

# Cross-Lingual Alignment of Medical Lexicons

Kornél Markó<sup>1</sup>, Robert Baud<sup>2</sup>, Pierre Zweigenbaum<sup>3</sup>, Magnus Merkel<sup>4</sup>,  
Maria Toporowska-Gronostaj<sup>5</sup>, Dimitrios Kokkinakis<sup>5</sup>, Stefan Schulz<sup>1</sup>

<sup>1</sup>Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany

<sup>2</sup>University Hospitals of Geneva, Service of Medical Informatics, Geneva, Switzerland

<sup>3</sup>Inserm, U729; Assistance Publique – Paris Hospitals, STIM; Inalco, CRIM, Paris, France

<sup>4</sup>Linköping University, Department of Computer and Information Science, Linköping, Sweden

<sup>5</sup>Göteborg University, NLP Section, Department of Swedish, Göteborg, Sweden

## Abstract

We present an approach for the creation of a multilingual medical dictionary for the biomedical domain. In a first step, available monolingual lexical resources are compiled into a common interchange format. Secondly, according to a linking format defined by the authors, the cross-lingual mappings of lexical entries are added. We show how these mappings can be generated using a morpho-semantic term normalization engine, which captures intra- as well as interlingual synonymy relationships on the level of subwords.

## 1. Introduction

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

Within the European Network of Excellence “Semantic Interoperability and Data Mining in Biomedicine”, a multinational team of researchers from (computational) linguistics, medicine, and medical informatics, including the authors, gathered in a series of meetings with the goal of building a European multilingual medical lexicon with high coverage and the integration of complete morpho-syntactic information.

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

## 2. Interchange Format Definition

The Interchange Format is a convention about the way to exchange linguistic information entering in the building process of a medical multilingual lexicon (Baud et al., 2005). The basic idea is that the exchange of information is performed through the Interchange Format only, and each contributor of lexical resources is converting his or her data into that representation.

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

- **Typ** The *basic entry* (B) encodes single words. The *subword entry* (S) is a marker for parts of words entering in the composition of a *compound entry* (C). Finally, a *term entry* (T) describes a sequence of words.
- **Lem** The lemma is the representation of the entry in its basic form. It is supposed to be recoverable from any occurring form by an inflectional morphology process.
- **Mul** The code for encoding morphological and syntactic information is defined as in the open standard MULTTEXT.<sup>1</sup>

<sup>1</sup>Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets (<http://nl.ijs.si/ME/V3/msd/related/msd-multext/>)

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)
Lem	Lemma	the entry in its basic form
Mul	Morpho-syntactic Features	the MULTEXT morpho-syntactic tag of the lemma
Frm	Inflected Form	any inflected form
Mfr	Features of Inflected Form	the MULTEXT 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)

Table 1: Fields of the Lexicon Interchange Format

- **Frm** 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 MULTEXT exactly as for the **Mul** field.

Table 2 shows an excerpt of different lexicons encoded in the Interchange Format. One obvious shortcoming is that the different lexical resources provide different amounts of information.

### 3. Monolingual Resources

After agreeing upon the Interchange Format, partners from five different institutions collected their monolingual lexical resources.<sup>2</sup> These are:

- the French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries) (Zweigenbaum et al., 2004)
- an English lexicon from Linköping University, Sweden (22,686 entries)
- a Swedish lexicon from Linköping University (23,223 entries)
- a Swedish lexicon from Göteborg University, Sweden (6,786 entries)
- the German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries) (Weske-Heck et al., 2002)

<sup>2</sup>The English Specialist Lexicon, which is part of the Unified Medical Language System (UMLS, 2005), will be included in future work.

Up until now, 127,730 lexical entries for the biomedical domain, fully encoded with morpho-syntactical features, were collected covering four European languages (cf. Table 2 for a sample<sup>3</sup>).

### 4. Linking Format Definition

The cross-lingual grouping of corresponding entries is the essence of a multilingual dictionary.

Unfortunately, this is not a straightforward process 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 words *Schere* and *Hose* (both noun, singular) and the English equivalents, *scissors* and *trousers* (both noun, plural). Singular forms of the latter examples do not exist, whilst for other examples, of course, singular forms can be translated to a corresponding singular form in the other language.

Different languages also make different use of grammatical gender or noun classes. Whilst in German, Greek or Latin, three grammatical gen-

<sup>3</sup>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* 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.

Lng	Id	Typ	Lem	Mul	Frm	Mfr	Prt	Str
FR	UMLF:10081	B	doigt	Ncms				
EN	LIU:EN8427	T	finger nail	Nc-sn				nail
SV	LIU:SV6663	B	digital	Afp-sn				
SV	UGOT:3373	C	fingeravtryck	Nc-sn			finger-avtryck	avtryck
DE	UKLFR:39556	B	Fingerpanaritium	Ncnsn	Fingerpanaritiem	Ncnpa		

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

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

Table 3: Fields of the Linking Format

ders are distinguished (masculine, feminine and neuter), French and Italian 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, *A* and *B*, can be regarded as translations (or, within one language, synonyms) of each other, we defined the following "levels" of relationships:

1. **Rel1:** *A* and *B* share the same part of speech (POS) and all MULTEXT features
2. **Rel2:** *A* and *B* share the same POS, but at least one MULTEXT feature differs
3. **Rel3:** *A* and *B* do not share the same POS

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

So far, the meaning of words and their possible translations have not been discussed. In the following section, we show how lexical entities can be aligned on the semantic level.

## 5. Cross-Lingual Alignment

For the medical domain, methods for the automatic search for translation candidates have already been explored. One promising idea is to use already existing translations at a subword level in order to support the acquisition of translations at a term level (Namer and Baud, 2005). For the linkage of lexemes on the semantic level, we make use of the MORPHOSAURUS system (Markó et

al., 2005), a text normalization engine using subword lexicons for different languages, as well as a multilingual thesaurus.

### 5.1. Morpho-Semantic Indexing

The MORPHOSAURUS system is based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages, we observe a high frequency of complex word forms such as in '*pseudo⊕hypo⊕para⊕thyroid⊕ism*'. To properly account for particularities of 'medical' morphology, the notion of subwords was introduced as self-contained, semantically minimal units.

Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them. Subwords are listed as entries together with their attributes such as language and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned to one or more morpho-semantic identifier(s) representing the corresponding synonymy classes (MIDs). Intra- and interlingual semantic equivalence are judged within the context of medicine only.

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized (top-right). Next, words are segmented into sequences of subwords or left unaffected when no subwords can be decomposed (bottom-right). Finally, each meaning-bearing

Src	Tar	Typ	Lng1	Lem1	Mul1	Lng2	Lem2	Mul2
LIU:EN147	LIU:SV151	REL1	EN	abdominal hernia	Nc-sn	SV	bukbråck	Nc-sn
LIU:EN143	UKLFR:34985	REL2	EN	abdominal aorta	Nc-sn	DE	Bauchaorten	Ncfn
LIU:EN947	UMLF:1123	REL3	EN	alveolar	Afp-n	FR	alvéole	Ncfs

Table 4: Sample Links between Lexical Items (some fields omitted). Additionally, the MULTEXT values of the corresponding items are depicted in Column four to nine. Also cf. Footnote 3 for the explanation of *Mul* values.

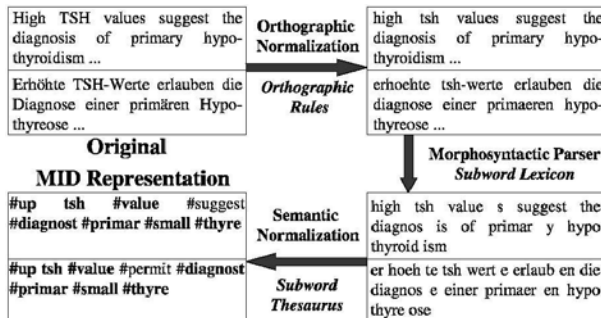


Figure 1: Morpho-Semantic Indexing Pipeline

subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). MIDs which co-occur in both document fragments appear in bold face.

## 5.2. Linking Algorithm

In a first step, all lexical entries are processed with the MORPHOSAURUS system. Afterwards, 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 (REL1, REL2 or REL3, cf. previous section) is determined, by comparing the lexical attributes.

## 6. Results and Conclusion

Using the algorithm introduced, we obtained a list of 300,894 bi-directional relations between lexemes, some of which are depicted in Table 4. For English-German, 30,716 translations have been generated, for English-French 16,123 and for English-Swedish 27,441. Furthermore, 21,619 relations have been extracted for French-Swedish, 32,805 for French-German and finally, 41,966 for German-Swedish. All their relations (130,224) cover intralingual synonymy. The distribution of different types of relations is 32,805 occurrences for REL1 (11%), 147,145 for REL2

(49%) and 120,944 for REL3 (40%). First examinations of the data proved many alignments to be valid. Of course, an extensive evaluation of the multilingual medical lexicon is still due. Further work will also examine relations with ongoing lexicon standardization efforts such as the Lexical Markup Framework of ISO/TC 37/SC 4.<sup>4</sup>

**Acknowledgments:** This work was supported by the European Network of Excellence “Semantic Mining” (NoE 507505).

## 7. References

- Robert Baud, Mikael Nyström, Lars Borin, Robert Evans, Stefan Schulz, and Pierre Zweigenbaum. 2005. Interchanging lexical information for a multilingual dictionary. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 31–35.
- Kornél Markó, Stefan Schulz, and Udo Hahn. 2005. Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545.
- Fiammetta Namer and Robert Baud. 2005. Guessing lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In *Proceedings of the 19th International Congress of the European Federation of Medical Informatics*.
- UMLS. 2005. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.
- Gesa Weske-Heck, Albrecht Zaiss, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rüdiger Klar. 2002. The German Specialist Lexicon. In Isaac S. Kohane, editor, *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 884–888.
- Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère, and Stéfan Darmoni. 2004. A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2):119–124.

<sup>4</sup><http://tagmatica.fr/doc/ISO24613cdRev7.pdf>