

SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D20.4

Report: Evaluation of the Multilingual Medical Dictionary

Report Version: 1.0

Report Preparation Date: 27-Feb-07

Classification: RE

Contract Start Date: 2004-01-01

Duration: 3.5 years (42 months)

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	OBJECTIVES	3
2.1	Objectives of the Semantic Mining Workpackage 20.....	3
2.2	Objectives of the Deliverable 20.4.....	3
3	BACKGROUND	4
4	METHODS	5
4.1	Interchanging Lexical Information.....	5
4.2	Resources.....	7
4.3	Linking Format Definition.....	8
4.4	Cross-Lingual Alignment.....	9
5	RESULTS	11
5.1	Coverage.....	11
5.2	Correctness	13
6	DISCUSSION	14
7	CONCLUSION	16



1 Administrative Information

Lead contractor: UKLFR

Responsible: UKLFR

Participants: IMT-LiU, IDA-LiU, SU, UGOT, DIM, INSERM, JENA

Deliverable: D20.4

Author: Stefan Schulz

2 Objectives

The European Network of Excellence “SemanticMining” instantiates collaborations between 25 organisations from 11 European countries. In this deliverable the progress in work package 20 “Multilingual Medical Dictionary” will be reported. This report covers part of the research activities in this workpackage which were initiated in month 30 after the start of the NoE. Target audience of this deliverable is the Commission, project partners, and the public.

2.1 Objectives of the Semantic Mining Workpackage 20

The objective of the work package 20 was to pool resources and to create a standard for multilingual medical dictionaries to facilitate cross-language mapping between medical terminologies, to support cross-language text retrieval and corpus-based machine translation. Work package 20 aimed at testing several approaches to terminology mapping and terminology creation, including the manual and the automatic acquisition of lexicon entries and lexeme mappings.

2.2 Objectives of the Deliverable 20.4

Whereas the objectives of creating interchange formats for language-specific lexicon entries, semantic links, and corpora have already been met (see previous deliverables D201. – D20.3, and related scientific publications^{1,2}), the objective of the present report is to provide

1. An overview of the lexicon structure used, summarizing the content of D20.2
2. A description of the semantic mapping of lexicon entries
3. A report on the evaluation of the coverage and the correctness of the experimental (full term) multilingual lexicon as generated according to the above premises.

¹ Baud, Robert, Mikael Nyström, Lars Borin, Robert Evans, Stefan Schulz & Pierre Zweigenbaum (2005). Interchanging lexical information for a multilingual dictionary. In AMIA 2005 - Proceedings of the Annual Symposium of the American Medical Informatics Association, pp. 31-35.

² Markó, Kornél, Robert Baud, Zweigenbaum Pierre, Lars Borin, Magnus Merkel & Stefan Schulz (2006a). Towards a multilingual medical lexicon. In AMIA'06 - Proceedings of the 2006 Annual Symposium of the American Medical Informatics Association, pp. 534-538. Washington, D.C., November 11-15, 2006. American Medical Informatics Association.



3 Background

Lexicons, especially designed for natural language processing purposes, can generally be characterized along several dimensions. Firstly, lexicons can provide different amount of lexical information, such as part-of-speech, number, gender and case.

Secondly, the coverage of a lexicon, which often captures the terminology of a specialized domain, indicates for how many words of a (domain specific) text collection lexical information is available. For translation dictionaries, finally, special attention is drawn on the multilingual dimension.

There is currently no large electronic dictionary in the medical domain which is characterized by a true multilingual dimension, relevant coverage, and substantial lexical information at the same time. Of course, with the UMLS Metathesaurus³ there already exists a widely used multilingual resource with high coverage in the medical domain. However, lexical information which reaches beyond the mapping of terms from multilingual medical terminology systems (MeSH, ICD) is missing for other languages than English.

For non-specialized domains, remarkable effort of developing mono- and multilingual dictionaries has been made.

For example, WordNet⁴ provides a good coverage for general English. It may be useful for covering lay terminology of (bio-)medicine^{5,6} for example within a consumer-oriented health information system. The European counterpart, EuroWordNet⁷ tend towards a multilingual system, but with considerable diverse levels of lexical coverage.

Whenever medical terminology has been addressed in the construction of a multilingual dictionary with substantial lexical information, it lacks convenient coverage or has been developed as a demonstrative prototype⁸.

The MorphoSaurus⁹ subword lexicons (being enhanced as part of this workpackage, but kept separately from the multilingual lexicon described here), which align medical words in different languages on the subword-level, provide high coverage of medical terminology in different languages. But morpho-syntactic information such as part-of-

³ UMLS (2005). Unified Medical Language System. Bethesda, MD: National Library of Medicine.

⁴ Fellbaum, Christiane (Ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

⁵ Burgun, Anita & Olivier Bodenreider (2001). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In Proceedings of the NAACL 2001 Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations', pp. 77-82. Pittsburgh, PA, June 3-4, 2001. New Brunswick, NJ: Association for Computational Linguistics.

⁶ Bodenreider, Olivier, Anita Burgun & Joyce A. Mitchell (2003). Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study. In Robert Baud, Marius Fieschi, Pierre Le Beux & Patrick Ruch (Eds.), Medical Informatics Europe 2003 - Proceedings of the 18th International Congress of the European Federation for Medical Informatics. The New Navigators: From Professionals to Patients., Studies in Health Technology and Informatics 95, pp. 379-384. St. Malo, France, May 4-7, 2003. Amsterdam: IOS Press.

⁷ Vossen, Piek (Ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

⁸ Chiao, Y.C. & P. Zweigenbaum (2002). Looking for French-English translations in comparable medical corpora. In Isaac S. Kohane (Ed.), AMIA 2002 - Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical Informatics: One Discipline, pp. 150-154. San Antonio, TX, November 9-13, 2002. Philadelphia, PA: Hanley & Belfus.

⁹ <http://www.morphosaurus.net>



speech, case, gender, etc. is missing completely in this resource. Nevertheless, morpho-semantic indexing, i.e. the extraction of the morphological atoms (mostly morphemes) can be used for linking different monolingual resources into a multilingual repository with high coverage¹⁰.

Multilingualism means at least that corresponding entries in different languages are connected, which is a complex task and raises simple questions and concerns open issues, like e.g., in which cases a translation relationship truly holds for lexical entities. Therefore, syntactic as well as semantic criteria have to be developed, or, at least, a consensus of different lexical input providers has to be found.

Of course, monolingual resources exist for different languages, so the first step to merge them is to create a common framework for the integration of lexical entities from different languages, with respect to their intrinsic peculiarities.

4 Methods

4.1 Interchanging Lexical Information

The Interchange Format has been developed as a convention about the way to exchange linguistic information entering in the building process of a medical multilingual lexicon. The basic idea is that the exchange of information is performed through this Interchange Format only, and each contributor of lexical resources is converting available data into that representation.

Table 1 lists the fields of the interchange format. The most important ones are the following:

Lng: The language field determines to which language a particular entry belongs. Up until now, the values are: *EN* for English, *FR* for French, *DE* for German, *LA* for Latin, *SV* for Swedish, *ES* for Spanish and *PT* for Portuguese.

Id: This argument specifies the unique identifier of the multilingual lexicon entry, made of the concatenation of the name of the input provider and a consecutive number.

Typ: The type of entry defines either a basic entry (B), a subword entry (S), a compound entry (C) or a term entry (T). By definition, these types are mutually exclusive. The *basic entry* encodes single words of the language, generally without a space character in their lemma. The *subword entry* is a marker for parts of words entering in the composition of a *compound entry*. Therefore, a *SubWordEntry* can generally not be used standalone and a *compound entry* is for words, which have been explicitly recognized as a composition of two or more *subword entries*. Finally, a *term entry* (T) describes a sequence of words, generally separated by the space character.

Lem: The lemma is the representation of the entry in its basic form (singular, nominative for nouns; infinitive for verbs). It is supposed to be recoverable from any occurring form

¹⁰ Schulz, Stefan, Kornél Markó, Philipp Daumke, Udo Hahn, Susanne Hanser, Percy Nohama, Roosevelt Leite de Andrade, Edson Pacheco & Martin Romacker (2006). Semantic atomicity and multilinguality in the medical domain: Design considerations for the MorphoSaurus subword lexicon. In LREC 2006 - Proceedings of the 5th International Conference on Language Resources and Evaluation. Genua, Italy, May 24-26, 2006.



by an inflectional morphology process which is language dependent. There is exactly one unique basic form for any entry.

Mul: The code for encoding morphological and syntactic information is defined as in the open standard MULTTEXT (Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets)¹¹. Language dependent extensions of MULTTEXT may be used.

Frm: An entry that describes a specific inflected form that is linked to an entry for its lemma through the **Ref** field.

Mfr: The morpho-syntactic features of the inflected form using MULTTEXT exactly as for the **Mul** field.

Prt: The decomposition of compound entries.

Ref: If the entry consists of an inflected form, a unique ID of its lemma entry is given.

Field	Description	Definition
Lng	Language	the language to which pertains the present entry
Id	Multilingual Identifier	the unique identifier of this entry
Typ	Entry Type	one of the 4 allowed types of entry (B,C,S,T)
Err	Correctness	flag for correctness of this entry
Lem	Lemma	the entry in its basic form
Mul	Morpho-syntactic Features	the MULTTEXT morpho-syntactic tag of the lemma
Frm	Inflected Form	any inflected form
Mfr	Features of Inflected Form	the MULTTEXT morpho-syntactic tag of the inflected form
Inf	Inflection Model	language specific information
Mis	Language Specific Argument	to be used freely by provider of entries
Prt	Decomposition	the decomposition of a compound entry into its parts
Str	Head	the head word of the term
Ref	Reference Lemma	ID of its lemma's entry (if inflection form)
Exa	Typical Usage	a sentence presenting a typical usage of this entry
Com	Comment	any comment or warning about

¹¹ <http://nl.ijs.si/ME/V3/msd/related/msd-multext/>

Table 1: Fields of the Lexicon Interchange Format

4.2 Resources

After agreeing upon the Interchange Format, the Semantic Mining Partners UKLFR, UGOT, LIU, INSERM, and DIM, collected their monolingual lexical resources. These are:

- the French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries)¹²
- an English medical lexicon from Linköping University, Sweden (22,686 entries)¹³
- a Swedish medical lexicon from Linköping University (23,223 entries)¹³
- a Swedish medical lexicon from Göteborg University, Sweden (12,430 entries)
- the German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries)¹⁴
- In addition, the English Specialist Lexicon¹⁵, which is part of the UMLS (96,621 entries, avoiding acronyms and chemical names), has also been converted into the Interchange Format.

Lng	Typ	Lem	Mul	Frm	Mfr	Prt
FR	B	doigt	Ncms			
EN	T	finger nail	Nc-sn			
SV	B	digital	Afp-sn			
SV	C	Fingeravtryck	Nc-sn			Finger- avtryck
DE	B	Finger	Ncmsn	Fingers	Ncmmsg	
DE	C	Fingerfraktur	Ncfsn	Fingerfrakturen	Ncfpn	Finger- frakturen

Table 2: Sample of Compiled Lexical Resources (some fields omitted)

¹² Zweigenbaum, Pierre, Robert Baud, Anita Burgun, Fiammetta Namer, Eric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-Fran,cois Forget, Magaly Douy'ere & Stéfan Darmoni (2005). UMLF - a unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2/4):119-124.

¹³ Nyström, M., M. Merkel, L. Ahrenberg, H. Petersson & H. Ahlfeld (2006). Creating a medical English-Swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making*.

¹⁴ Weske-Heck, Gesa, Albrecht Zaiss, Stefan Schulz, Wolfgang Giere, Michael Schopen & Rüdiger Klar (2002). The German Specialist Lexicon. In Isaac S. Kohane (Ed.), *AMIA 2002 - Proceedings of the Annual Symposium of the American Medical Informatics Association: Biomedical Informatics: One Discipline*, pp. 884-888. San Antonio, TX, November 9-13, 2002.

¹⁵ <http://www.nlm.nih.gov/pubs/factsheets/umlslex.html>



Up until now, 224,351 lexical entries for the biomedical domain, fully encoded with morpho-syntactic features, were collected covering four languages (cf. Table 2 for a sample: The first character of the *Mul* field encodes the part-of-speech: *N* (noun), *A* (adjective). In case of nouns, *c* denotes common nouns, *m* masculine, *s* singular, *n* neuter or nominative, depending on the position. For adjectives, *f* stands for qualitative, *p* for positive.

The character “–” indicates that a particular feature does not fit into the language given (e.g. gender in English) or is unspecified for this entry. The number of different lemmas (thus, ignoring ambiguous lexical information for an entry such as, e.g., case) is 105,317 for English, 29,822 for French, 27,480 for German, and 27,093 for Swedish (a total of 189,712, therefore, 1.2 morpho-syntactic variants are given per lexical entry, in average).

4.3 Linking Format Definition

The cross-lingual grouping of corresponding entries is the essence of a multilingual dictionary. This operation transforms a set of monolingual lexicons into a multilingual dictionary. Before this operation, the dictionary entries are independent; afterwards, they are organized as clusters of synonyms or translations. Multiple lexical entries, either in the same language or in different languages, are the expression of the same object in the reality with a common part of speech argument (POS). Typically, *clavicle* in English and *clavicule* in French denote unambiguously the same object (a bone of the pectoral girdle) and they share the same POS: a common noun. The two corresponding entries are candidates to be linked by a translation relation. A similar relation could be defined with the corresponding adjectives, *clavicular* and *claviculaire*. Unfortunately, the process of translating lexical items is not that straightforward, and a couple of cross-lingual phenomena are problematic to capture, especially regarding the different characteristics of case, gender and number in different languages, as well as multiple derivations, e.g. for adjectives, dependent on whether a definite or indefinite object follows or whether their use is attributive or predicative.

Consider the German (Swedish) words *Schere* (*sax*), *Hose* (*bralla*) (both noun, singular), *Scheren* (*saxar*), *Hosen* (*brallor*) (both noun, plural) and the English equivalents, *scissors* and *trousers* (both noun, plural). Singular forms of the latter examples do not exist¹⁶, whilst for other pairs of lexemes, of course, singular forms can be translated to a corresponding singular form in the other language. This information should be kept in a multilingual lexicon, e.g. for the use in machine translation applications.

Field	Description	Definition
Src	Source Entry ID	ID of the source entry to be linked to a target entry
Tar	Target Entry ID	ID of the target entry linked from the source entry
Typ	Link Type	Type of relation

¹⁶ except for noun compounds, as evidenced by “trouser board” or “scissor kick”

Table 3: Fields of the Linking Format

Different languages also make different use of grammatical gender or noun classes. Whilst in German, Greek or Latin, three grammatical genders are distinguished (masculine, feminine and neuter), French, Portuguese, and Spanish only use two (masculine, feminine). Swedish and Danish discriminate the classes *common* and *neuter*. Finally, English does not account for any of these features at all. In a first version, in order to find an agreement on the question, in which cases two lexical items from different languages, *A* and *B*, can be regarded as translations (or, within one language, synonyms) of each other, the following "grades" of bidirectional relationships are defined:

Synonymy/Translation (S/T): *A* and *B* share the same part of speech (POS) and all MULTEXT features, except of gender

Synonymy/Translation, inflected (S/T-i): *A* and *B* share the same POS, but at least one MULTEXT feature differs;

Synonymy/Translation, derived (S/T-d): *A* and *B* do not share the same POS

Having these types of relations in mind, a simple Linking Format was created, which is depicted in Table 3.

Having defined the morpho-syntactic framework, in which lexical relationships can be coded, two methods are used for the cross-lingual alignment of lexical entities on the semantic level.

4.4 Cross-Lingual Alignment

A great deal of work has already been done for the fully automatic cross-lingual alignment of lexical items, most of them using aligned corpora and employing statistical methods, such as context vector comparison^{17, 18, 19} or mutual information statistics²⁰. Considering the medical domain, in which multilingual resources are available, e.g. within the UMLS, methods for the automatic search for translation candidates have also already been explored. One idea was to use already existing translations at a subword level in order to support the acquisition of translations at a term level^{21, 22}. Therefore, the

¹⁷ Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 519–526. College Park, MD, USA, 20–26 June 1999. San Francisco, CA: Morgan Kaufmann.

¹⁸ Widdows, Dominic, Beate Dorow & Chiu-Ki Chan (2002). Using Parallel Corpora to enrich Multilingual Lexical Resources. Third International Conference on Language Resources and Evaluation, Las Palmas, May 2002, pages 240–245.

¹⁹ Déjean, Hervé, Eric Gaussier & Fatiha Sadat (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In COLING 2002 - Proceedings of the 19th International Conference on Computational Linguistics, pp. 218–224. Taipei, Taiwan, August 24 -September 1, 2002. Association for Computational Linguistics.

²⁰ Fung, Pascale (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, Third Conference of the Association for Machine Translation in the Americas, pages 1–16. Springer, October 1998.

²¹ Namer, Fiammetta & Robert Baud (2005). Predicting Lexical Relations between Biomedical Terms: towards a Multilingual, Morphosemantics-based system'. Medical Informatics Europe Congress MIE 2005 (28 aout au 1er septembre) - Studies in Health and Technology Information, Genève, Pages 793–798

²² Daumke, Philipp, Stefan Schulz & Kornél Markó (2005). Searching multilingual medical content in the Web. Technology and Health Care, 13(5).

MorphoSaurus^{23, 24, 25} system seems particularly well suited for the cross-lingual linkage of available monolingual lexicons.

MorphoSaurus is a semantic indexing engine based on a multilingual thesaurus of subwords. It takes medical text as input and transforms them in three steps, viz. orthographic, morphological and semantic normalization. This transformation yields, finally, a mapping to unique term class identifiers contained in an intermediary subword lexicon (cf. Fig. 1).

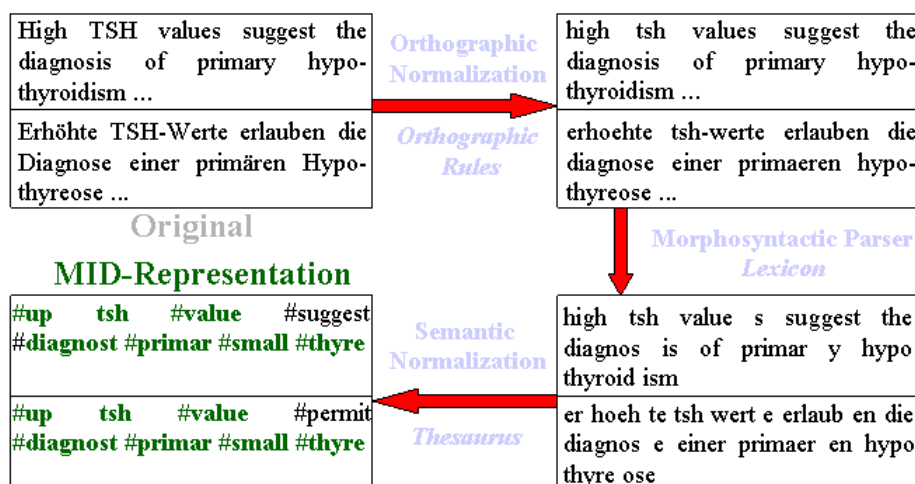


Fig. 1 Morphosemantic Normalization using the MorphoSaurus system

The result of MorphoSaurus is a simplified, language-independent representation of the input text. It has proved useful in cross-language document retrieval experiments²⁶, as well as in the automatic assignment of MeSH codes to documents²⁷.

In a first step, all lexical entries are processed with the morpho-semantic indexing procedure MSI, as described in Section 3.2. After resolving ambiguous MIDs (Chapter 7), a quite simple algorithm was used to perform the mappings between all entries: Every lexeme i and its attributes is compared to any other lexeme j in the list. If their representations in the interlingua format are identical, they are considered as potential translations or synonyms and linked. Then the relation type (S/T , $S/T-i$, $S/T-d$, cf. previous section) is determined, by comparing the lexical attributes of the items involved.

²³ <http://www.morphosaurus.net>

²⁴ Schulz, Stefan, Kornél Markó, Philipp Daumke, Udo Hahn, Susanne Hanser, Percy Nohama, Roosevelt Leite de Andrade, Edson Pacheco & Martin Romacker (2006). Semantic atomicity and multilinguality in the medical domain: Design considerations for the MorphoSaurus subword lexicon. In LREC 2006 - Proceedings of the 5th International Conference on Language Resources and Evaluation. Genua, Italy, May 24-26, 2006.

²⁵ Markó, Kornél, Stefan Schulz, Alyona Medelyan & Udo Hahn (2005f). Bootstrapping dictionaries for cross-language information retrieval. In SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 528-535. Salvador, Brazil, August 15-19, 2005. New York, NY: ACM.

²⁶ Daumke, Philipp, Stefan Schulz & Kornél Markó (2005a). A CLIR interface to a Web search engine. In SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brasil, August 15-19, 2005.

²⁷ Markó, Kornél, Phillip Daumke, Stefan Schulz & Udo Hahn (2003). Cross-language MeSH indexing using morpho-semantic normalization. In Mark A. Musen (Ed.), AMIA'03 - Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications, pp. 425-429. Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus.

Alternatively, an alignment technique was used for the English/Swedish lexicon, exploring a parallel collection of rubrics from the medical terminology systems ICD-10, ICF, MeSH, NCSP and KSH97-P, as described in²⁸.

5 Results

Using the algorithm introduced, 651,542 bi-directional relations between lexemes were obtained, a sample of which is depicted in Table 4. For English-German, 126,504 translations have been generated (31,544 when only different lemmas are taken into account, thus ignoring ambiguous lexical information), for English-French 70,680 (24,368, respectively) and for English-Swedish 86,655 (34,030). Furthermore, 21,604 (8,312) relations have been extracted for French-Swedish, 32,659 (10,458) for French-German and finally, 41,469 (12,105) for German-Swedish. All other relations (271,971) cover intralingual synonymy. The distribution of different types of relations is 66,641 occurrences for S/T (10%), 286,880 for S/T-i (44%) and 298,021 for S/T-d (46%).

Typ	Lng-1	Lem-1	Mul-1	Lng-2	Lem-2	Mul-2
S/T	EN	abdominal hernia	Nc-sn	SV	bukbräck	Nc-sn
S/T-i	EN	abdominal aorta	Nc-sn	DE	Bauchaorten	Ncfn
S/T-d	EN	alveolar	Afp-n	FR	alvéole	Ncfs

Table 4: Sample Links between Lexical Items

5.1 Coverage

The UMLS Metathesaurus is the most comprehensive resource for medical terminology. Therefore, it is particularly interesting how many terms of the UMLS are covered by the multilingual lexicon. Table 5 (second column) gives the numbers for those items in the Metathesaurus, which are marked as a preferred entry and only contain alphabetic characters (thus, multi-word entries and chemical compounds are not considered in the following discussion). Column three gives the number of those UMLS entries, which are covered by the multilingual lexicon. Values range between 13% for German up to 71% for Swedish. The numbers in Column four show how many synonyms and morpho-syntactic variants of UMLS terms are listed in the lexicon which are not part of the Metathesaurus, and, therefore, could be added. This consideration only takes those variants into account, which share at least the same part of speech with the corresponding UMLS entry (only S/T and S/T-i). Finally, the number of additional lexemes in the lexicon that are neither found in the Metathesaurus, nor constitute morpho-syntactic variants of existing UMLS entries, is depicted in Column five. In total, the multilingual

²⁸ Nyström, M., M. Merkel, L. Ahrenberg, H. Petersson & H. Ahlfeld (2006). Creating a medical English-Swedish dictionary using interactive word alignment. BMC Medical Informatics and Decision Making.



lexicon contains 189,712 different lemmas, i.e. 24,243 more than the part of the UMLS considered here.

Language	UMLS	Covered	Synonyms	Additional
English	122,035	32,668	3,807	68,842
German	21,162	2,832	1,269	23,379
French	10,260	3,590	309	25,923
Swedish	12,012	8,520	994	17,579
Σ	165,469		165,469	

Table 5: Comparison of Lexical Entries: UMLS Metathesaurus and Multilingual Lexicon

Language Pair	UMLS	Covered	Synonyms	Additional
English-German	15,979	1,259	8,801	21,484
English-French	12,589	1,783	6,974	15,611
English-Swedish	9,554	3,403	10,124	20,503
German-French	9,859	850	773	8,835
German-Swedish	10,063	810	1,699	9,596
French-Swedish	6,793	1,109	1,911	5,292
	64,837		120,817	

Table 6: Comparison of Cross-Lingual Mappings

For the language pairs considered, the UMLS Metathesaurus already contains between 6,700 and 16,000 translations (cf. Table 6, Column two). Within a range of 8% (EN-DE and DE-SV) to 36% (EN-SV), these mappings are also included in the multilingual lexicon (Column three). A total of 30,282 synonymous entries (Column four) could be added to 64,837 existing UMLS translations. Finally, those cross-lingual mappings which are captured in the multilingual lexicon but not in the UMLS Metathesaurus, sum up to 81,321 alignments (again, only considering the relations S/T and S/T-i). While there are 64,837 word-to-word translations in the UMLS for the languages considered, the multilingual lexicon contains 120,817 different translations.

5.2 Correctness

In order to assess the correctness of linkage between semantically equivalent lexicon entries, the following method was used: for the language pair DE-EN, FR-DE, EN-SV a random sample (n=500) was generated and analyzed by persons both familiar with medical terminology and the languages involved.

The following parameters were considered:

- Correct equivalent: Yes/No (i.e. the source and the target entry correspond semantically). Style issues are not considered. For example, a layman term may very well correspond to a professional medical term. As correct equivalents are considered synonyms (e.g. English “alcohol” can be given at least two correct entries in Swedish, “alcohol” and “sprit”), polysemous (e.g. the polysemous German “Bruch” can translate to English “hernia” and “fracture”) , and homographs (e.g. “alcoholic” adj. “alkoholhaltig”, “alcoholic” noun “alkoholist”)
- Medical term entry: Yes/No (if no general language item is assumed). Requires that both the source and the destination entries are medical. The criterion for “medical term” is that the term denotes an entity that requires medical knowledge to be fully described (this rules out common language terms such as headache, hand, stomach)
- Correct lemma: Yes/No (i.e. the term pair may be correct semantically but the form of at least one of the lemmas is incorrect). As was pointed in the Interchange Format Definition paper, the source entry should be in basic form, defined there as singular number, masculine gender, nominative case, and infinitive mode for verbs, whatever applies, depending on the part of speech and the given language. For example, the English „alginolyticu“ vs. the Swedish „alginolyticus“ is in principle a correct entry semantically, but the lemma (base form of the English entry) lacks an “s”, making it incorrect formally.
- Thesaurus variant: Yes/No. (i.e. the term pair shares significant conceptual content), such as in
DIM:20743|alcohol|LIU_SV835_A|alcoholism|CCT||morphoSaurusAlign

Due to the heterogeneous sources of the lexicon entries, in the English-Swedish lexicon, many doublets and triplets are found, e.g.

DIM:20743|alcohol|LIU_SV8231_A|etanol|CCT||morphoSaurusAlign

DIM:20743|alcohol|UGOT:564-564|etanol|CCT||morphoSaurusAlign

destination-ID will be anyhow kept in the database, as they are needed
It was decided to count these entries only once.

Two modes of cross-language mapping were tested, the MorphoSaurus based mapping as introduced above for all languages, and for Swedish-English, alternatively, a mapping based on Swedish lemmas.

The results are given in Table 7.

Language Pair / Mapping methods	Fully Correct mappings (%)	Fully and Partly Correct mappings (%)	Medical Entries (%)	No. of Entries
English-German (MorphoSaurus mapping)	66.8	78.2	89.4	
German-French (MorphoSaurus mapping)	67.4	83.6	90.2	
English-Swedish (MorphoSaurus mapping)	67.2	77.7	80.0	24 395
English-Swedish (LiU lexicon alignment)	94.8	96.6	89.6	23,430

Table 7: Comparison of Cross-Lingual Mappings

6 Discussion

The interlingual term mapping studies proved the usefulness for the previously defined mapping format, so that this format can be recommended as a standard for the encoding of lexical information in or domain, and for the languages we were working with. Regarding the content, the resources available in SemanticMining were, by far, not sufficient for the generation of a ready-to-use medical lexicon. However, we reached the lexical coverage as proposed in the original description of work.

Analyzing the correctness, first of all, we observe a considerable difference between the two methods, namely the use of alignment techniques between parallel corpora on the one hand (which yield a good result) and the subword mapping on the other hand (which yields a less impressive result). This finding is explained by the several facts:

- The terms in aligned corpora are used within their context and need not to be reduced to their constituent parts, which, in many cases do not explain the meaning of the compound terms, such as in the translation error. “hemisphere”(FR) – “Hemiballismus”(GE);
- The model of the MorphoSaurus approach is deliberately coarse-grained since it is tailored for text retrieval and not to exact translation. Many suffixes are ignored for indexing, which explains, for instance, the erroneous translation pair “therapist”(EN) – “Therapie”(GE);

- Some equivalences are still underspecified. For instance, the erroneous translation “hair”(EN) – “Trichiasis”(GE) could be avoided by adding the subword “trichias-”, since it refers to a very special disorder of a special kind of hair.
- In some long words there still occur segmentation errors.

On the other hand we have to recognize that the alignment method inevitably fails when the words under scrutiny are not in the corpora, which is especially to be expected with highly complex nominal compounds such as “pseudopsychopathique”, “rhinoscléromatis”, “macroductyly”, “Hyperglyzeridämie”, or “subependymoma”, word that can only be aligned by the MorphoSaurus approach.

Since coverage and correctness are basic requirements imposed on the lexicons, some form of post-processing of MorphoSaurus-aligned lexicons is required to further improve the evaluation scores. Here are some possible NLP-based checks which could be performed before a final multi-lingual compilation takes place. These checks are aimed in the first place at verification of the lexical layer of the aligned lexicons.

- Correct spelling of input and output lemmas in source and target language against available data in the collected lexicons and other on-line available medical glossaries or corpora;
- Part of speech checking of the equivalent pairs against available monolingual and bilingual dictionaries to avoid alignment between incongruent parts of speech;
- Prefix/suffix check in concerned languages (to avoid links of type “alcohol” / “alcoholism”)
- Interlingua check. If English is assumed as interlingua and a semi-automatic or manual check would be performed on pairs of languages with English as either input or target language, considerable improvement could be obtained even for remaining language pairs;
- Ontological check. Equivalence pairs whose nodes in an ontological classification show deviating results could be automatically subjected to manual revision for either rejection or re-structuring in case they are polysemous or homographic items (e.g. English /polyposia/ and Swedish /polyp/ would be automatically marked as semantically incongruent);
- Identify synonym spelling variants in each language by using simple string similarity checks. Sometimes there are variant medical spellings used in different sources, often only differing in one or two characters.
- Register label expansion. Given monolingual medical corpora, tagged with the distinction between professional and layman authorship, it would be feasible to arrive at register attributes which could label medical usage of specific terms as either professional or patient terms (or possibly both).



7 Conclusion

For the construction of multilingual medical lexicons we have proposed a standardized description format for Western European languages suited to encode the morphosyntactic information of lexicon entries, as well as a linkage format for expressing synonymy and translation relations. In a pilot study we have used two different kinds of techniques for automated mapping between translation, a probabilistic approach exploiting parallel corpora on the one hand, and a heuristic approach using a subword thesaurus. We discussed strengths and weaknesses of either approach, and suggest future research which yields to a stricter validation of automatically proposed translations. Although we do not expect to build semantic links between lexicon entries in a fully automated manner, our method is suited to speed up the time and cost intensive manual lexicon building process.

A possible way to exploit the experience gained in WP 20 is to seek industrial partnership where we could apply our methodology on a larger scale and with the resources necessary.

We should mention that commercial exploitation already takes place with regard to the MorphoSaurus lexicon. Although the development of this resource has not constituted the main focus of WP 20, it has consumed some of the WP20 effort (as already described in the DoW), exactly for supporting the lexicon mapping as described in this report.