AMIA 2006: Biomedical and Health Informatics
November 11-15, 2006 Hilton Washington & Towers Washington, DC

Towards a Multilingual Medical Lexicon

Kornél Markó[1], Robert Baud[2], Pierre Zweigenbaum[3], Magnus Merkel[4], Lars Borin[5], Stefan Schulz[1]

[1] Freiburg University Hospital, Department of Medical Informatics, Germany
[2] University Hospitals of Geneva, Service of Medical Informatics, Switzerland
[3] Inserm, U729; Assistance Publique -- Paris Hospitals, STIM; Inalco, CRIM, France
[4] Linköping University, Department of Computer and Information Science, Sweden
[5] Göteborg University, NLP Section, Sweden

Information Society Technologies

Semantic Mining
—— *Semantic Interoperability and Data Mining in Biomedicine*

What matters for a multilingual dictionary ?

- Entries: Base forms + all lexical variants
- Morpho-syntactic information
  - POS, number, gender, case, tense, …
- Coverage
  - With respect to a given domain
- Multilingualism
  - Translation dictionaries

Available Sources

	Lexical Information	Coverage	Multilingualism
WordNet	POS	•155,000 words •General language	English
EuroWordNet	POS	•30,000-50,000 words •General language	Dutch, Italian, Spanish, German, French, Czech, Estonian, …
EuroDicAutum		•630,000 words and phrases •EU activities, finance, agriculture, legislation, transport,…	Dutch, French, German, Italian, Danish, English, Greek, Portuguese, Spanish, Finnish, Swedish
UMLS Metathesaurus	-	• > 1M words and phrases •Biomedicine	English, German, French, Spanish, French, Swedish, Russian, …
UMLS Specialist Lexicon	POS, number, case, tense, person, inflection types, …	•250,000 words and acronyms •Biomedicine	English

Multilingual Lexicon Creation

1. Create/Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Multilingual Lexicon Creation

1. Create/Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Collect Monolingual Lexical Resources with full lexical information

- French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries)
- English medical lexicon from Linköping University, Sweden (22,686 entries)
- Swedish medical lexicon from Linköping University (23,223 entries)
- Swedish medical lexicon from Göteborg University, Sweden (6,786 entries)
- German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries)
- English Specialist Lexicon, which is part of the UMLS (96,621 entries, avoiding acronyms and chemical names)
- A total of 224,351 entries

Collect Monolingual Lexical Resources with full lexical information

- French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries)
- English medical lexicon from Linköping University, Sweden (22,686 entries)
- Swedish medical lexicon from Linköping University (23,223 entries)
- Swedish medical lexicon from Göteborg University, Sweden (6,786 entries)
- German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries)
- English Specialist Lexicon, which is part of the UMLS (96,621 entries, avoiding acronyms and chemical names)
- A total of 224,351 entries

Large variability in scope and granularity

Multilingual Lexicon Creation

- Collect available (monolingual) resources
- Define an interchange format
- Convert resources to that format
- Define a cross-lingual linking format
- Define criteria for semantic linkage
- Semantically align lexical entries

Lexicon Interchange Format

(cf. Baud et al. AMIA 05)

Field	Description	Definition
ID	Unique Identifier	
Frm	**Inflected Form**	Inflection of Lem
Mfr	Morphosyntactic features of inflected form	Coded in MULTEXT
Lem	Lemma	The basic form of the entry
Mul	Morpho-syntactic features	Coded in MULTEXT
Lng	Language	EN,DE,FR,SW
Typ	Entry Type	•Basic entry (B) •Subword entry (S) •Compound entry (C) •Term entry (T)
Prt	Decomposition	Parts of a compound entry
Str	Head	Head of a compound/term entry
Ref	Reference lemma	ID of lemma‘s entry

MULTEXT Standard

Part-of-Speech

Noun	N
Verb	V
Adjective	A
Adverb	R
…	

Nouns (N)

P	ATT	VAL	C
1	Type	common	c
		proper	p
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
4	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Verbs (V)

P	ATT	VAL	C
1	Type	main	m
		auxiliary	a
		modal	o
2	Mood/VForm	indicative	i
		subjunctive	s
		imperative	m
		conditional	c
		infinitive	n
		participle	p
		gerund	g
		supine	s
		base	b
3	Tense	present	p
		imperfect	i
		future	f
		past	s
4	Person	first	1
		second	2
		third	3
5	Number	singular	s
		plural	p
6	Gender	masculine	m
		feminine	f
		neuter	n

Adjectives (A)

P	ATT	VAL	C
1		indefinite	i
		possessive	s
2	Degree	positive	p
		comparative	c
		superlative	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Articles (T)

P	ATT	VAL	C
1	Type	definite	d
		indefinite	i
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p

MULTEXT Standard

Part-of-Speech

Noun	N
Verb	V
Adjective	A
Adverb	R
…	

Nouns (N)

P	ATT	VAL	C
1	Type	common	c
		proper	p
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
4	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Lemma	MULTEXT
nail (EN)	Nc-sn
doigt (FR)	Ncmsn
digital (SW)	A-pfsn

Verbs (V)

P	ATT	VAL	C
1	Type	main	m
		auxiliary	a
		modal	o
2	Mood/VForm	indicative	i
		subjunctive	s
		imperative	m
		conditional	c
		infinitive	n
		participle	p
		gerund	g
		supine	s
		base	b
3	Tense	present	p
		imperfect	i
		future	f
		past	s
4	Person	first	1
		second	2
		third	3
5	Number	singular	s
		plural	p
6	Gender	masculine	m
		feminine	f
		neuter	n

Adjectives (A)

P	ATT	VAL	C
1		indefinite	i
		possessive	s
2	Degree	positive	p
		comparative	c
		superlative	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a

Articles (T)

P	ATT	VAL	C
1	Type	definite	d
		indefinite	i
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p

Multilingual Lexicon Creation

1. Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Monolingual Resources

```
EN|DIM:20501|B||abdomen|Ncns|||||||||
EN|DIM:20502|B||abdominal|Afpns|||||||||
EN|DIM:20503|B||abduction|Ncns|||||||||
EN|DIM:20504|B||abductor|Ncns|||||||||

SV|LIU_SV142_A|B||buk|nc0sn||||||||||
SV|LIU_SV143_A|B||abdominal|afpusn||||||||||
SV|LIU_SV144_A|T||från buk|nc0sn||||||buk||||
SV|LIU_SV145_A|C||bukabscess|nc0sn|||||buk__abscess|||||
SV|LIU_SV146_A|T||abdominell aktinomykos|nc0sn||||||aktinomykos||||
SV|LIU_SV147_A|C||bukaorta|nc0sn|||||buk__aorta|||||

DE|UKLFR:1045|B||Atrophie|Ncfsn|||19||||||
DE|UKLFR:1046|B||Atropin|Ncnsn|||11||||||
DE|UKLFR:1047|B||Atropinvergiftung|Ncfsn|||20||||||
DE|UKLFR:1048|B||Attacke|Ncfsn|||19||||||
DE|UKLFR:1049|B||Attest|Ncnsn|||11||||||

FR|UMLF:10012|B||distiller|Vmn|||||||||
FR|UMLF:10013|B||distinct|Afms|||||||||
FR|UMLF:10014|B||distinguer|Vmn|||||||||
FR|UMLF:10015|C||distocclusion|Ncfs|||||dist__occlusion|occlusion|||
```

POS: Common Noun

```
EN|DIM:20501|B||abdomen|Ncns|||||||||
EN|DIM:20502|B||abdominal|Afpns|||||||||
EN|DIM:20503|B||abduction|Ncns|||||||||
EN|DIM:20504|B||abductor|Ncns|||||||||

SV|LIU_SV142_A|B||buk|nc0sn||||||||||
SV|LIU_SV143_A|B||abdominal|afpusn||||||||||
SV|LIU_SV144_A|T||från buk|nc0sn||||||buk||||
SV|LIU_SV145_A|C||bukabscess|nc0sn|||||buk__abscess|||||
SV|LIU_SV146_A|T||abdominell aktinomykos|nc0sn||||||aktinomykos||||
SV|LIU_SV147_A|C||bukaorta|nc0sn|||||buk__aorta|||||

DE|UKLFR:1045|B||Atrophie|Ncfsn|||19||||||
DE|UKLFR:1046|B||Atropin|Ncnsn|||11||||||
DE|UKLFR:1047|B||Atropinvergiftung|Ncfsn|||20||||||
DE|UKLFR:1048|B||Attacke|Ncfsn|||19||||||
DE|UKLFR:1049|B||Attest|Ncnsn|||11||||||

FR|UMLF:10012|B||distiller|Vmn|||||||||
FR|UMLF:10013|B||distinct|Afms|||||||||
FR|UMLF:10014|B||distinguer|Vmn|||||||||
FR|UMLF:10015|C||distocclusion|Ncfs|||||dist__occlusion|occlusion|||
```

POS: Adjective

```
EN|DIM:20501|B||abdomen|Ncns|||||||||
EN|DIM:20502|B||abdominal|Afpns|||||||||
EN|DIM:20503|B||abduction|Ncns|||||||||
EN|DIM:20504|B||abductor|Ncns|||||||||

SV|LIU_SV142_A|B||buk|nc0sn||||||||||
SV|LIU_SV143_A|B||abdominal|afpusn||||||||||
SV|LIU_SV144_A|T||från buk|nc0sn||||||buk||||
SV|LIU_SV145_A|C||bukabscess|nc0sn|||||buk__abscess|||||
SV|LIU_SV146_A|T||abdominell aktinomykos|nc0sn||||||aktinomykos||||
SV|LIU_SV147_A|C||bukaorta|nc0sn|||||buk__aorta|||||

DE|UKLFR:1045|B||Atrophie|Ncfsn|||19||||||
DE|UKLFR:1046|B||Atropin|Ncnsn|||11||||||
DE|UKLFR:1047|B||Atropinvergiftung|Ncfsn|||20||||||
DE|UKLFR:1048|B||Attacke|Ncfsn|||19||||||
DE|UKLFR:1049|B||Attest|Ncnsn|||11||||||

FR|UMLF:10012|B||distiller|Vmn|||||||||
FR|UMLF:10013|B||distinct|Afms|||||||||
FR|UMLF:10014|B||distinguer|Vmn|||||||||
FR|UMLF:10015|C||distocclusion|Ncfs|||||dist__occlusion|occlusion|||
```

POS: Compound

```
EN|DIM:20501|B||abdomen|Ncns|||||||||
EN|DIM:20502|B||abdominal|Afpns|||||||||
EN|DIM:20503|B||abduction|Ncns|||||||||
EN|DIM:20504|B||abductor|Ncns|||||||||

SV|LIU_SV142_A|B||buk|nc0sn||||||||||
SV|LIU_SV143_A|B||abdominal|afpusn||||||||||
SV|LIU_SV144_A|T||från buk|nc0sn||||||buk||||
SV|LIU_SV145_A|C||bukabscess|nc0sn|||||buk__abscess|||||
SV|LIU_SV146_A|T||abdominell aktinomykos|nc0sn||||||aktinomykos||||
SV|LIU_SV147_A|C||bukaorta|nc0sn|||||buk__aorta|||||

DE|UKLFR:1045|B||Atrophie|Ncfsn|||19||||||
DE|UKLFR:1046|B||Atropin|Ncnsn|||11||||||
DE|UKLFR:1047|B||Atropinvergiftung|Ncfsn|||20||||||
DE|UKLFR:1048|B||Attacke|Ncfsn|||19||||||
DE|UKLFR:1049|B||Attest|Ncnsn|||11||||||

FR|UMLF:10012|B||distiller|Vmn|||||||||
FR|UMLF:10013|B||distinct|Afms|||||||||
FR|UMLF:10014|B||distinguer|Vmn|||||||||
FR|UMLF:10015| C||distocclusion|Ncfs|||||dist__occlusion|occlusion|||
```

Multi Word Nouns

```
EN|DIM:20501|B||abdomen|Ncns|||||||||
EN|DIM:20502|B||abdominal|Afpns|||||||||
EN|DIM:20503|B||abduction|Ncns|||||||||
EN|DIM:20504|B||abductor|Ncns|||||||||

SV|LIU_SV142_A|B||buk|nc0sn||||||||||
SV|LIU_SV143_A|B||abdominal|afpusn||||||||||
SV|LIU_SV144_A|T||från buk|nc0sn||||||buk||||
SV|LIU_SV145_A|C||bukabscess|nc0sn|||||buk__abscess|||||
SV|LIU_SV146_A|T||abdominell aktinomykos|nc0sn||||||aktinomykos||||
SV|LIU_SV147_A|C||bukaorta|nc0sn|||||buk__aorta|||||

DE|UKLFR:1045|B||Atrophie|Ncfsn|||19||||||
DE|UKLFR:1046|B||Atropin|Ncnsn|||11||||||
DE|UKLFR:1047|B||Atropinvergiftung|Ncfsn|||20||||||
DE|UKLFR:1048|B||Attacke|Ncfsn|||19||||||
DE|UKLFR:1049|B||Attest|Ncnsn|||11||||||

FR|UMLF:10012|B||distiller|Vmn|||||||||
FR|UMLF:10013|B||distinct|Afms|||||||||
FR|UMLF:10014|B||distinguer|Vmn|||||||||
FR|UMLF:10015|C||distocclusion|Ncfs|||||dist__occlusion|occlusion|||
```

Multilingual Lexicon Creation

1. Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Linking Format

Links represent possible translations.

Field	Description	Definition
Src	Source	Source ID to be linked with target entry
Tar	Target	Target ID to be linked with source entry
Lnk	Type of Relation	???

Relation Type:

- Language-specific characteristics of
  - number,
  - case,
  - gender
- Multiple derivations, e.g.
  - attributive or predicative adjectives,
  - definite or indefinite objects

Multilingual Lexicon Creation

1. Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Simple Relation Types

- REL1
  - A and B share the same POS and MULTEXT features
- REL2
  - A and B share the same POS, but at least one MULTEXT feature differs
- REL3
  - A and B do not share same POS

Multilingual Lexicon Creation

1. Collect available (monolingual) resources
2. Define an interchange format
3. Convert resources to interchange format
4. Define a cross-lingual linking format
5. Define criteria for semantic linkage
6. Semantically align lexical entries

Linking using Subword Indexing

- MorphoSaurus engine extracts meaningful subwords from medical text
- Maps them to language-independent concept IDs
- Covers English, German, Portuguese, French, Spanish, Swedish
- Validated for cross-lingual text retrieval

Original Document	Orthographic Normalization	Morphological Segmentation	Semantic Normalization (MIDs)
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#** .

Semantic Enrichment

- All inflected forms (*Frm-field*, if available) or base forms (*Lem-field*) are processed with MorphoSaurus.
- Resulting representations are added to lexicon entries:

  – …
  – EN|DIM:20501|B||abdomen|Ncns|||||||||| #abdom
  – DE|UKLFR:1383|B||Bauch|Ncmsn|||1u||||||| #abdom
  – …

Linking Algorithm

FOR EVERY lexeme *i* and its attributes in the list DO
  FOR EVERY lexeme *j* (starting from *i*) and its attributes in the list DO
    IF MorphoSaurus-Representation of input is identical
    THEN
      IF POS and MUL-values of *i* and *j* are identical
        THEN *type* = REL1 # synonymy/translation
      IF POS-values are identical, but not MUL-values
        THEN *type* = REL2
      IF POS-values differ
        THEN *type* = REL3 # derivation
      PRINT "ID(*i*) | ID(*j*) | *type*"

Results

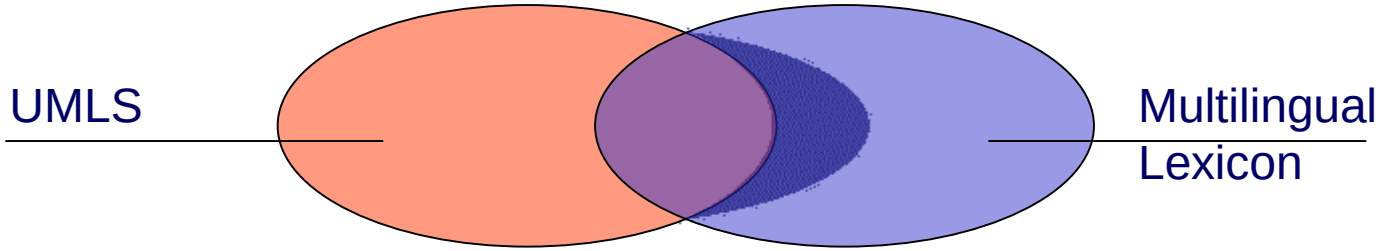
Language Pair	# Relations	Different Lemmas
English-German	126,504	31,544
English-French	70,680	24,368
English-Swedish	86,655	34,030
French-Swedish	21,604	8,312
French-German	32,659	10,458
German-Swedish	41,469	12,105
TOTAL	379,571	120,817

- Additional 271,971 intralingual synonymy relationships found (sums up to 651,542)
- REL1 (10%), REL2 (44%), REL3 (46%)

Preliminary Evaluation: Coverage

- Most comprehensive resource for medical terminology: UMLS Metathesaurus
- How many terms are covered?
- How many relations are covered?

Coverage: Terms



Language	UMLS *	Covered	Synonyms**	Additional**
English	122,035	32,668 (27%)	3,807	68,842
German	21,162	2,832 (13%)	1,269	23,379
French	10,260	3,590 (35%)	309	25,923
Swedish	12,012	8,520 (71%)	994	17,579
TOTAL	165,469	189,712		

* Only single-word preferred entries ** only REL1 and REL2 (no derivation)

Coverage: Relations

Language	UMLS	Covered	Synonyms*	Additional*
English-German	15,979	1,259 (8%)	8,801	21,484
English-French	12,589	1,783 (14%)	6,974	15,611
English-Swedish	9,554	3,403 (36%)	10,124	20,503
German-French	9,859	850 (9%)	773	8,835
German-Swedish	10,063	810 (8%)	1,699	9,596
French-Swedish	6,793	1,109 (16%)	1,911	5,292
TOTAL	64,837	120,817		

* only REL1 and REL2

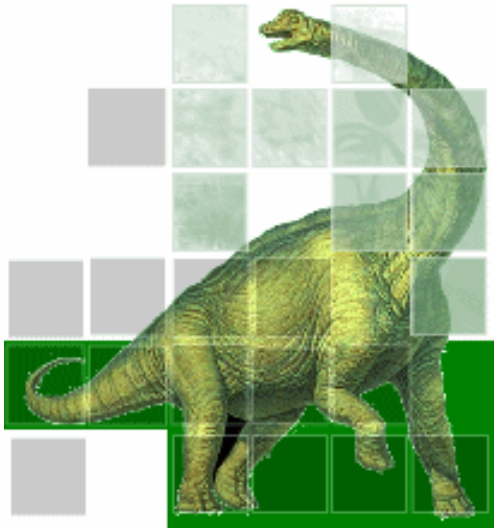
Conclusion

- Framework for integrating heterogenous lexical resources
- Integration of English, German, French and Swedish sources
- Simple Linkage Format for coding lexical relations
- Generating substantial amount of synonym mappings and translations using MorphoSaurus subword indexing
- Partly validated using UMLS Metathesaurus
- Extensive evaluation is ongoing (presentation at MEDINFO)

www.morphosaurus.net

Morphosaurus Resources

- Subword-Lexicon:
  - Organizes subwords in several languages (English, German, Portuguese, Spanish, Fench, Swedish)
- Subword-Thesaurus:
  - Groups synonymous subwords (within and between languages)
- Subword-Segmeter:
  - Extraction of Subwords and Assignment of *Equivalence Classes*



Morphosaurus

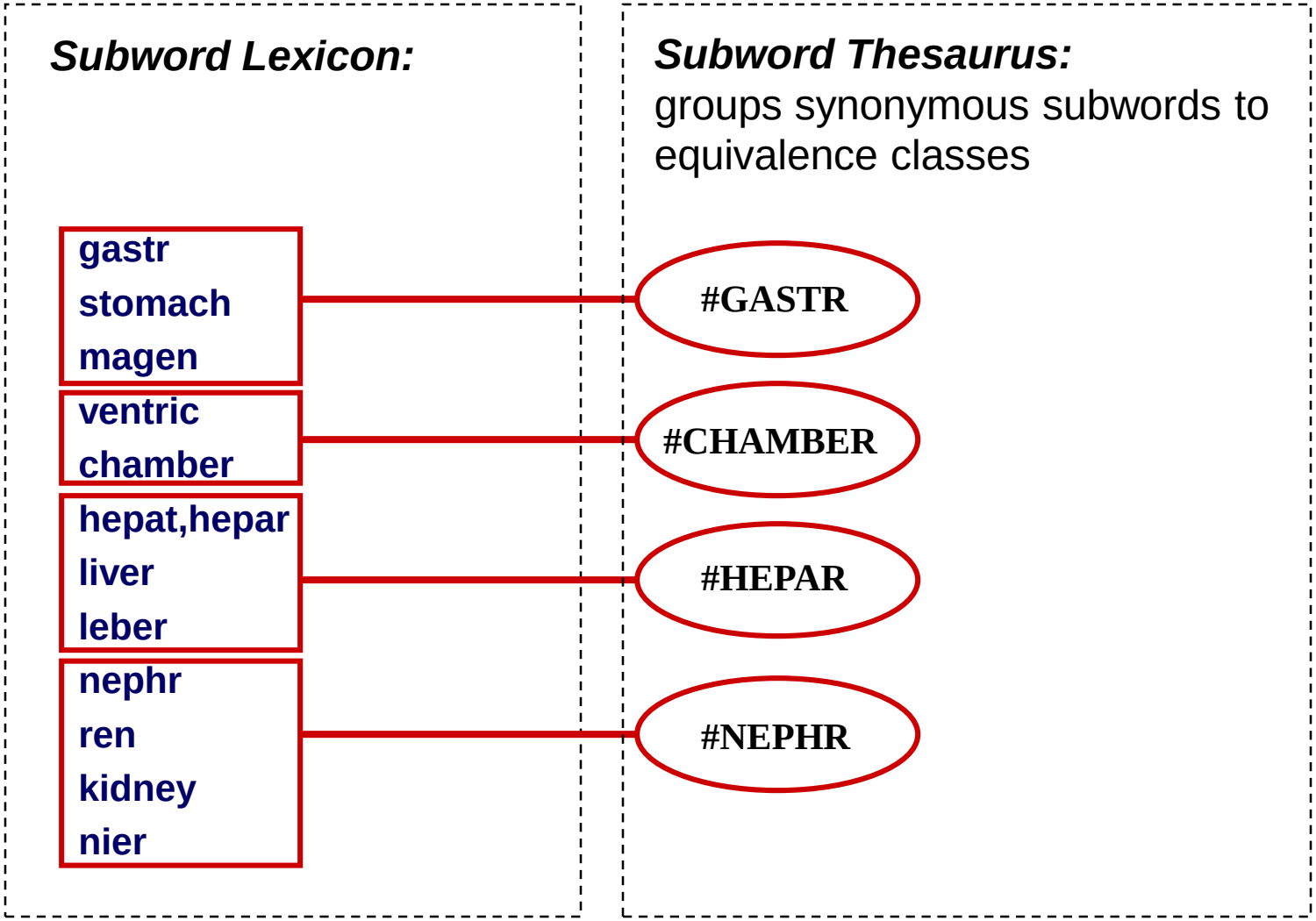
→ **Morphosaurus-Identifier (MID)**

Subword Approach

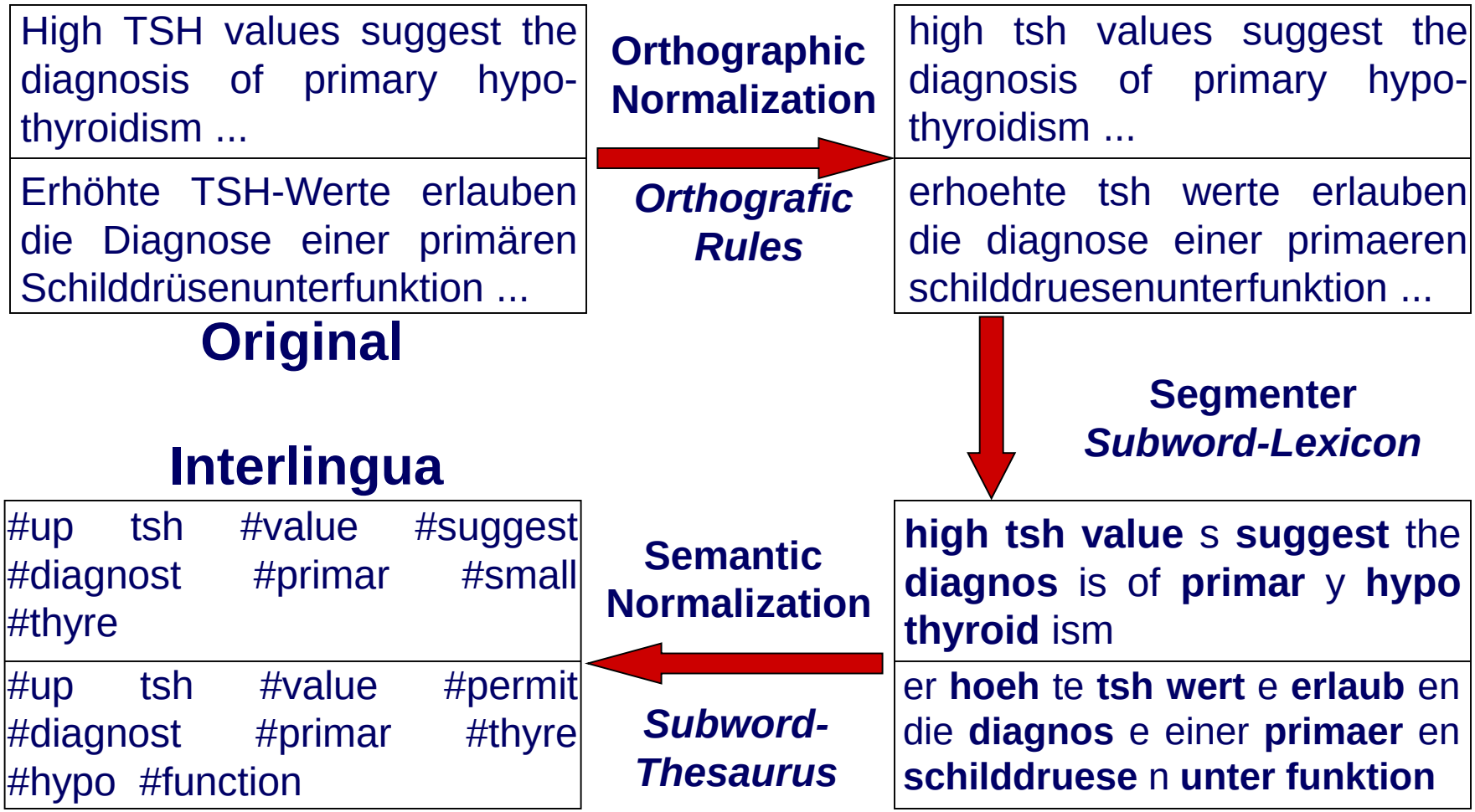
- Subwords are atomic, conceptual *or* linguistic units:
  - Stems: *stomach*, *gastr*, *diaphys*
  - Prefixes: *anti-, bi-, hyper-*
  - Suffixes: *-ary, -ion, -itis*
  - Infixes: *-o-, -s-*
- Equivalence classes contain synonymous subwords and their translations:

#derma = { derm, cutis, skin, haut, kutis, pele, cutis, piel, … }
#inflamm = { inflam, -itic, -itis, entzuend, -itis, -itisch, inflam, flog, inflam, flog, -iolitis, ... }

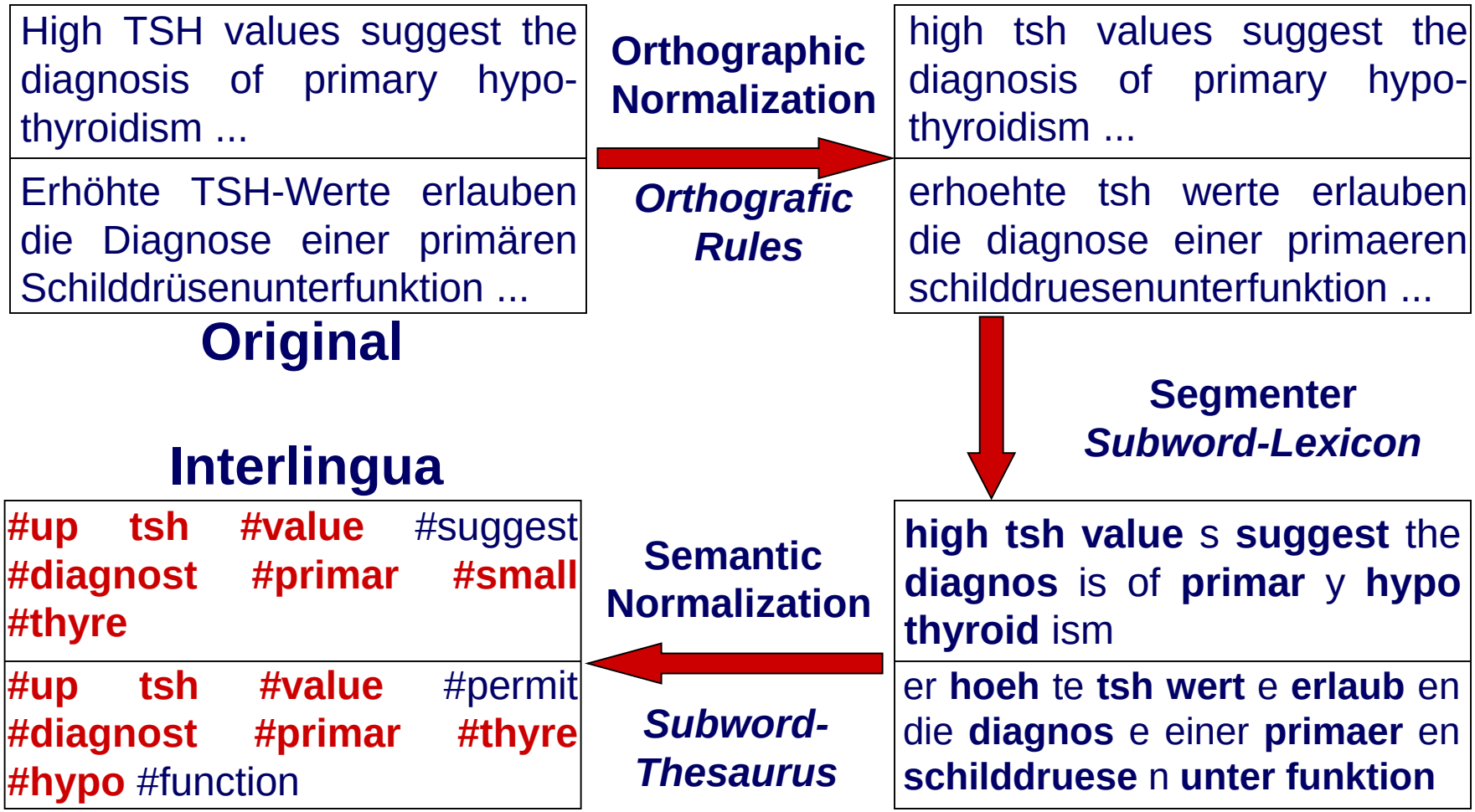
Subwords: Lexicon & Thesaurus



Example


High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...
Original
Orthographic Normalization
Orthografic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...
Segmenter
Subword-Lexicon
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en schilddruese n unter funktion
Semantic Normalization
Subword-Thesaurus
Interlingua
#up tsh #value #suggest #diagnost #primar #small #thyre
#up tsh #value #permit #diagnost #primar #thyre #hypo #function

Example


High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...
Original
Orthographic Normalization
Orthografic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...
Segmenter
Subword-Lexicon
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en schilddruese n unter funktion
Semantic Normalization
Subword-Thesaurus
Interlingua
#up tsh #value #suggest #diagnost #primar #small #thyre
#up tsh #value #permit #diagnost #primar #thyre #hypo #function