



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D20.1

Report: Multilingual Medical Dictionary

Report Version: 0.1

Report Preparation Date:

Classification: RE

Contract Start Date: 2004-01-01

Duration: 3 years

Project Co-ordinator: Hans Åhlfeldt

Department of Biomedical Engineering / Medical Informatics

S-581 83 Linköping University, Sweden

<hans.ahlfeldt@imt.liu.se>



**Project funded by the European Community under the FP6
programme "Integrating and Strengthening the European Research
Area" (2002-2006)**



Table of Contents

1	ADMINISTRATIVE INFORMATION	3
2	INTRODUCTION	3
3	OBJECTIVES OF THE WORKPACKAGE	3
4	QUALITY INDICATORS RELEVANT TO WP20	3
4.1	Q1 Workshops and symposiums	4
4.2	Q2 Sharing of resources and use of research software tools	4
4.3	Q6 Short-and medium-term visits of staff members	4
4.4	Q7 Co-authoring of research papers, reports and educational materials	4
5	SUMMARY OF ACTIVITIES IN 2004	5
5.1	Workshops and Symposiums	5
5.1.1	Workpackage Meetings	5
5.1.2	International Conference Participation	5
5.2	Co-authoring	6
5.3	Joint Research Programme	6
5.4	Exchange Visits	7
6	WORKPACKAGE ASSESSMENT	8
6.1	Assessment against Quality Indicators	8
6.2	Assessment Against Workpackage Objectives	8
6.2.1	Exchange of Methods and Resources	8
6.2.2	Elaboration of a Common Data Structure and Integration of Biomedical Terminologies	8
6.2.3	Population of Lexicons by Manual and Semi-automated Approaches	9
6.2.4	Integration with other Workpackages	9

1 Administrative Information

Lead contractor: UKLFR (Freiburg University Hospital)

Responsible: UNIFR, LiU (IMT), DIM, ITRI, UGOT, SU, INSERM

Participants: KI, LiU (IDA). CNR, KITH, SOS, STAKES, NBH

Deliverable: D20.1

Author: Stefan Schulz

2 Introduction

SemanticMining is a EU network of excellence which comprises 25 participant organisations from 11 European countries. This deliverable presents a progress report for work package 20 “Research Activity Multilingual Medical Dictionary”, covering the period from the start of the research activity in June 2004 until the end of November 2004, including the WP20 kick-off meeting end of March 2004.

The deliverable target audience is the Commission and other project participants. It is not intended for public distribution.

The work package description focuses on multilingual approaches to medical terminology, tackling several aspects such as terminology cross-mapping, cross-language text retrieval, lexicon acquisition and corpus-based machine translation. A central focus of this work package is the construction of a multi-lingual medical base lexicon providing a broad coverage of the main European languages.

3 Objectives of the Workpackage

The work package objectives are to enable greater knowledge transfer integration and collaborative terminology development between its participants. The following objectives are stated in the project proposal:

- To facilitate the exchange of methods and resources by short study visits of members of each others’ groups
- To propose a common data structure for medical language resources, linking lexicons of different scope, language and granularity, covering both morphosyntactic and semantic aspects
- To enhance, integrate and populate existing lexicons, focussing on atomic (non-decomposable) lexical entities, using manual, semi-manual and automated lexical acquisition approaches.

4 Quality Indicators Relevant to WP20

Annex I of the contract defines those quality indicators by which the consortium seeks to assess its progress. A subset of these indicators are listed below; these have been adopted in this report in order to structure the assessment of progress in work package 20.

4.1 Q1 Workshops and symposiums

Special Interest Group Workshops at Workpackage Level

Number: at least 3 per year; length 3 days; Location: variable sites

Subjects covered: each workshop should cover one or more work packages

Participation in national and international conferences

WP20 should be represented with several member-institutions presenting as a unit via workshops, information meetings, and the like in at least 3 national and 1 international conference per year

4.2 Q2 Sharing of resources and use of research software tools

Software Resources

Goal: At least two software tools developed at one location but used by different partners

Baseline: N/A

Terminological / Lexical Resources

Goal: At least two resources exchanged between partners

Baseline: N/A

4.3 Q6 Short-and medium-term visits of staff members

Medium-term visits = 2 weeks or longer, for preparation of joint publications, student supervision, conference organization etc.

short-term visits = less than 2 weeks, for conferences and workshops, meetings with students

Baseline: assumed 0

Goal: all institutions should participate at least one visit per year for WP20.

4.4 Q7 Co-authoring of research papers, reports and educational materials

Participants in all institutions at senior researcher, post doc and PhD student levels should be involved with participants from other institutions in the preparation of co-authored scientific papers (original research paper, reviews, conference proceedings etc.) and reports.

Baseline: assumed 0

Goal: all research institutions should have at least 1 and an average of 2 co-authored scientific papers per year

5 Summary of Activities in 2004

5.1 Workshops and Symposiums

5.1.1 Workpackage Meetings

The following two meetings with their scope restricted to WP 20 were organized and hosted within the network:

Title	Date	Location	Network Delegates	Comments
WP 20 Kick off meeting	29/30 March	Freiburg, Germany	17	Delegates from UKLFR, UNIFR, LiU, DIM, ITRI, UGOT, SU, INSERM
Lexicon Acquisition Workshop	17/18 May	Göteborg, Sweden	14	Delegates from UKLFR, UNIFR, LiU, DIM, ITRI, UGOT, SU, INSERM

Future Plans

The next WP20 internal meeting will be held at ITRI, probably in February.

A larger lexical acquisition workshop is planned for end of 2005.

5.1.2 International Conference Participation

WP 20 relevant participation in international conferences include:

Title	Date	Location	Network Delegates	Comments
EUROMISE 2004	12-15 April	Prague, CZ	5	International conference; Keynote on Medical Language and Knowledge Engineering by WP20 leader
RIAO 2004	26-28 April	Avignon, FR	3	International conference; Papers presented by network members
FLAIRS 2004	17-19 May	Miami, US	1	International conference; Papers presented by network members
LREC 2004	26-28 May	Lisbon, PT	5	International conference; Papers presented by network members
COLING 2004	23-27 August	Geneva, CH	5	International conference; Paper presented by network members
MEDINFO 2004	7-11 September	San Francisco, USA	6	International conference; Papers, posters, tutorials and panels presented by network members
CLEF 2004	5-17 September	Bath, UK	1	International workshop; paper presented by network member

5.2 Co-authoring

The following publications relevant to WP 20 were written by authors from at least two SemanticMining partners:

- Hahn U, Markó K, Schulz S: Learning Indexing Patterns from One Language for the Benefit of Others. AAAI 2004. The Nineteenth National Conference on Artificial Intelligence, July 25-29, 2004, San Jose, California.
- Poprat M, Hahn U, Wermter J, Markó K, Schulz S: An Experimental Assessment of Direct vs. Interlingual Translation for Cross-Language Information Retrieval. The 17th International FLAIRS Conference, Miami Beach, Florida, May 17-19, 2004.
- Schulz S, Markó K, Sbrissia E, Nohama P, Hahn U: Cognate Mapping – A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. COLING 2004. The 20th International Conference on Computational Linguistics. Geneva, August 23-27, 2004.
- Hahn U, Poprat M, Schulz S, Wermter J, Markó K: Crossing Languages in Text Retrieval via an Interlingua. Proceedings of RIAO'04 - 7th International Conference "Recherche d'Information Assistée par Ordinateur", Avignon, April 26-28, 2004.
- Markó K, Hahn U, Schulz S, Daumke P: Interlingual Indexing across Different Languages. Proceedings of RIAO'04 - 7th International Conference "Recherche d'Information Assistée par Ordinateur", Avignon, 26. – 28.4.2004: 82-99.
- Markó K, Schulz S, Wermter J, Poprat M, Hahn U: Cross-Language Document Retrieval with MORPHOSAURUS. gmds2004, 26.-30. Sept. 2004, Innsbruck, 2004: 184-186.

5.3 Joint Research Programme

As a result of the WP 20 kickoff meeting in March, three subtasks have been defined:

- a) specification of a common framework for multilingual medical dictionaries, coordinated by ITRI and DIM,
- b) inclusion of additional languages into the MorphoSaurus lexicon, coordinated by UKLFR
- c) lexical acquisition, coordinated by UGOT

For these subtasks the following activities took place until now:

- Elaboration of a draft specification of a common framework for multilingual medical dictionaries (DIM).
- Adaptation of the MorphoSaurus Lexicon Editor by UKLFR to additional languages (Spanish, French, Swedish).
- Update of the specification for MorphoSaurus Lexicon development and training of MorphoSaurus input providers by UKLFR.
- Population of the MorphoSaurus lexicon with French, Swedish and Spanish lexemes (UKLFR).
- Preparatory work for quality assessment for lexicon development (UKLFR).
- Improvement of interactive word alignment tools as regards precision in proposals and support tools. (LiU-IDA)

-
- Semi-automated population of the MorphoSaurus lexicon by generating candidate Swedish and Spanish subwords by character translation from German, English and Portuguese subwords, with sanity check in comparable corpora (SU, UKLFR, UNIFR)
 - Exchange of lists of French and Swedish suffixes (UGOT, DIM)
 - Work on named entity recognition and compounds (UGOT).
 - Collection of biomedical corpora (> 2 Gbyte) for German, English, Swedish, French, Spanish and Portuguese (UKLFR, UNIFR) from different sources (Web, eBooks, journals, medical narratives).
 - Updating of a Swedish general purpose lexicon with medical terms (UGOT).
 - Integration of medical terminology in a Swedish named entity recognition system (UGOT).
 - Ongoing collaboration with regional healthcare authorities on lexical resources for computer-assisted Swedish LSP instruction to non-Swedish healthcare staff (UGOT).
 - Discussions of possible involvement of Lexical Computing Ltd as a new partner in order to promote the use of the Sketch Engine by WP 20 participants (ITRI).
 - Initial design for a DATR-inspired Java-based generic interface for multilingual lexical resources (ITRI) 5.3.
 - Enlargement and quality improving work on a bilingual terminology collection of English and Swedish medical terminology systems. (LiU-IMT) .
 - Word alignment in English and Swedish medical terminology systems with the purpose of building an English-Swedish medical lexicon. (LiU-IMT & LiU-IDA).

5.4 Exchange Visits

The number and duration of exchange visits between network members is another indicator of network effort spent towards future joint research. The summer school and several conferences have provided opportunity for many informal bilateral and group discussions. Up to now there have been only short exchange visits:

- Regular meetings between the Göteborg groups UGOT the SU SemanticMining group
- Regular meetings between UKLFR and UNIFR/JENA groups
- Visit of Dietrich Rebholz-Schuhmann and Harald Kirsch, EBI (March 15)
- Visit of Kornél Markó (UKLFR) at Göteborg (May 18-19)
- Visit of Anders Thurin (SU) at Freiburg (Sept 13-17)
- Regular meetings between LiU-IMT and LiU-IDA groups.

6 Workpackage Assessment

6.1 Assessment against Quality Indicators

Considering that the scheduled WP20 activities for year one spanned only eight months, the target for specialist workshops organised was reached. In terms of participation in international conferences, WP20 has exceeded its targets.

Co-authoring activity in the first year was limited to the UKLFR and UNIFR (now JENA) groups which, however, have been closely worked together for many years. Regarding the other partners, time is required to become familiar with each other before opportunities for joint publication become apparent. This familiarisation process is expected to be successfully completed at the next WP meeting (scheduled for 02/2005), so joint publications will appear over the coming year.

Co-tutoring of doctoral students is so far being practised in three cases between UKLFR and UNIFR (now JENA).

The sharing of lexical resources and corpora has already produced good results in the ongoing process of lexicon population.

Although the communication between the different groups has developed well using e-mail and the common MERMIG platform, it should be improved by more frequent visits by staff members, as well as by doctoral students, encouraged by the mobility program which will start in 2005.

6.2 Assessment Against Workpackage Objectives

6.2.1 Exchange of Methods and Resources

So far, the principal method exchanged has been a lexical acquisition routine adapted to medical subwords. Based on character translation rules, translation candidates are generated and checked against a corpus in the target language. At present, the focus of research lies on the identification of false cognates and new (valid) translations, by exploiting context and token (subword) frequency. The principal resources exchanged were lists of domain-specific affixes. The large collection of corpora are currently being reorganized and documented in order to share them within the network through the common database (WP3).

6.2.2 Elaboration of a Common Data Structure and Integration of Biomedical Terminologies

A specification of a common structure was worked out by DIM and submitted to the WP20 partners. The purpose is to translate the terminological resources developed at various locations into this format until the next WP20 meeting, thus meeting the requirements for the next deliverable (May 2005) and providing sufficient content for joint publications.

In a further phase, MorphoSaurus subwords and corpora will be used to link the heterogeneous terminological sources. To this end, experiments have already been performed with the MeSH thesaurus. Machine learning techniques were used to map (sequences of) MorphoSaurus identifiers to MeSH terms, using Medline abstracts (with manually assigned MeSH descriptors) as a training set. New experiments of this kind will be possible after the integration of multilingual resources.

6.2.3 Population of Lexicons by Manual and Semi-automated Approaches

An English-Swedish medical lexicon is being built by LiU-IMT and LiU-IDA. The source for the lexicon is 40 000 medical terms from medical terminology systems existing in both English and Swedish. The terms are collected by LiU-IMT. The lexicon is built in an iterative process with automatic word alignment and manual revision. The tools used in the process are developed at LiU-IDA.

At Geneva, considerable effort is being put into the population of an English / French / German / Latin lexicon with anatomy terms.

The MorphoSaurus lexicon, maintained by UKLFR and UNIFR (JENA) was populated by several thousands of Spanish, Swedish and French lexemes which were automatically acquired and linked to existing synonym classes. They are currently being manually validated, refined and corrected. The quality of this work is constantly being assessed by mapping parallel texts to MorphoSaurus identifiers. An outcome-oriented evaluation has recently performed using a standardized framework which measures precision and recall of cross-language retrieval using the OHSUMED text collection. This test has shown a considerable improvement for Portuguese.

6.2.4 Integration with other Workpackages

Although not specified in the original work program, we see interesting prospects of linkage to the following workpackages:

- WP 21 (ontology): MorphoSaurus equivalence classes (which group (quasi-synonym) subwords) are atomic entities of meaning. A linkage to lexical ontologies such as UMLS and WordNet would constitute an important add-on. For this purpose a visit of Alessandro Oltramari from the CNR ontology group is planned for the first quarter of 2005.
- WP 24 (information retrieval): The MorphoSaurus indexer is, principally, neutral with regard to specific retrieval environment or search engines. However, search engines should be optimized to best support document retrieval mediated by subword identifiers. Several experiments were run using SMART and Boolean search engines (UKLFR). An agreement was made with WP 24 to use MorphoSaurus for indexing Medline abstracts.