Final Wordnets for Czech, Estonian, French and German

Version 1, Final

September 26, 1999

Contributors:

Piek Vossen, University of Amsterdam
Claudia Kunze, Andreas Wagner, Universität Tübingen
Karel Pala, Pavel Sevecek, Masaryk University Brno
Kadri Vider, Leho Paldre, University of Tartu
Laurent Catherin, Université d'Avignon
Dominique Dutoit, Memodata

Deliverable 2D014, WP3, WP4
EuroWordNet, LE4-8328

Identification number	LE4-8328-2D014
Type	Document and Lingware
Title	Final wordnets for Czech, Estonian, French and German
Status	Final
Deliverable	2D014
Work Package	WP3 and WP4
Task	T3.6,T4.6
Period covered	April 1998 – October 1998
Date	September 26, 1999
Version	1
Number of pages	
Authors	♉ Piek Vossen, University of Amsterdam ♉ Claudia Kunze, Andreas Wagner, Universität Tübingen ♉ Laurent Catherin, Université d'Avignon ♉ Dominique Dutoit, Memodata ♉ Kadri Vider, Leho Paldre, University of Tartu ♉ Karel Pala, Pavel Sevecek, Masaryk University Brno
WP/Task responsible	TUE
Project contact point	Piek Vossen University of Amsterdam Spuistraat 134 1012 VB Amsterdam The Netherlands tel. +31 20 525 4669 fax. +31 20 525 4429 e-mail: **Fout! Verwijzingsbron niet gevonden.**
EC project officer	Ray Hudson
Status	Public
Actual distribution	Project Consortium, the EuroWordNet User Group, the world via **Fout! Verwijzingsbron niet gevonden.**.
Suplementary notes	n.a.
Key words	Linguistic Resources, Multilingual Wordnets, Language Engineering

Abstract	This deliverable describes the final wordnets for Czech, Estonian, French and German.
Status of the abstract	Final
Received on	
Recipient's catalogue number	

Executive Summary

This deliverable describes the final wordnets for Czech, Estonian, French and German. The complete wordnets consist of 15K to 30K word senses per language, which more or less corresponds with 10K to 20K synsets. This confirms to the specifications of the Technical Annex.

All synsets have at least a hyperonym relation and an equivalence link to the Inter-Lingual-Index that connects all the wordnets. Additional relations have been expressed up to 2.4 relations average per synset.

This document first gives some general background and overview tables and next consists of 4 parts describing the separate wordnets.

Table of Contents

1. General approach for building the wordnets

The EuroWordNet database was 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, buidling of subset 1 and subset 2. 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
   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.

The first wordnet subsets of EuroWordNet 2 have been created from a set of 1310 Base Concepts (extended from the set of 1024 Base Concepts selected by EWN 1 languages). These Base Concepts play an important role in at least two wordnets, where importance is measured in terms of numbers of relations and position in the hierarchy. The Base Concepts have been represented (as far as possible) by synsets in Czech, Estonian, French and German, and have been extended with other synsets that are important in these languages. These sets have been encoded and extended to form the first subset (minimally 7,500 resp. 3,500 synsets due to the different obligations). These first subsets represent the cores of the different wordnets on which the meanings of more specific concepts depend. 2D007008 contains a detailed description of the first subset and the results of comparing them. The core wordnets have been extended to full coverage, taking the results of comparison into account. The process of extending the wordnets is described in the individual wordnet reports attached to this deliverable 2D014. In 2D011D012014, AMS/FUE and AVI describe the results of comparing the final wordnets. In this deliverable, we describe the overall building process and results of the individual wordnets:

2D001	Set of Common Base Concepts in EuroWordNet-2
2D007008	Building and restructuring of core wordnets for EuroWordNet-2 languages
D011012[1]	Extending the core wordnets
2D011012014	Comparing the final wordnets
2D014	Final wordnet reports for Czech, Estonian, French and German

[1] The extension of the core wordnets of EWN-2 languages (2D011012) has not been documented separately in a deliverable. The outcome of subset 2 is being integrated in the comparison part of this deliverable (2D014) initial to the final wordnet reports for Czech, Estonian, French and German.

The total set of synsets aimed at for the final wordnets is 15-30K word meanings, which more or less corresponds with 10-20K[2] synsets, due to different building conditions of the EuroWordNet-2 languages. For each of these synsets, the following information has to be minimally specified:

〈 Hyperonym
〈 Synonyms (synset members)
〈 Equivalence relations to the Inter-Lingual-Index (WordNet1.5)

Optionally, any other relation has been added. The addition of other relations first of all depends on the relevance of the relation for the synset. Secondly, we have been limited by the project resources. Given the project-funding, it is not possible to comprehensively encode all relevant relations.

Attached to this document there are 4 reports describing the Czech, Estonian, French and German wordnet. Here we will give some overview tables of the results with some comments.

[2] The number of correlating synsets depends on the definition of synonymy that is used within a wordnet. Since these synonyms may be extracted automatically the ratios may differ across the wordnets.

2. Quantitative overview of the Czech, Estonian, French and German wordnet

The aimed size of the wordnets is 15-30K word meanings, which roughly corresponds to 10-20K synsets (where synsets can contain multiple synonyms, but different meanings of a word cannot occur in the same synset). For each synset, at least one hyperonym relation and one equivalence relation is required, other relations are optional. The next table gives an overview of the results for all 4 wordnets:

Table 1: Overview figures for the Czech, Estonian, French and German wordnets

		Synsets	*No. of senses*	*Sens./ syns.*	*Entries*	*Sens./ entry*	*LIRels.*	*LIRels/syns*	*EQRels -ILI*	*EQRels /syn*	*Synsets without ILI*	*% without ILI*
Czech Wordnet	**Nouns**	9727	13829	1.42	9277	1.49	19856	2.04	9729	1.00	0	0.00%
	Verbs	3097	6120	1.98	3006	2.04	6403	2.07	3097	1.00	0	0.00%
	Other	0	0	0	0	0	0	0	0	0	0	0.00%
	Total	12824	19949	1.56	12283	1.62	26259	2.05	12824	1.00	0	0.00%
Estonian Wordnet	**Nouns**	5028	8226	1.64	7209	1.14	10873	2.16	5683	1.13	0	0.00%
	Verbs	2650	5613	2.12	3752	1.50	5445	2.05	3321	1.25	0	0.00%
	Other	0	0	0	0	0	0	0	0	0	0	0.00%
	Total	7678	13839	1.80	10961	1.26	16318	2.13	9004	1.17	0	0.00%
French Wordnet	**Nouns**	17826	24499	1.37	14879	1.65	39172	2.20	17815	1.00	16	0.09%
	Verbs	4919	8310	1.69	3898	2.13	10322	2.10	4915	1.00	4	0.08%
	Other	0	0	0	0	0	0	0	0	0	0	0.00%
	Total	22745	32809	1.44	18777	1.75	49494	2.18	22730	1.00	20	0.09%
German Wordnet	**Nouns**	9951	13656	1.37	12746	1.07	23856	2.40	10570	1.06	0	0.00%
	Verbs	5166	6778	1.31	4333	1.56	10960	2.12	5762	1.12	0	0.00%
	Other	15	19	1.27	19	1	2	0.13	15	1.00	0	0.00%
	Total	15132	20453	1.35	17098	1.20	34818	2.30	16347	1.08	0	0.00%

With respect to the numbers of synsets, we can state that every language covers (at least) the aimed size which is 15 K for German and French and 7,5 K for Estonian and Czech. French (almost 23K) and Czech (almost 13K) cover 50% resp. 70% more than the required size. In terms of senses, French (32K) and Czech (20K) have more than the aimed size (30K resp. 15K). The ratio of synonyms per synset is quite balanced across the wordnets, the highest for Estonian (1.8) and the lowest for German (1.35), whereas WordNet15 has a ratio of 1.84.
The degree of language internal relations is also rather balanced across the wordnets: averages of 2.05 (Czech), 2.13 (Estonian), 2.18 (French) and 2.30 (German) relations per synset. In terms of the equivalence relations, we see that the Czech and French wordnet have a 1:1 ratio with the Inter-Lingual-Index and that Estonian and German have an average a bit more than 1 equivalence per synset.

The next column gives the distribution of these language internal relations. The first column gives the absolute number of relations per type, the second column for each language gives the relative percentage.

Table 2: Overview of Language Internal Relations

Language Internal Relations	*Czech*		*Estonisn*		*French*		*German*		*Wordnet15*	
HAS_HYPERONYM	12824	48,3%	7804	47,8%	22741	45,9%	15778	45,3%	71902	33,9%
HAS_HYPONYM	12824	48,3%	7804	47,8%	22741	45,9&	15778	45,3%	71902	33,9%
HAS_XPOS_HYPERONYM	0	0,0%	6	0,0%	0	0,0%	0	0,0%		
HAS_XPOS_HYPONYM	0	0,0%	6	0,0%	0	0,0%	0	0,0%		
NEAR_SYNONYM	0	0,0%	78	0,5%	0	0,0%	0	0,0%	20014	9,4%
XPOS_NEAR_SYNONYM	0	0,0%	116	0,7%	0	0,0%	0	0,0%		
NEAR_ANTONYM	213	0,8%	42	0,3%	754	1,5%	668	1,9%	10070	4,7%
XPOS_NEAR_ANTONYM	0	0,0%	4	0,0%	0	0,0%	0	0,0%		
HAS_HOLONYM	0	0,0%	16	0,1%	50	0,1%	1228	3,5%		
HAS_HOLO_LOCATION	0	0,0%	3	0,0%	0	0,0%	0	0,0%		
HAS_HOLO_MADEOF	0	0,0%	4	0,0%	51	0,1%	0	0,0%		
HAS_HOLO_MEMBER	24	0,1%	8	0,0%	131	0,3%	0	0,0%	11471	5,4%
HAS_HOLO_PART	84	0,3%	43	0,3%	1067	2,2%	0	0,0%	5690	2,7%
HAS_HOLO_PORTION	0	0,0%	1	0,0%	1	0,0%	0	0,0%	366	0,2%
HAS_MERONYM	0	0,0%	16	0,1%	50	0,1%	1228	3,5%		
HAS_MERO_LOCATION	0	0,0%	3	0,0%	0	0,0%	0	0,0%		
HAS_MERO_MADEOF	0	0,0%	4	0,0%	51	0,1%	0	0,0%		
HAS_MERO_MEMBER	24	0,1%	8	0,0%	131	0,3%	0	0,0%	11471	5,4%
HAS_MERO_PART	84	0,3%	43	0,3%	1067	2,2%	0	0,0%	5690	2,7%
HAS_MERO_PORTION	0	0,0%	1	0,0%	1	0,0%	0	0,0%	366	0,2%
INVOLVED	0	0,0%	60	0,4%	2	0,0%	0	0,0%		
INVOLVED_AGENT	66	0,2%	7	0,0%	4	0,0%	0	0,0%		
INVOLVED_PATIENT	1	0,0%	4	0,0%	0	0,0%	0	0,0%		
INVOLVED_INSTRUMENT	4	0,0%	21	0,1%	10	0,0%	0	0,0%		
INVOLVED_LOCATION	3	0,0%	1	0,0%	1	0,0%	0	0,0%		
INVOLVED_RESULT	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
INVOLVED_DIRECTION	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
INVOLVED_SOURCE_DIRECT	0	0,0%	0	0,0%	10	0,0%	0	0,0%		
INVOLVED_TARGET_DIRECT	0	0,0%	4	0,0%	0	0,0%	0	0,0%		
ROLE	0	0,0%	55	0,3%	2	0,0%	0	0,0%		
ROLE_AGENT	66	0,2%	7	0,0%	4	0,0%	0	0,0%		
ROLE_PATIENT	1	0,0%	4	0,0%	0	0,0%	0	0,0%		
ROLE_INSTRUMENT	4	0,0%	21	0,1%	10	0,0%	0	0,0%		
ROLE_LOCATION	3	0,0%	1	0,0%	1	0,0%	0	0,0%		
ROLE_RESULT	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
ROLE_DIRECTION	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
ROLE_SOURCE_DIRECTION	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
ROLE_TARGET_DIRECTION	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
CO_ROLE										
CO_AGENT_PATIENT										
CO_AGENT_INSTRUMENT										
CO_AGENT_RESULT										
CO_PATIENT_AGENT										
CO_PATIENT_INSTRUMENT										
CO_PATIENT_RESULT										
CO_INSTRUMENT_AGENT										
CO_INSTRUMENT_PATIENT										
CO_INSTRUMENT_RESULT										
CO_RESULT_AGENT										
CO_RESULT_PATIENT										
CO_RESULT_INSTRUMENT										
CAUSES	0	0,0%	60	0,4%	311	0,6%	60	0,2%	204	0,1%
IS_CAUSED_BY	0	0,0%	60	0,4%	311	0,6%	60	0,2%	204	0,1%
HAS_SUBEVENT	17	0,1%	16	0,1%	1	0,0%	5	0,0%	435	0,2%
IS_SUBEVENT_OF	17	0,1%	16	0,1%	1	0,0%	5	0,0%	435	0,2%
IS_MANNER_OF	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
IN_MANNER	0	0,0%	0	0,0%	0	0,0%	0	0,0%		
BE_IN_STATE	0	0,0%	2	0,0%	0	0,0%	0	0,0%	636	0,3%
STATE_OF	0	0,0%	2	0,0%	0	0,0%	0	0,0%	1272	0,6%
ANTONYM	0	0,0%	0	0,0%	0	0,0%	0	0,0%		

HAS_DERIVED	0	0,0%	0	0,0%	0	0,0%	2	0,0%		
IS_DERIVED_FROM	0	0,0%	0	0,0%	0	0,0%	2	0,0%		
FUZZYNYM	0	0,0%	0	0,0%	0	0,0%	4	0,0%		
XPOS_FUZZYNYM	0	0,0%	2	0,0%	0	0,0%	0	0,0%		
Total	26259		16318		49494		34818		212128	

In all 4 languages, hyponymy is the most important relation, French and German have about 90%, Czech and Estonian even 90% hyperonym-hyponymy-relations. The distribution of the other relations is also rather balanced. Some of the language internal relations have not been encoded in the wordnets at all, eg. the different ROLE-relations for German.
The CO -roles which have been added to the EuroWordNet specification very recently are not used by anyone of the EWN-2 partners.
The German wordnet exhibits the highest percentage of part-of relations[3], the Estonian wordnet is the only one to provide for the LI-relations NEAR_SYNONYM and XPOS_NEAR_SYNONYM, Czech has made a relative frequent use of the ROLE_AGENT relation (taking into account the distribution of Czech LI-relations and in comparison to the other wordnets).

Finally, Table 3 lists the equivalence relations for Czech, Estonian, French and German.

Table 3: Overview of Equivalence Relations

	Czech		***Estonian***		***French***		***German***	
Equivalence Relations	Nouns	Verbs	Nouns	Verbs	Nouns	Verbs	Nouns	Verbs
EQ_SYNONYM	9727	3097	3989	1682	17815	4915	8186	4206
EQ_NEAR_SYNONYM			551	1102			652	792
EQ_HAS_HYPERONYM			592	166			1153	541
EQ_HAS_HYPONYM			227	134			316	71
EQ_INVOLVED			69	87			20	10
EQ_ROLE			112	9			17	38
EQ_IS_CAUSED_BY			27	80			3	25
EQ_CAUSES			2	24			7	36
EQ_HAS_HOLONYM			6	0			125	0
EQ_HAS_MERONYM			47	0			52	0
EQ_HAS_SUBEVENT			0	0			3	5
EQ_IS_SUBEVENT_OF			0	3			0	2
EQ_BE_IN_STATE			58	34			36	36
EQ_IS_STATE_OF			0	0			0	0
EQ_CO_ROLE			0	0			0	0
EQ_GENERALIZATION								
EQ_METONYM								
EQ_DIATHESIS								
Total	9727	3097	5683	3321	17815	4915	10570	5762

The main conclusions we can make here are:

- exclusive mapping of EQ_SYNONYM in the French and Czech wordnets which have been built by translating WordNet 1.5 synsets so that a 1:1 ratio is achieved;
- quite frequent usage of EQ_NEAR_SYNONYM and EQ_ROLE and EQ_INVOLVED in the Estonian wordnet. Note that 33% of the Estonian verbs have an EQ_NEAR_SYNONYM link but only 14% of the German verbs. Only 50% of the Estonian verbs compared to 72% of the German verbs are linked via a 'pure' synonymy relation. This may partly correlate to the observation that in Estonian lexicalization patterns expressing causativity are very productive and often have no counterpart in WordNet 1.5 (see Estonian Wordnet report).
- quite frequent usage of EQ_HAS_HOLONYM in the German wordnet (eg. in case of establishing multiple EQ_links in addition to a EQ_HAS_HYPERONYM link).

[3] Note that the German wordnet does not distinguish among different meronymy relations.

References

Vossen, P., C. Kunze, A. Wagner, D. Dutoit, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Olav, H.Õim.
1998 Set of Common Base Concepts in EuroWordNet-2: Subset1. EuroWordNet (LE4-8328), Deliverable 2D001, University of Amsterdam.

Kunze, C., A. Wagner, D. Dutoit, L. Catherin, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Olav, H. Õim
1998 First Wordnets for Base Concepts in French, German, Czech and Estonian. EuroWordNet (LE4-8328), Deliverable 2D007, University of Amsterdam.

Vossen, P., G. Escudero, H. Rodriquez., C. Kunze, A. Wagner, D. Dutoit, L. Catherin, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Olav, H. Õim.
1998 Comparison of the Core Wordnets in EuroWordNet-2, EuroWordNet (LE4-8328), Deliverable 2D008, University of Amsterdam.

Vossen, P., G. Escudero, H. Rodriquez., L. Catherin, D. Dutoit, C. Kunze, A. Wagner, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Olav, H. Õim.
1999 Comparison of the Final Wordnets German, French, Czech and Estonian, EuroWordNet (LE4-8328), Deliverable 2D011012014, University of Amsterdam.

Part A: The Czech wordnet

Part B: The Estonian Wordnet

Part C: The French Wordnet

Part D: The German Wordnet