Multilingual Biomedical Dictionary

Freiburg University Hospital[1] MediLOG
Jena University[2] JULIE
Pontificial University of Paraná[3] HTMP

Philipp Daumke[1], Kornél Markó[1,2], Michael Poprat [1,2], Stefan Schulz[1,3]

Morphosaurus

The development and maintenance of a conventional multilingual dictionary is a time-consuming and expensive task which requires both domain and linguistic knowledge. We present an alternative approach based on the **Morphosaurus System** by which time and cost can be considerably reduced due to the use of subwords.

Morpho-Semantic Indexing (MSI) is a term normalization methodology developed by the authors which deals with various morphological processes in different languages. MSI uses a special type of dictionary, whose entries consist of **subwords**, i.e. semantically minimal units. Subwords are grouped into **language independent equivalence classes**, represented by morpheme identifiers (MIDs). A morphosyntactic parser extracts subwords from texts and assigns MIDs in a three step procedure (cf. Figure 1).

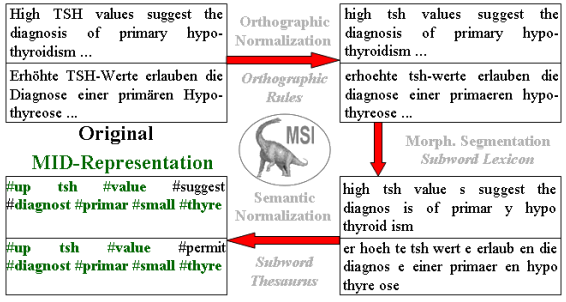

Figure 1: Morpho-Semantic Indexing (MSI)

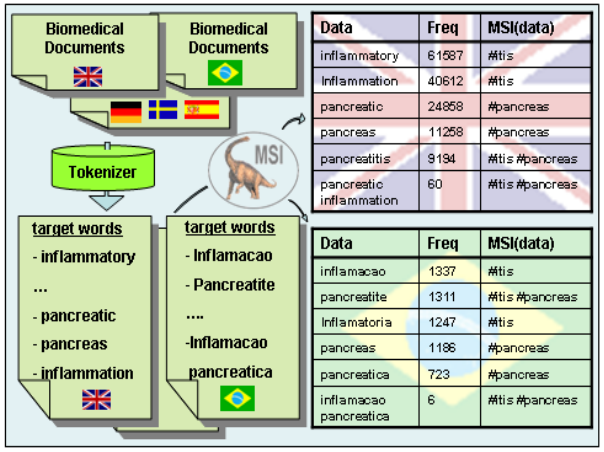

Figure 2: Generation of target word databases

Multilingual Biomedical Dictionary

We acquired **domain and language specific corpora** from various medical sources in the WWW. Using a tokenizer we then created large **lists** of surface words, bigrams and trigrams **of adjacent words** containing their frequencies within these corpora (target words).
All target words are translated to a set of MIDs and stored in language specific databases. These databases consist of about 3 M entries each.

A user can **query the dictionary** via a web interface. Again, this query is firstly altered to a **set of corresponding MIDs**. This MID set is used to create a list of possible reading variants (partitions). Each partition consists of one or more subwords which are now **compared to the relevant databases**. All matching records are finally sorted using several heuristics and presented to the user on the web interface.

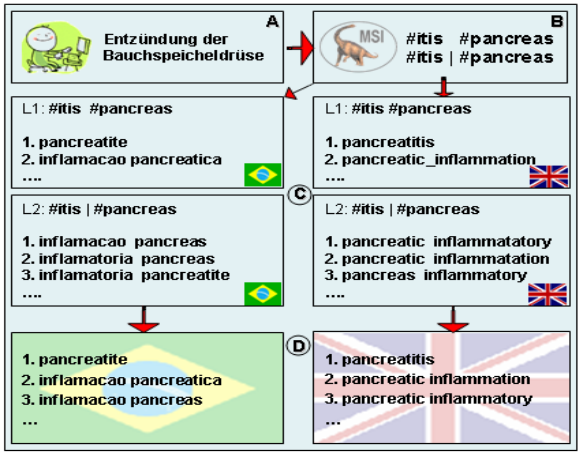

Figure 3: Output of the dictionary
(here in Portuguese and English)

www.morphosaurus.net -> Web Tools -> Medical Dictionary

Contact: Philipp Daumke, Freiburg Medical University, Dept. Medical Informatics, Freiburg, Germany, http://www.morphosaurus.net, email: daumke@coling.uni-freiburg.de