

The German Specialist Lexicon

- University of Freiburg
 - Gesa Weske-Heck
 - Susanne Hanser
 - Albrecht Zaiss
 - Stefan Schulz
 - Rüdiger Klar
- University of Frankfurt
 - Wolfgang Giere
- DIMDI Cologne
 - Michael Schopen

Aims of the Project

- German Specialist Lexicon similar to the UMLS „Specialist Lexicon“
- Lexical data model covering the German biomedical terminology
- „good“ lexical coverage focusing the clinical sublanguage
- Mappings between spelling variants, synonyms and abbreviations
- Software toolkit for lexicon querying and manipulation similar to the LVG functions

German biomedical language

“Specialties”

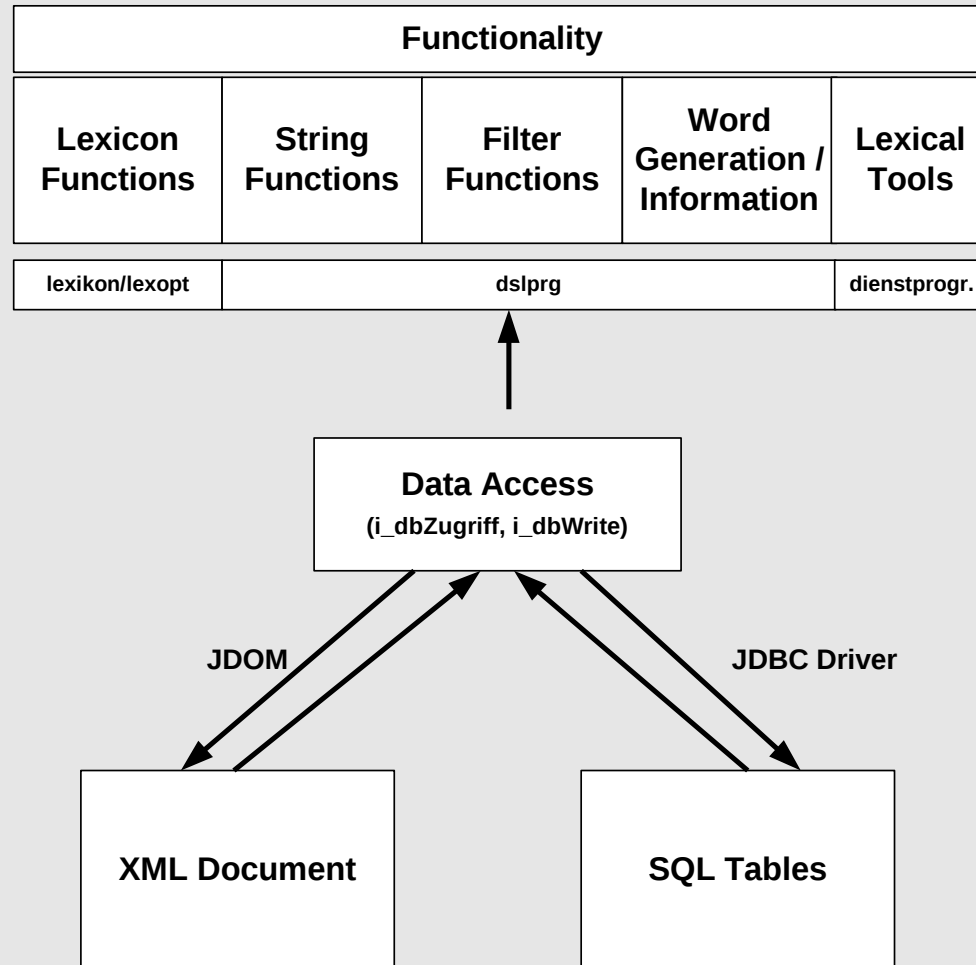
- Numerous inflection forms
 - nouns, adjectives, verbs depending on
 - gender
 - inflection classes
 - word order
 - syntactic context
- Spelling variants
 - Karzinom, Carcinom
- German/Greek/Latin synonyms
 - Nieren-, Nephro-, Ren-
- Homonyms
- Latin phrases with complete Latin inflection
 - Ulcus duodeni
- nominal compounds
 - Duodenalulcus

Suppositions of the Project

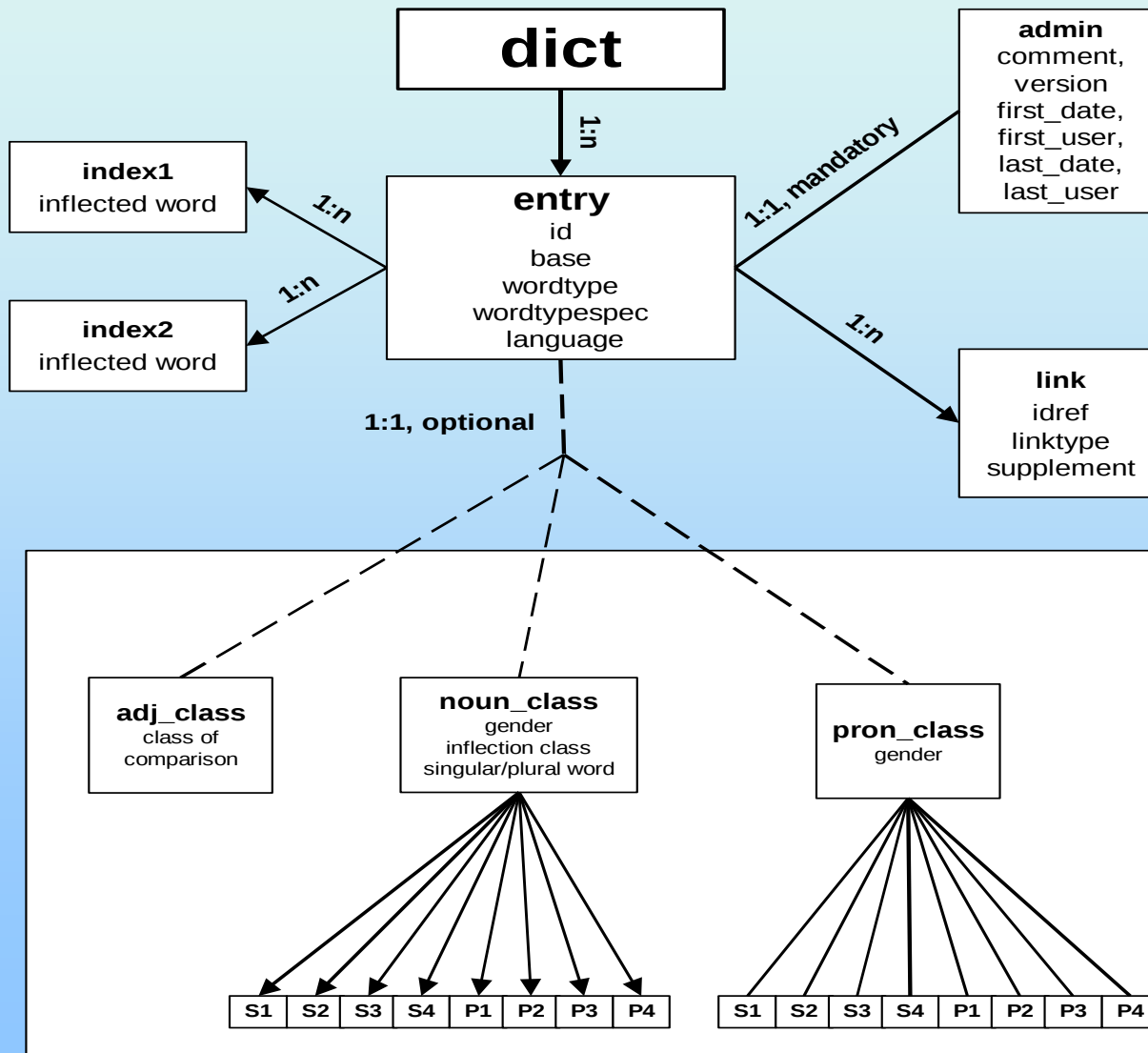
- Modelling structure and functions of the „Specialist Lexicon (UMLS)“ as close as possible
- Data storage in XML and in SQL
- JAVA-Programming

General Architecture

JAVA



XML Data Structure



XML Document Type Definition (DTD)

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!ELEMENT dict (e+)>

<!ELEMENT e (cont,(noun_class|pron_class|adj_class),link*)>
<!-- ATTLIST e -->
  id ID #REQUIRED
  language (d|lat|eng|gr|lu) #REQUIRED
  wordtype CDATA #REQUIRED
  wordtypespec CDATA #REQUIRED
  version CDATA #REQUIRED
  sg_pl (true|false|null) #IMPLIED
  casus_prep CDATA #IMPLIED
  ambig (true|false) #REQUIRED

<!-- ELEMENT cont -->
  (#PCDATA)

<!-- ELEMENT noun_class -->
  (exception*)
<!-- ATTLIST noun_class -->
  gen (m|f|n|u) #REQUIRED
  nclass CDATA #REQUIRED

<!-- ELEMENT pron_class -->
  (exception*)
<!-- ATTLIST pron_class -->
  gen (m|f|n|u) #REQUIRED

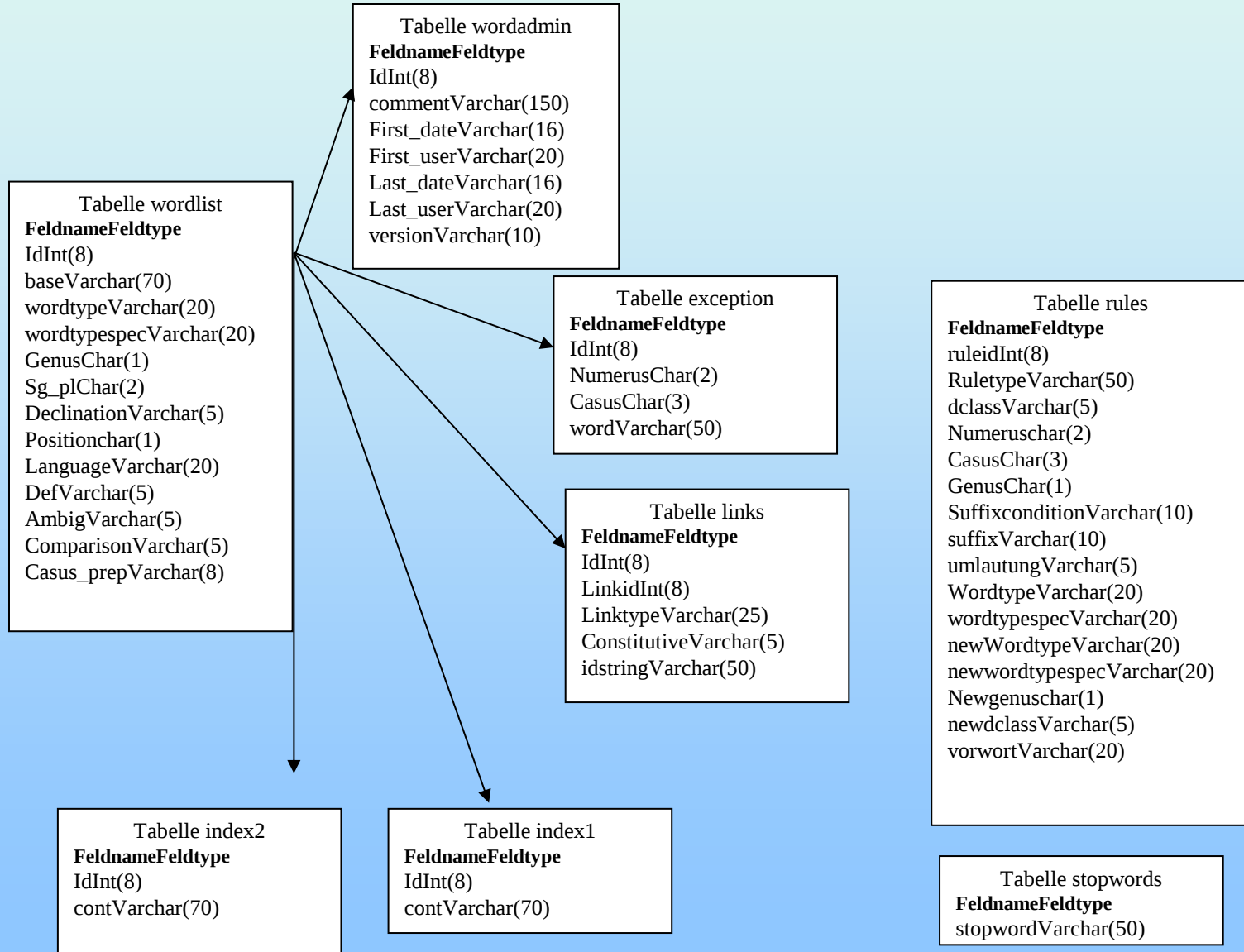
<!-- ELEMENT exception -->
  EMPTY
<!-- ATTLIST exception -->
  numerus (Sg|Pl) #REQUIRED
  casus (Nom|Dat|Gen|Akk) #REQUIRED
  word CDATA #REQUIRED

<!-- ELEMENT adj_class -->
  EMPTY
<!-- ATTLIST adj_class -->
  aclass CDATA #IMPLIED
  position (rel|abs|colour) #IMPLIED

<!-- ELEMENT link -->
  EMPTY
<!-- ATTLIST link -->
  idref IDREF #IMPLIED
  linktype CDATA #REQUIRED
  constitutive CDATA #REQUIRED
  idstring CDATA #REQUIRED
```

(Abstract) Classes	Wordtype (Part of Speech)	Wordtype Specification
Noun-Class	Noun	Noun
		Proper Name
		Acronym
		Numeral
		Adjective
		Abbreviation
		Phrase
Adj-Class	Adjective	Adjective
		Abbreviation
		Participle
		Numeral
		Proper Name
		Comparative
		Superlative
Pron-Class	Determiner	Defined Determiner
		Undefined. Determiner
	Pronoun	Reflexive Pronoun
		Relative Pronoun
		Possessive Pr. etc.
Verb	Verb	Verb
Non-Inflected	Preposition	Preposition
		Abbreviation
	Conjunction	Conjunction
		Abbreviation
	Adverb	Adverb
		Abbreviation
		Phrase
	Symbol	Arithmetic operator

SQL Table Structure



DSL Functions

INPUT

STRING

FILE

BATCHFILE

FUNCTIONS

STRING funct.

toLowerCase, czk-normalization, replaceumlaut, replaceß, sortAlphabetic, filterdoublewords etc.

FILTER funct.

filterEntries, filterNoEntries, filterWordtype, filtrateWorttypeSpec, filter Phrase

GENERATION AND INFORMATION

Print (Wordtype, wordtypespec, id, language)
Print (Spelling_variants, synonym, superlative)
Print (specialcase like "genitive, plural") etc.

tools:

add entry,
delete entry etc.

lexopt:

addstopword,
addpunct,
setIndexfile etc.

lexicon functions:

listStopwords
listWordofWordtype,
listRules,
listLinks,
listWordtype
etc.

OUTPUT

STRING

FILE

Input - Output

```
dsl -in eine große absolute Arrhythmie -out e 0 1
```

```
eine|eine|32677  
eine|eine|32758  
eine|ein|33028  
große|groß|25851  
absolute|absolut|25390  
Arrhythmie|Arrhythmie|864
```

```
dsl -konfig ktest.txt
```

Batch

```
dsl -infile test.txt -out text
```

dieser Satz abbilden drei aktive Stadien mit weil seine gebildeten Addison AHB A

```
dsl -in Haus -out 0 1 -outfile output.txt
```

Piping

Functions

dsl -in eine große absolute Arrhythmie **-fct 0 1** -out e 0 1

eine|eine|32677
eine|eine|32758
eine|ein|33028
grosse|groß|25851
absolute|absolut|25390
arrhythmie|Arrhythmie|864

-fct

0 = lowerCase

1 = replaceß

dsl -in eine große absolute Arrhythmie **-fct 14** -out 0 1

eine|32677
eine|32758
ein|33028
groß|25851
absolute Arrhythmie|32952

-fct

14 = filterPhrase

dsl -in eine große absolute Arrhythmie **-fct 12(adj|noun)** -out 0 1

groß|25851
absolut|25390
Arrhythmie|864

-fct

12 = getWordtype

Functions

dsl -lexfct 6

functionname / id for function

lowercase / 0

replaceß / 1

replaceumlaut / 2

czkNorm / 3

filtercomments / 4

filterdiacret / 5

filterpunctuation / 6

filterspecchar / 7

sortalphabetic / 8

convertCharEnc / 81

inUnicode / 82

filterDoublewords / 9

getEntries / 10

getNoEntries / 11

getWordtype / 12

getWordtypespec / 13

filterPhrase / 14

filterStopwords / 15

filterWordtype / 16

filterWordtypespec / 17

Output-Parameter

```
dsl -infile test.txt -out text
```

complete text

dieser Satz abbilden drei aktive Stadien mit weil seine gebildeten Addison AHB A

```
dsl -in eine große absolute Arrhythmie -out e 0 1
```

one line for each id

eine|eine|32677

eine|eine|32758

eine|ein|33028

große|groß|25851

absolute|absolut|25390

Arrhythmie|Arrhythmie|864

```
dsl -infile test.txt -outln e4 e2 e7 e12
```

one line for each input line

drei|Satz|mit|AHB

```
dsl -infile test.txt -outln e4.1 e2.38 e7 e12.46
```

one line for each input line

33036|Sätze|mit|Anschlussheilbehandlung/8451

with desired parameters

Output-Parameter

dsl -lexfct 5

Outputname / Id for output

baseform / 0

id / 1

language / 2

version / 3

ambig / 4

first_date / 5

first_user / 6

last_date / 7

last_user / 8

comment / 9

worttype / 10

worttypespec / 11

dclass / 12

aclass / 13

genus / 14

cas / 15

Sg/Pl / 16

S1 / 31

S2 / 32

S3 / 33

S4 / 34

P1 / 35

P2 / 36

P3 / 37

P4 / 38

comparative / 39

superlative / 40

spellingvariant / 41

synonym / 42

participlepresent / 43

participleperfect / 44

shortform / 45

longform / 46

parts / 47

adjective / 48

generatedwords / 49

inflectinginformation / 50

baseform / 51

linksfrom / 52

linksto / 53

allinks / 54

Lexicon Functions

dsl -lexfct 7

lexicalfunctionname / id

listLinks / 1

listWordtype / 2

listWordtypespec / 3

listRules / 4

listOutputfields / 5

listDSLfunction / 6

listLexfct / 7

listeDSLOpt / 8

listtools / 9

listpunctuation / 10

listHybrid / 11

listspecsion / 12

listcommentchar / 13

listspecchar / 14

listSeparator / 15

listPipechar / 16

ListCharacterSet / 17

listSCreplacement / 18

listDiacret / 19

listHybridopt / 20

listDB / 21

listDBwrite / 22

listIndexfile / 23

listminimumlength / 24

listVersion / 25

listStopwords / 26

listambigwords / 30

listwordsstartswith / 31

listwordendswith / 32

listwordsOFwordtype / 33

listwordsOFwordtypespec / 34

DSL - Tools

dsl -lexfct 9

tools

add index1

add index2

modify index

add entry

delete entry

modify entry

delete rule

add rule

delete exception

add exception

modify exception

add link

delete link

add stopword

delete stopword

xmlsql admin

xmlsql entry

xmlsql exception

xmlsql link

xmlsql rules

xmlsql stopwords

switchDB (1 = SQL, 2 = XML)

switchDBWrite

switchIndex

Options

dsl -lexfct 8

optionname / Information

addspecsign / ErgänzeSonderzeichen z.B. \$

delspecsign / LöscheSonderzeichen

adddia / ErgänzeDiakretikum

deldia / LöscheDiakretikum

addpunct / ErgänzeInterpunktion

delpunct / LöscheInterpunktion

addhyb / ErgänzeZwitter

delhyb / LöscheZwitter

addstopword / ErgänzeStopword

delstopword / LöscheStopword

addcomm / ErgänzeKommentar

delcomm / LöscheKommentar

addspecchar / ErgänzeSonderbuchstabe z.B. â

delspecchar / LöscheSonderbuchstabe

setIndexfile / SetzeAktuelleIndexdatei (1 = Standard, 2 = erweiterte Indexdatei)

setPipechar / SetzePipezeichen

setSCreplacement / Sonderzeichen werden ersatzlos gestrichen True/False

setSeparator / Setze Worttrennzeichen

setHybridropt / 0 wird belassen, 1 wird ersatzlos ersetzt, 2 wird mit Blanc ersetzt

Summary

Results

- 30.000 entries
 - most of them nouns and adjectives
 - determiners, pronouns, adverbs, conjugations, prepositions and numbers
 - infinitives of verbs as base for participles
 - taken from the ICD-10 tabular list and alphabetical index
- Lexical functions like LVG
 - command line interface
- Available from DIMDI for scientific research, no costs

Missing:

- Functions for decomposition of compound nouns
- Verbs and conjugation ...
- An user friendly graphical interface
- Connection to UMLS