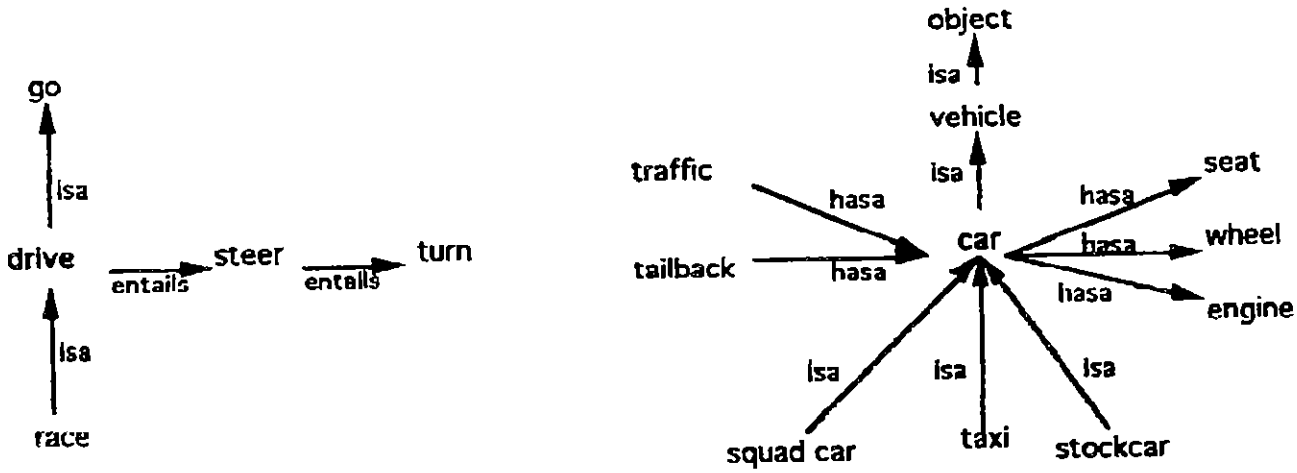
Abstract

The project aims at developing a multilingual database with basic semantic relations between words for several European languages (Dutch, Italian and Spanish). The wordnets will be linked to the American wordnet for English and a shared top-ontology will be derived, while language specific properties are maintained in the individual wordnets. The database can be used for multilingual information retrieval which will be demonstrated by Novell Linguistic Development.

1 Objectives

Currently, information is massively stored in electronic form and can be accessed from anywhere in the world via electronic networks. Although access to this information is constantly being improved by new interfaces and facilities, information retrieval from large electronic resources is still mainly determined by key word matching or fixed indexing and menu systems. Likewise, a user cannot simply use his own words to find information but has to make use of the wordings and rationale of the classification system. As the detail and amount of information increases a non-expert user will have more and more difficulty to use the right terminology to gain access to it. The situation in Europe is even worse since its diversity of languages and cultures constitutes an extra barrier, while the available linguistic tools to support textual search are mostly restricted to English. As a result of this, the information society is becoming restricted to a small group of people that speak English and have good knowledge of the access system and the stored data.

To provide non-expert searchers flexible access to the information society it is therefore crucial to develop tools that can expand his general and common words in a specific language to any possible variant or term in any other language. The user should be able to get around the choice of words in a document or the choice of key words by matching meanings rather than words. Such tools depend on the availability of generic resources with basic semantic relations between words, like the Princeton WordNet (Miller et al 1990). The American WordNet database consists of semantic relations between English word meanings (so-called synsets) which can be accessed as a kind of thesaurus in which words with related meanings are grouped together. For example, a noun like "car" is linked to, among others, all words that have a *hyponymy* or *isa* relation or a *meronymy* or *hasa* relation with it, and a verb like "drive" to, among others, all words that have a *hyponymy* or an *entailment* relation with it[1]:

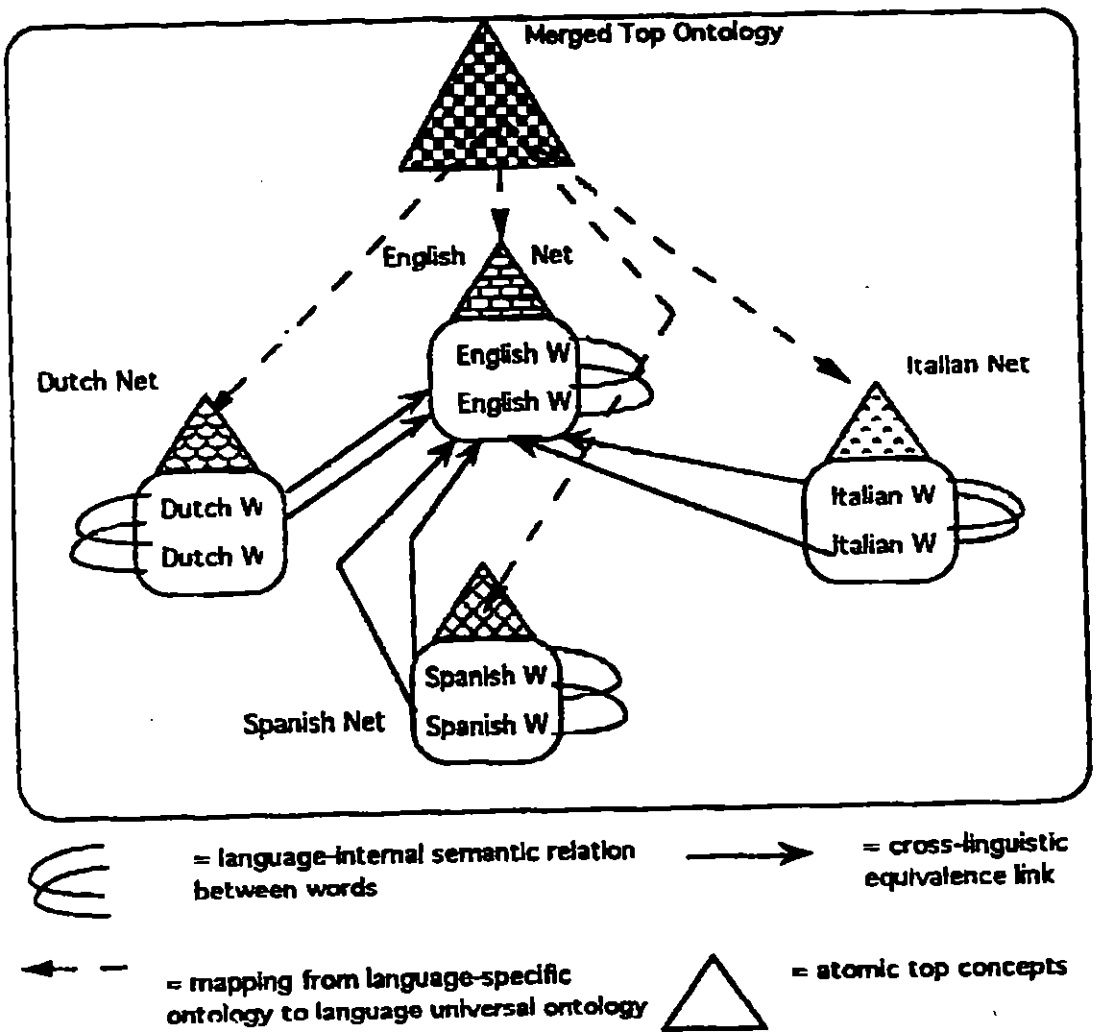


With such a database the query-terms of a user can be expanded to any set of closely related terms in a language, leading to better retrieval of information in terms of recall. For example a query with the terms "drive" and "car" will be expanded to combinations such as "go + car", "race + vehicle", "steer + car", "turn + wheel", "race + engine".

Unfortunately, such resources are not available for other languages than English, let alone a resource in which multiple wordnets are combined and interlinked. This severely holds back developments in language engineering and the information society in Europe. The aim of this project is therefore to develop such a *multilingual* database with wordnets for several European languages (Dutch, Italian and Spanish) which can be used to improve recall of queries via

1 Here a simplified example is given. In practice, different subtypes of "isa" and "hasa" relations will be distinguished as well as various other types of relations.

semantically linked variants in *any* of these languages. These European wordnets will as much as possible be built from available existing resources and databases with semantic information developed in various projects. This will not only be more cost-effective but will also make it possible to combine information from independently created resources, making the ultimate database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages. The wordnets will be stored in a central lexical database system and the word meanings will be linked to meanings in the Princeton WordNet1.5. Furthermore, we will merge the major concepts and words in the individual wordnets to form a common language-independent ontology (an ontology is the set of semantic relations between concepts). This will guarantee compatibility and maximise the control over the data across the different wordnets while language-dependent differences can be maintained in the individual wordnets.



What this implies is best illustrated with an example. Consider, for example, all the words that are related to *body parts*. All the wordnets will share the top-ontology concept BODY but each language has different lexicalizations for body parts. Whereas English words like "head" and "leg" can name the same parts of "animals" and "humans", in Dutch different words are used for animal parts and humans parts ("kop" (head) and "poot" (leg) for all animals except horses and "hoofd" (head) and "been" (leg) for humans and horses respectively). Similarly, in English and Dutch there are different words for "finger" and "toe" whereas Italian and Spanish have a single word to name both types of body parts ("dito" and "dedo" respectively). Each wordnet will thus reflect a unique lexicalization pattern. Equivalence in the lexicalization will be reflected by parallelism in the wordnet structure and simple equivalence relations with the English words, whereas differences in lexicalization will be reflected by divergence of the wordnet structures and

partial equivalence relations with the closest English word (different types of partial-equivalence relations will be distinguished).

Builders of the wordnets are: the University of Amsterdam (Co-ordinator of the project), the University of Sheffield, the University of Pisa, and a research team belonging to three Spanish Universities: University of Barcelona, Technical University of Catalunya (UPC) and U.N.E.D. The resulting data will be stored in a multilingual database system developed by Novell Linguistic Development in Antwerp in which relations can be traversed, selections can be made and data can be exported. Special facilities are made to get a multinlingual view on the wordnets.

2 User Involvement

The multilingual database will be tested and demonstrated by Novell who will also act as a user in the project and who is interested in developing such a multilingual information retrieval tool for the European Community. After full delivery of the resource Novell will load it into their Information Retrieval System (IRS) and test the adequacy in the demonstration phase. The Novell retrieval system does not use key words. Instead the user can enhance the query with a ConceptNet to search for all possible related terms. At this point, an English ConceptNet has been developed using WordNet1.5 as the main data source.

The development and testing of an IRS as such goes beyond the scope of this project. The user-requirements of such a full system include aspects such as flexibility, on-line help, visual representation of the results, allowing for different retrieval techniques and criteria (such as thesaurus or fixed-indexing systems, automatic key-word browsing, information on authors, publishers, institutes, reviews, date, etc.). The aim of this project is not to develop such a system but to provide a generic basic resource that could be included in such a broader IRS. Given the current State of the Art in IRSs the availability of general, generic resources such as a wordnet is typically expected to help non-expert users when retrieval by indexing is problematic because:

1) the indexing system does not cover the desired aspect or facet that a user is looking for,
2) the words chosen by the user are not included in the indexing key-word list,
3) the user speaks another language.

Therefore, the usefulness of the resources will not be tested by end-users in a real-life environment but by the developers of the IRS (in this project Novell) in controlled test situations that reveal the quality and added value for an IRS.

In addition to the direct scope of the project, we will also form a European User-Group of wordnet-builders and users that cover a wider range of languages and applications. The members of the User-Group will have the possibility to give feed-back to early releases of the project results (including sample, databases, documentation, definition of standards and data formats) which will be taken into account in the incremental building of the resources. Furthermore, we hope to create a wider awareness of the project results and to pave the way for the extension of the resources to other languages, larger vocabularies and other types of applications. The User-Group will contribute to a more complete understanding and description of the different user-needs depending on the kind of resources developed in this project (or developed on the basis of the project results).

The User-Group currently comprises:

Publishers	**Application area**
• Van Dale Lexicografie B.V. (NL) (provider of data)	(electronic) dictionaries, language generation tools, learning tools.
• Bibliograf (ES) (provider of data)	(electronic) dictionaries.
• Garzanti (IT)	(electronic) dictionaries.
Software Developers	
• SENA Athens (GR)	information retrieval
• CapVolmac, Utrecht (NL)	authoring tools, Grammar checkers
• INCYTA Barcelona (ES)	machine Translation
• Novell Linguistic Development, Antwerp (BE)	information retrieval, authoring tools, natural language interfaces
• LOGOS (IT)	technical translations, desktop publishing, technical writing.
• EBSCO (ES)	products for automated library systems, publishing of reference databases, retrieval systems for citation and full text databases, document delivery
• BERTIN (FR)	information retrieval in textual databases, concept-based indexing
• DATAMAT (IT)	information retrieval, document processing
Non-profit users	
• RKD, National Institute for Art-Historical Documentation (NL)	information retrieval, electronic libraries
• University of Madrid (ES)	machine translation, corpus linguistics, electronic dictionaries.
• VPRO, Broadcasting Organization (NL)	information retrieval, Internet services
Builders	
• University of Heidelberg (DE)	German wordnet
• University of Tuebingen (DE)	German wordnet
• University of Athens (GR)	Greek wordnet
• University of Goetheborg (ES)	Swedish wordnet
• University of Euskal Herriko	Bask wordnet
• University of Tartu, Estonia	Estonian, Latvian and Lithuanian wordnet
• University of Nantes, (FR)	French wordnet

During the project the user-group will be extended to achieve a maximal coverage in the different interested parties in Europe, where coverage relates to spreading in national interest, organisation type and type of application. Via exhibitions, the distribution of documents, electronic mailings and workshops we want to create an awareness of the electronic-linguistic services that can be developed using the multilingual wordnets as a starting point.

3 Results and exploitation

The most important deliverables will be a user-guide on the tools to develop the resources, the wordnets in each separate language (Dutch, Italian and Spanish) linked to the English WordNet, the shared top-ontology, the database in which all this can be viewed and selections can be exported and a report on the demonstration of the results in information retrieval tasks.

On a longer term the wordnets will become the backbone of any semantic database of the future and will open up a whole range of new applications and services in Europe at a trans-national and trans-cultural level. It will enhance the fundamental understanding of lexicalisation patterns across languages which will be crucial for machine translation and language learning systems. It will give non-native users and non-skilled writers the possibility to navigate or browse through the vocabulary of a language in new ways, giving them an overview of expression which is not feasible in traditional alphabetically organised resources. Finally, it will stimulate the development of sophisticated lexical knowledge bases which are crucial for a whole gamut of future applications, ranging from basic information retrieval to question/answering systems, language understanding and expert systems, summarizers to automatic translation tools and resources.

The results of the project will be publicly available where licensing contracts for background and foreground material will be drawn up at an early stage of the project. Non-commercial use will be free, commercial use will be charged for the background costs. The results will be stored on a CD and will be announced and distributed via commercial channels and via the academic networks. Information on the project can be obtained from a WWW home-page (see the contact person above), such as:

- general information on the project, such as progress, partnership, goals and aims, public documents, sponsorship
- forms to become registered as a member of the User Group.
- licensing forms for obtaining the project results
- public data samples, tools and databases

4 Work Parts and Time Schedule

The project can be characterized both as an LE-resource project and as a longer-term 'leading-edge' application project. The main focus will therefore be on Stage II (Development and Verification) of a project life-cycle, with minimal work parts for Stage I (Preparatory Activities) and Stage III (Demonstration). The main body of work will involve the building of the wordnets and the verification and demonstration in an information retrieval setting. The work packages (WPs) are organized around the 5 stages of a life-cycle:

Phase 1& 2:	User requirements and functional specification	WP1
Phase 3:	Building of the demonstrator	WP2, WP3, WP4, WP5, WP6
Phase 4:	Validation	WP7
Phase 5:	Exploitation	WP8

Separate work packages are devoted to managament (WP0), awareness and dissemination (WP9) and concertation (WP10).