EuroWordNet Tools and Resources Report

Piek Vossen, , Laura Bloksma, Paul Boersma, University of Amsterdam
Felisa Verdejo, Julio Gonzalo, UNED, Madrid
Horacio Rodriquez, German Rigau, Politecnica de Catalunya, Barcelona
Nicoletta Calzolari, Carol Peters, Eugenio Picchi, Simonetta Montemagni , ILC, Pisa
Wim Peters, University of Sheffield

Version 1, 15 July 1998
Final

Deliverable D021D025, WP2,
EuroWordNet, LE2-4003

Identification number	LE-4003-D-021D025
Type	Document
Title	EuroWordNet Tools Report
Status	
Deliverable	D021D025
Work Package	WP2
Task	T2.2
Period covered	
Date	15 July 1998
Version	1
Number of pages	9
Authors	• Piek Vossen,Laura Bloksma, Paul Boersma, University of Amsterdam • Felisa Verdejo, Julio Gonzalo, UNED, Madrid • Horacio Rodriquez, German Rigau, Politecnica de Catalunya, Barcelona • Nicoletta Calzolari, Carol Peters, Eugenio Picchi, Simonetta Montemagni , ILC, Pisa • Wim Peters, University of Sheffield
WP/Task responsible	AMS
Project contact point	Piek Vossen Computational Linguistics Department University of Amsterdam Spuistraat 134 1012 VB Amsterdam The Netherlands tel. +31 20 525 4669 fax. +31 20 525 4429 e-mail: Piek.Vossen@hum.uva.nl
EC project officer	Ray Hudson
Status	Public

Actual distribution	Project Consortium
Suplementary notes	n.a.
Key words	
Abstract	
Status of the abstract	
Received on	
Recipient's catalogue number	

Executive Summary

This deliverable gives an overview of the Tools and Resources used by each site to build their local wordnets. All sites have different resources and database and also the languages have different properties giving each one a very different starting point. Consequently, the partners apply different methodologies for building their wordnet.

Attached to the general introduction are 4 documents describing the tools and resources for each site. These documents can be seen as catalogues of tools and resources that are available and/or have been used. Other builders of wordnets can use the documents to find out what tools and methods fit their situation best and how they can be obtained.

1. Introduction

This deliverable is an introduction to the tools and resources that have been used by the sites in EuroWordNet-1 (LE2-4003) for building wordnet fragments for their national languages. In this document, we summarize the general approach for building the wordnets with local tools. Attached to this document are 4 other documents describing the tools and resources for each site. These documents can be seen as catalogues of tools and resources that are available and/or have been used. Other builders of wordnets can use the documents to find out what tools and methods fit their situation best and how they can be obtained.

The general EuroWordNet database, Polaris, is described in D024 (Louw 1998). In this database all the results are stored and compared. This document describes the methodologies and tools needed to get the data in the EuroWordNet format that Polaris can handle. Samples of the final database format for Dutch, Italian and Spanish, together with a complete conversion of WordNet1.5 to the EuroWordNet format can be downloaded from the EuroWordNet WWW-side: http://www.hum.let.uva.nl. The sample data are available in ASCII format and in EuroWordnet database format. The latter can be viewed with the Periscope program (Cuypers and Adriaens 1997, D008), which can be downloaded from the WWW-side for free as well.

Each site is responsible for the description and distribution of their own tools. The following contact persons should be addressed for further inquiries:

Dutch WordNet and Tools at AMS	Dr. Vossen, Piek Faculteit der Geesteswetenschappen Universiteit van Amsterdam Spuistraat 134 Amsterdam, 1012 VT, NL	tel: +31 20 525 4669 fax: +31 20 525 4429 e-mail: piek.vossen@let.uva.nl
Italian Wordnet and Tools, at PSA (ILC)	Prof. Calzolari, Nicoletta Instituto di Linguistica Computazionale C.N.R. Via Della Faggiola 32 Pisa, 56100, IT	tel: +39 50 560 481 fax: +49 50 55 62 85 e-mail: glottolo@ilc.pi.cnr.it
English Wordnet additions and Tools, at SHE	Prof. Wilks, Yorick Wim Peters University of Sheffield Computer Science Department Portobello Street 211, Regent Court Sheffield, S1 4DT, GB	tel: +44 1142 825 571 fax: +44 1142 780 972 e-mail: yorick@dcs.sheffield.ac.uk e-mail: w.peters@dcs.sheffield.ac.uk
Spanish Wordnet and Tools, at FUE	Verdejo, Felisa Fundación Universidad Empresa Serrano Jover 5,7[a] Madrid, 28015, ES	tel: +34 1 398 6484 fax: +34 1 398 6028 e-mail: felisa@iluvatar.ieec.uned.es
EuroWordNet Database Polaris, LHS	Dr. Adriaens, Geert Lernout and Hauspie Postleihof 1 Antwerp, BE	tel: +32 3 230 51 03 Adriaens, Geert e-mail: Geert.Adriaens@lhs.be

2. General approach for building the wordnets

The EuroWordNet database is being built (as much as possible) from available existing resources and databases with semantic information developed in various projects. In general, the wordnets are built in two major cycles as indicated by I and II in Figure 1 below. Each cycle consists of a building phase and a comparison phase:

1. Building a wordnet fragment
1.1. Specification of an initial vocabulary
1.2. Encoding of the language-internal relations
1.3. Encoding of the equivalence relations
2. Comparing the wordnet fragments
2.1. Loading of the wordnets in the EuroWordNet database (Polaris)
2.2. Comparing and restructuring the fragments
2.3. Measuring the overlap across the fragments

The building of a fragment is done using local tools and databases which are tailored to the specific nature and possibilities of the available resources. The available resources differ considerably in quality and explicitness of the data. Whereas some sites have the availability of partially structured networks between word senses, others start from genus words extracted from definitions that still have to be disambiguated in meaning.

After the specification of a fragment of the vocabulary, where each site uses similar criteria (there may again be differences due to the different starting points), globally, two approaches are followed for encoding the semantic relations:

> **Merge model**: the selection is done in a local resource and the synsets and their language-internal relations are first developed separately, after which the equivalence relations are generated to WordNet1.5 . This approach is followed for the Dutch and Italian wordnets.
>
> **Expand model**: the selection is done in WordNet1.5 and the WordNet.1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets in the other language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets. This approach is followed for the Spanish wordnet.

The Merge model results in a wordnet which is independent of WordNet1.5, possibly maintaining the language-specific properties. The Expand model will result in a wordnet which is very close to WordNet1.5 but which will also be biased by it. Whatever approach is followed also depends on the quality of the available resources.

After a production phase (step Ib and Ic in Figure 1) the results are converted to the EuroWordNet import format and loaded into the common database (step Ic). At that point various consistency checks are carried out, both formally and conceptually. By using the specific options in the EuroWordNet database it is then possible to further inspect and compare the data, to restructure relations where necessary and to measure the overlap in the fragments developed at

the separate sites. Those meanings not covered by a site may be included in the extension of the vocabulary in the next building phase.

The overall design of the EuroWordNet database makes it possible to develop the individual language-specific wordnets relatively independently while guaranteeing a minimal level of compatibility. Nevertheless, some specific measures have been taken to enlarge the compatibility of the different resources:

1. The definition of a common set of so-called Base Concepts that is used as a starting point by all the sites to develop the cores of the wordnets. Base Concepts[1] are meanings that play a major role in the wordnets: i.e. have many relations or high positions in the hierarchies.
2. The classification of the Base Concepts in terms of a Top Ontology.
3. The exchange of problems and possible solutions for encoding the relations for the Base Concepts.

The Base Concepts and the Top Ontology are further described in Deliverable D017D034D036 [Vossen et al. 1997] and in [Rodriguez et al. fc.]. The results of the first building phase are described in Deliverable D014D015 [Vossen et al. 1998].

[1] The notion of Base Concepts should not be confused with Basic-Level Concepts as defined by Rosch (1977). Base Concepts are technically defined as the concepts with most relations. In most cases, they are more general than the Basic Level Concepts.

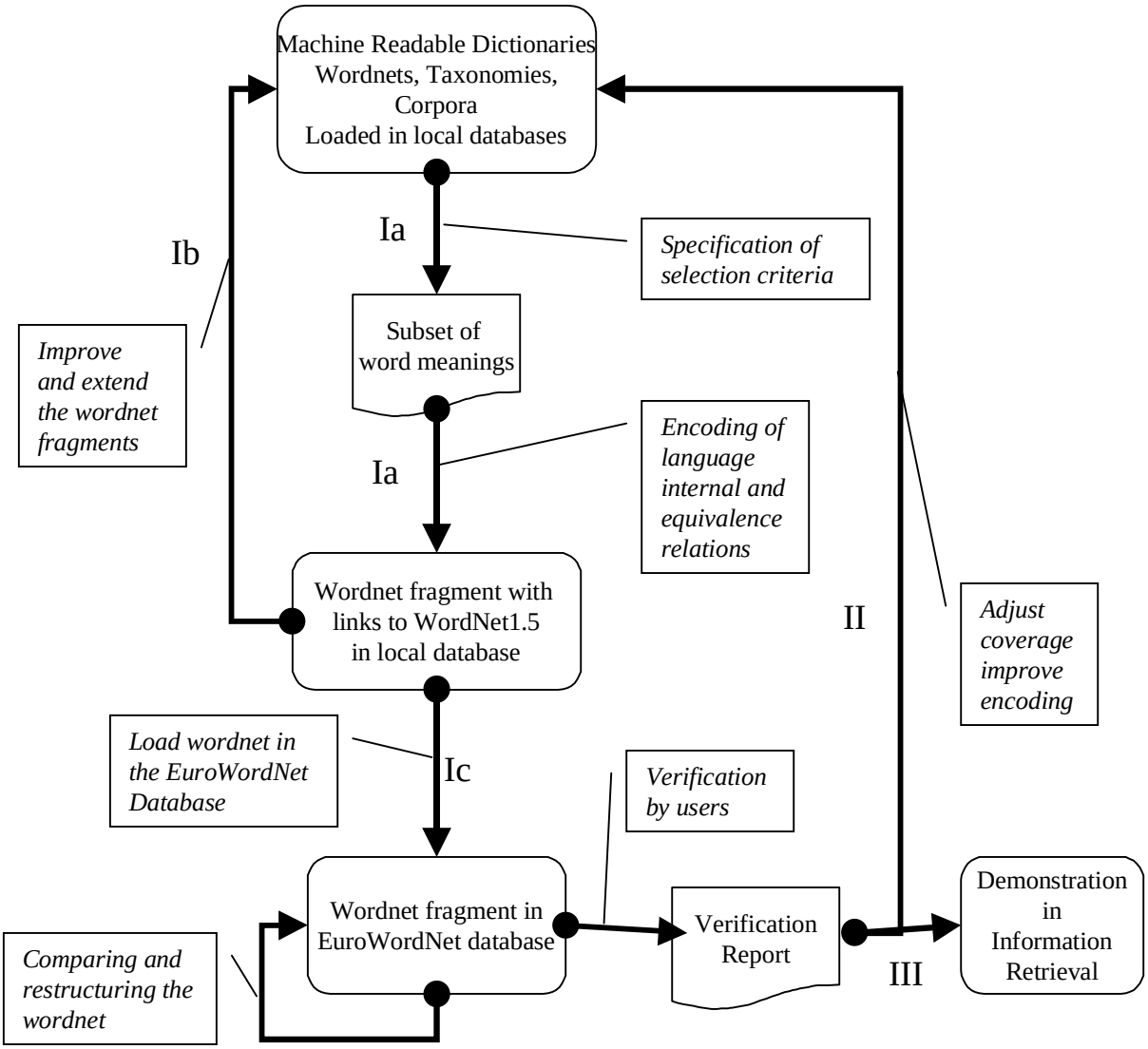

Figure 1: Overview of the building phases in EuroWordNet

References

Cuypers, I. And G. Adriaens
1997 *Periscope: the EWN Viewer*, EuroWordNet (LE2-4003), Deliverable D008d012. University of Amsterdam, Amsterdam.

Louw, M.
1998 *Polaris User's Guide: the EuroWordNet Database Editor*, EuroWordNet (LE2-4003), Deliverable D024, University of Amsterdam, Amsterdam.

Rodriguez, H., S. Climent, P. Vossen, L. Bloksma; A. Roventini, F. Bertagna, A. Alonge, W. Peters,
(fc) The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In: Computer and the Humanities, Special Issue on EuroWordNet.

Rosch, E.
1977 *Human Categorisation.* In N. Warren (Ed.) Studies in Cross-Cultural Psychology, Vol. I, pp. 1-49. Academic Press. London. 1977

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Wim Peters
1997 The EuroWordNet Base Concepts and Top Ontology, EuroWordNet (LE2-4003) D017D034D036, University of Amsterdam, Amsterdam.

Vossen, Piek, Laura Bloksma, Salvador Climent, Maria Antonia Marti, Gabriel Oreggioni, Gerard Escudero, German Rigau, Horacio Rodriguez, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Carol Peters, Wim Peters
1998 The Restructured Core wordnets in EuroWordNet: Subset1, EuroWordNet (LE2-4003) D014D015, University of Amsterdam, Amsterdam.