Semantic Atomicity and Multilinguality in the Medical Domain: Design Considerations for the **MorphoSaurus** Subword Lexicon

Stefan Schulz, Kornél Markó, Philipp Daumke, Udo Hahn, Susanne Hanser, Percy Nohama, Roosewelt Leite de Andrade, Edson Pacheco, Martin Romacker

Medical Informatics, Freiburg University Hospital, Freiburg, Germany, Health Informatics Laboratory, Paraná Catholic University, Curitiba, Brazil, Jena University Language & Information Engineering (JULIE) Lab, Jena, Germany, Text Mining in Life Sciences Informatics, Novartis, Basel, Iexitzerland

Context: Subword indexing for multilingual semantic document indexing

Medical sublanguage: large, dynamic, multi-lingual, heterogeneous user community, rich morphology, highly derivative, single-word compounds, expert-layperson language gap

Subwords as atomic sense units

Atomic senses (in a given language and a given domain context) cannot be univocally derived from the sense(s) of its lexical constituents. Atomic senses can inhere in word stems: (*hepat-*), affixes (*anti-*, *hyper*, *-ectomy*, *-logy*), word fragments (*diagnost-*, *hypophys-)*, straight words (*milz*, *spleen*), combinations of words (*yellow fever*, *vitamin C*).

Representation of (sub)word senses

Each sense is represented by one *MID* (MorphoSaurus ID)
D = (lexeme, MID, domain, language)

- Synonymy:*(lex_1; MID_1; dom_1; $lang_1$);(lex_2; MID_1; dom_1; $lang_1$); (lex_3; MID_1; dom_1; $lang_1$)*
  Example*: nephr-, ren-, kidney*
- Translation*: (lex_1; MID_1; dom_1; $lang_1$); (lex_2; MID_1; dom_1; $lang_2$)*
  Example*: nephr-, riñon*
- Ambiguity*: (lex_1; MID_1; dom_1; $lang_1$); (lex_1; MID_2; dom_1; $lang_1$)*
  Example*: head* (body part vs. chief)
- Coincidence:*(lex_1; MID_1; dom_1; $lang_1$); (lex_1; MID_2; dom_1; $lang_2$)*
  Example*: era* (epoch vs. Spanish past of "to be")
- Domain specificity*: (lex_1; MID_1; dom_1; $lang_1$); (lex_1; MID_2; dom_2; $lang_1$)*
  Example*: aspirin* (in dom_2 brand name ≠ substance)

MIDs can be interrelated by two relations:

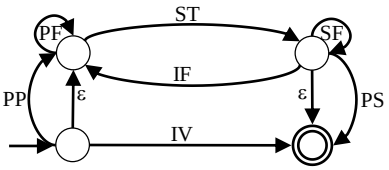
- *Expands* (MID_0; [MID_1; MID_2; :::; MID_3])
  Use: express composed meaning which cannot be suitably expressed by the word composition.
  Example: *Expands*($MID_{urinanalysis}$; [MID_{urine}; $MID_{analysis}$])
- *Has-Sense* (MID_0; {MID_1; MID_2; :::; MID_3})
  Use: treatment of lexical ambiguities.
  Example: *Has-Sense* (MID_{head}; {MID_{caput}; MID_{chief}})

Implementation in MorphoSaurus

Classification and description of lexicon entries in terms of:

- Lexeme classes:
  - Stems (ST), e.g. *hepat*, *enferm*, *diaphys*, *head*
  - Prefixes (PF), *de-*, *re-*, *in-*,
  - Proper Prefixes (PP) cannot be prefixed, e.g. *peri-*, *hemi-*, *down*
  - Infixes (IF), like *-o-*, e.g., in *gastr-o-intestinal*,
  - Sufixes (SF) e.g. *-a*, *-io*, *-ion*, *-tomy*, *-itis* follow a
  - Proper Sufxes (PS) cannot be suffixed, *-ing*, *-ieron*, *-ção*,
  - Invariants (IV), occur isolated e.g. *ion* or *gene*
- Language (English, French, German, Swedish, Spanish, Portuguese)
- MID (equivalence class identifier), only assigned to semantically relevant lexemes
- Inter-MID relations *Expands* and *Has-Sense (see above)*

Word parser



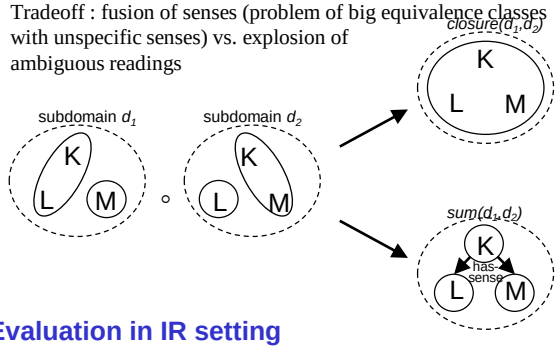
Morphosemantic indexing example

Original Document	Orthographic Normalization	Morphological Segmentation	Semantic Normalization
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value#** #suggest# **#diagnost# #primar# #small# #thyre# #suppress# tsh #nivell#** #suggest# **#up# #thyre#** .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	**#up# tsh - #value#** #permit# **#diagnost# #primar# #small# #thyre#** , **#suppress# tsh** - {#mirror# **#nivell#**} #speak# **#thyre# #up#** #function# .
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# **#value# #up# tsh** #suggest# **#diagnost# #small# #thyre# #primar#** , **#nivell# #suppress# tsh** #suggest# **#up# #thyre#** .

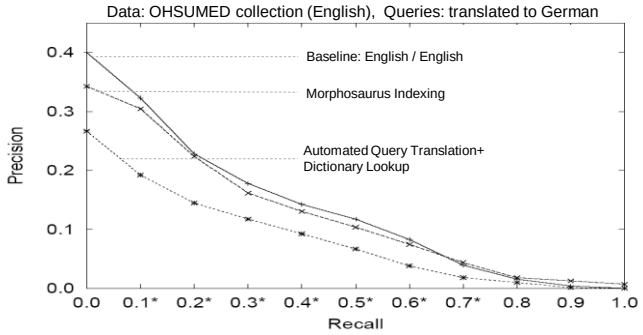
Pragmatics of lexicon building and maintenance

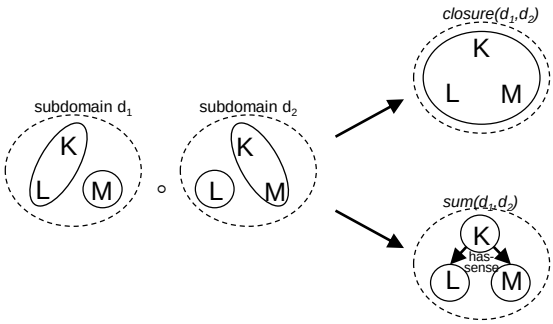
- Delimitation of subwords
  Generation of raw list of morphemes by automated affix stripping.
  - Morpheme candidates are eliminated when utterly short and ccurring as accidental substrings (causing parsing errors), e.g. *ov-*, *gen-*
  - Morpheme combinations are added when composed form has a non-compositional sense, e.g. *bauch|speichel|drüs- de|cubit-*, *neur|os-*
  - Delimitation decisions driven by performance function:
    Precoding of suffix combinations, e.g. *–ibilities*, *-alitäten*
    Prevention of known segmentation errors:
    nephrotomy -> nephr-oto-my (correct: *nephr-o-tomy*)
    addition of *-otomy* solves the problem.

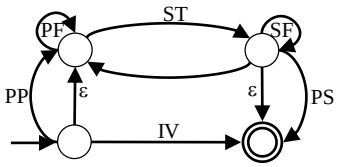
- Grouping of lexemes
  - Creation of equivalence (synonym, translation) classes by incremental fusion of MIDs.
  - Tradeoff : fusion of senses (problem of big equivalence classes with unspecific senses) vs. explosion of ambiguous readings



Evaluation in IR setting




closure(d1,d2)
K
L
M
subdomain d1
K
L
M
∘
subdomain d2
K
L
M
sum(d1,d2)
K
has-sense
L
M


PF
ST
SF
PP
ε
ε
PS
IV

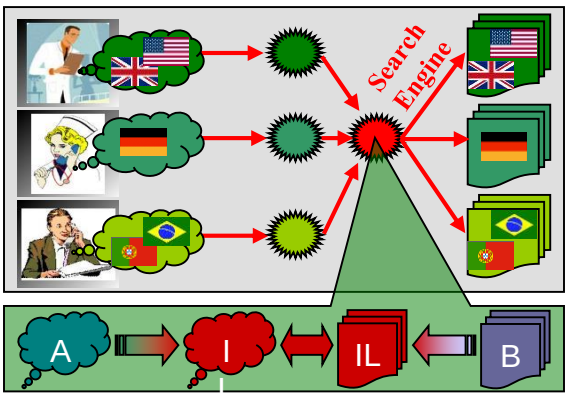
Semantic Atomicity and Multilinguality in the Medical Domain: Design Considerations for the **MorphoSaurus** Subword Lexicon

S. Schulz, K. Markó, P. Daumke, U. Hahn, S. Hanser, P. Nohama, R. L. de Andrade, E. Pacheco, M. Romacker

Medical Informatics, Freiburg University Hospital, Freiburg, Germany, Health Informatics Laboratory, Paraná Catholic University, Curitiba, Brazil, Jena University Language & Information Engineering (JULIE) Lab, Jena, Germany, Text Mining in Life Sciences Informatics, Novartis, Basel, Iexitzerland

Medical document collections are very large, dynamic, multi-lingual, multi-genre and used by a heterogeneous user community.

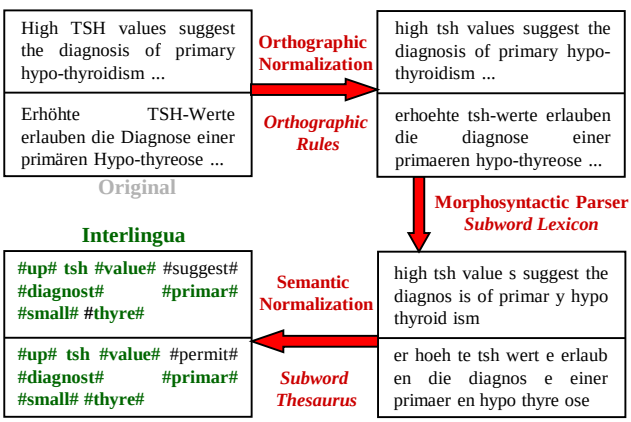
We respond to these challenges for **medical information retrieval** in terms of the **MorphoSaurus system** which is based upon using an **interlingua** representation of both queries and documents.



Interlingual representation: Queries from language A as well as documents from language B are both translated into a language-independent interlingua (IL) on which matching procedures apply.

The Morphosaurus system uses a special type of **dictionary**, with entries consisting of **subwords**, i.e., semantically minimal units. Subwords are grouped into **equivalence classes** which capture **intralingual** as well as **interlingual** synonymy.

A **morphosyntactic parser** extracts subwords and assigns equivalence class identifiers.



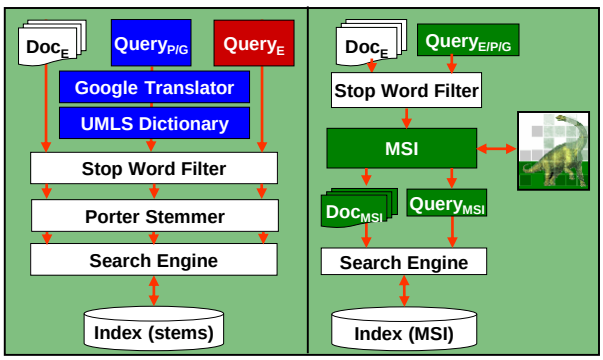
Interlingual **Morpho-semantic Normalization** is achieved by a three-step procedure: orthographic normalization, morphological segmentation and semantic normalization.

Evaluation with the **OHSUMED Corpus** (~233,000 English documents, 106 English queries – translated to German and Portuguese by medical experts)

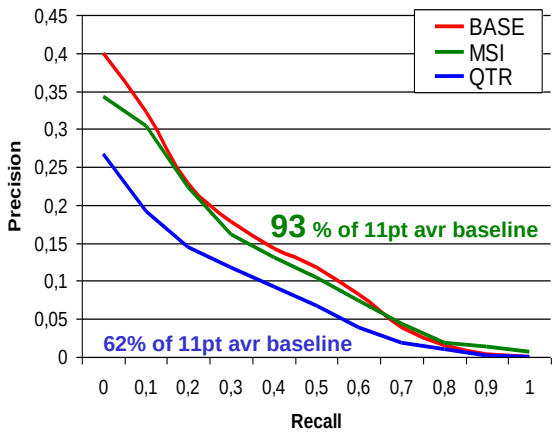
Baseline: monolingual retrieval, $Query_E \leftrightarrow Doc_E$

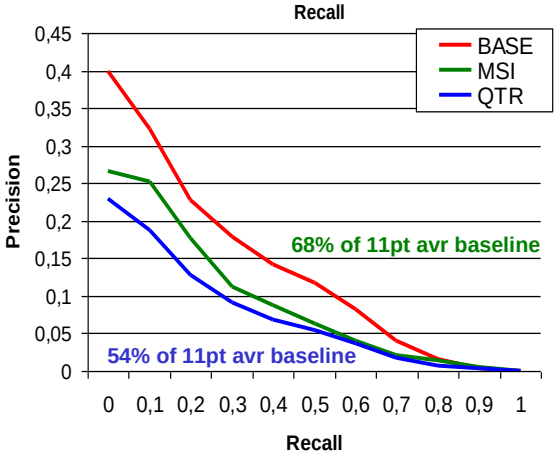
QTR: Query translation - GOOGLE translator & bilingual UMLS dictionary

MSI: Morphosaurus - morpho-semantically indexed queries and documents



Evaluation scenarios: Baseline (left), **query translation** (left), **morpho-semantic indexing (MSI)** (right)





Contact: Kornél Markó, Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany, http://www.coling.uni-freiburg.de/~marko, E-mail: marko@coling.uni-freiburg.de