metaDictionary – Towards a Generic e–Infrastructure for Detecting Variance in Language by Exploiting Dictionary Information

Dietmar Seipel and Werner Wegstein

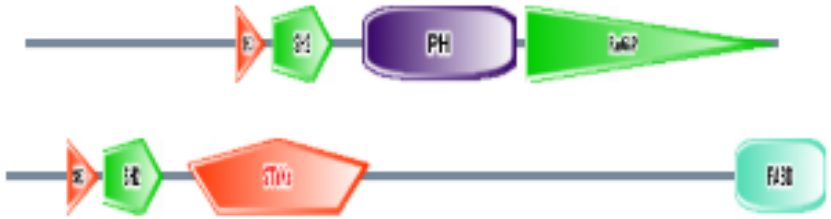
University Würzburg
Computer Science / Digital Humanities

ISGC 2011 – Taipei, 23.03.2011

Variance in Language and Genome

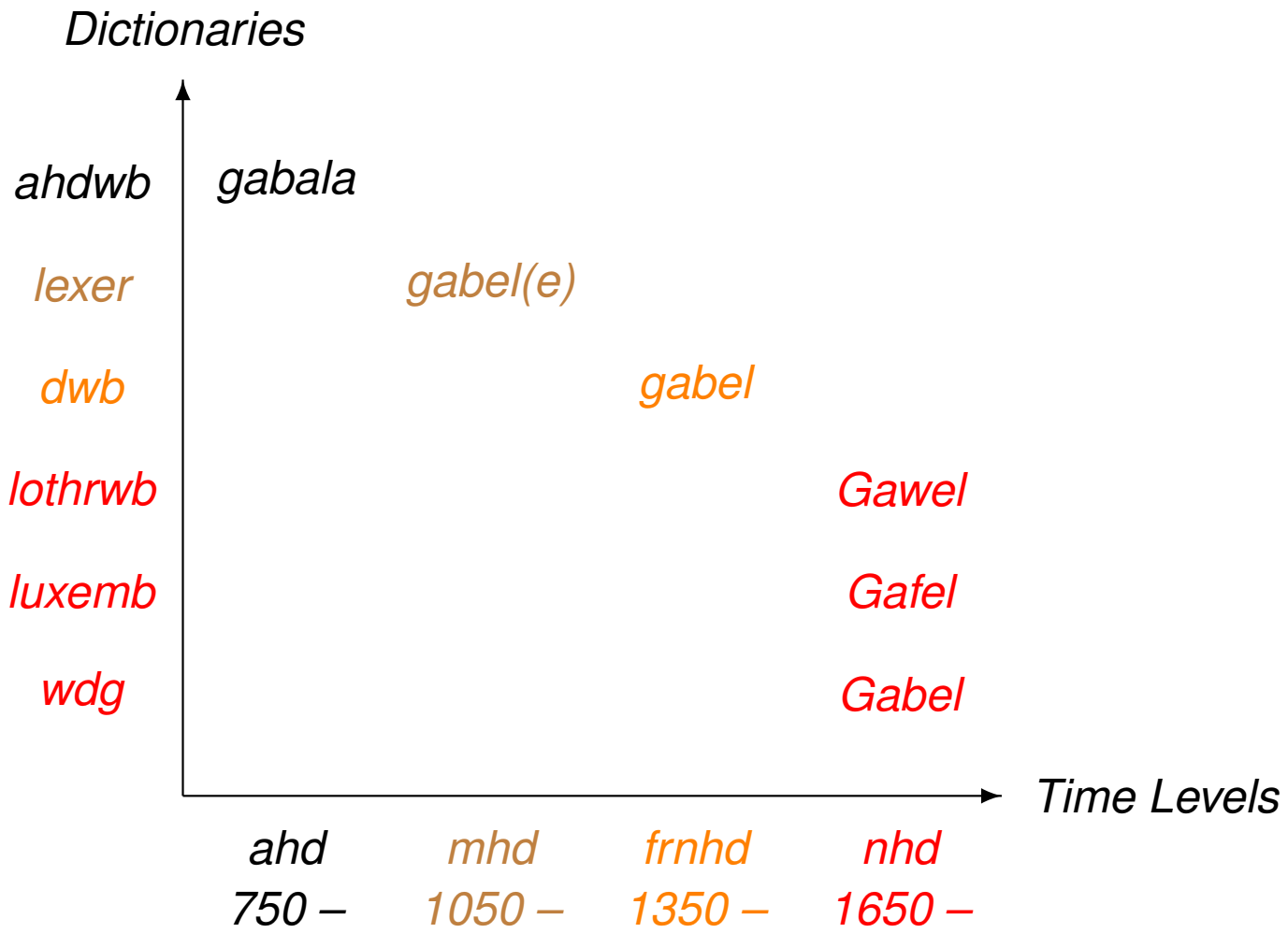
Project goals:

- development of a metaDictionary
- analysis of morpheme decomposition networks
- comparison with structural properties of genomes

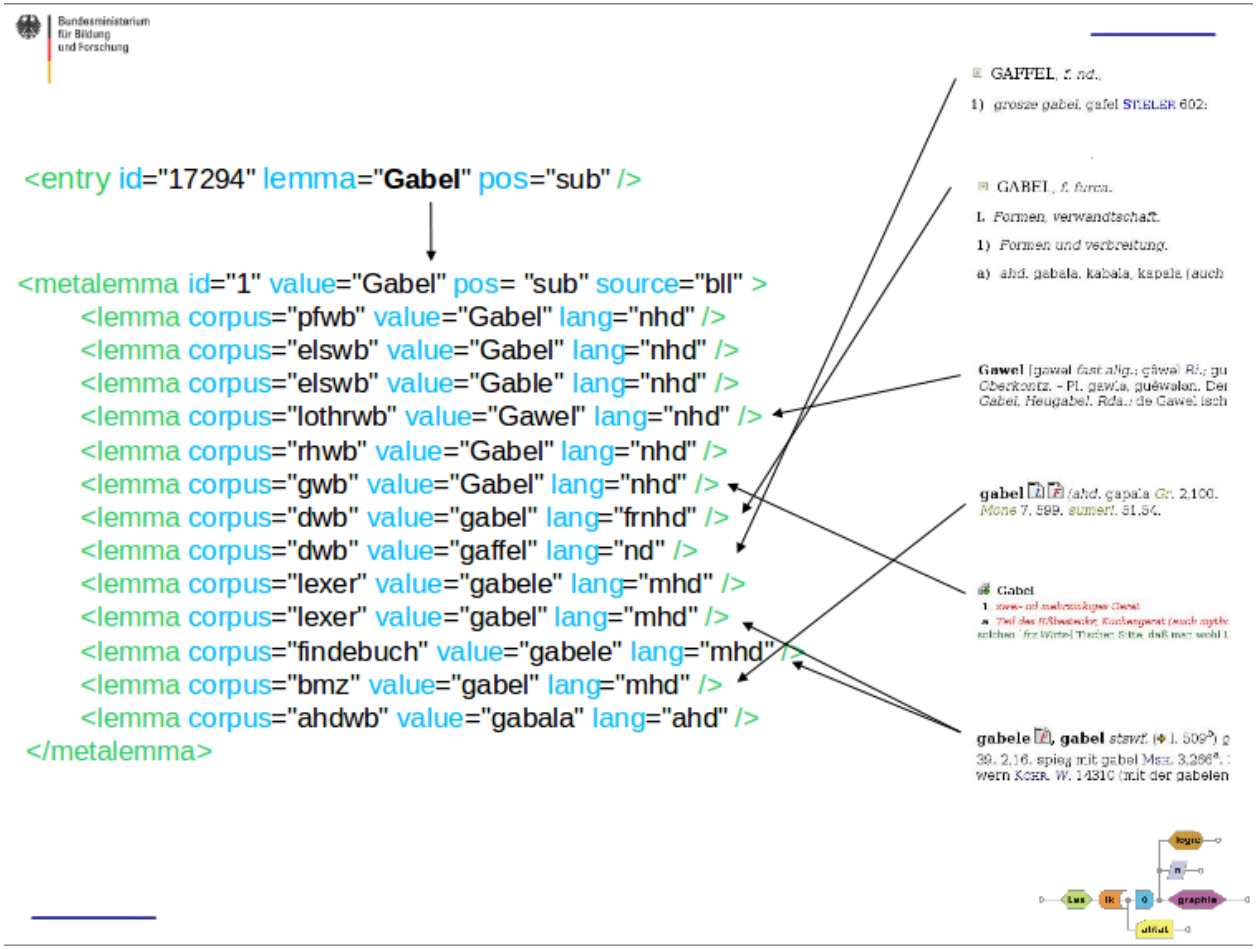


The project is funded in a BMBF framework focussing on interdependencies.

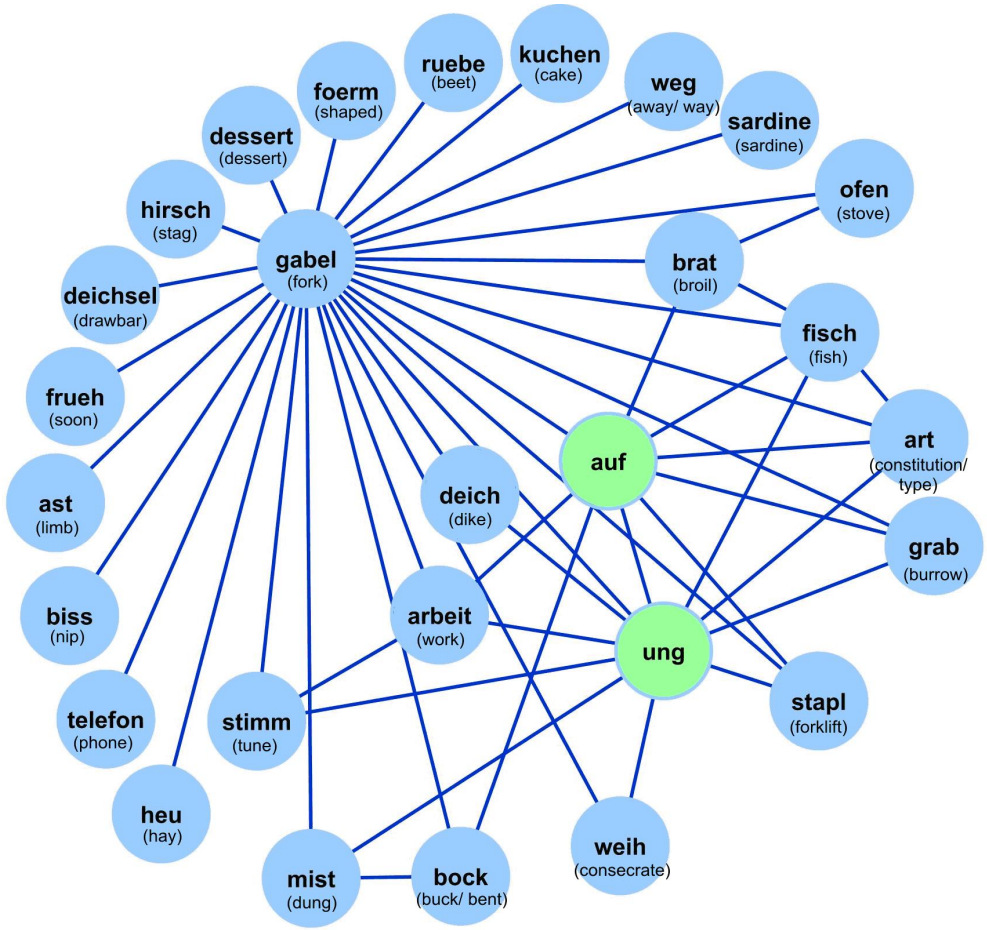
Variance in Space and Time



The metaLemma ”Gabel” (Fork)

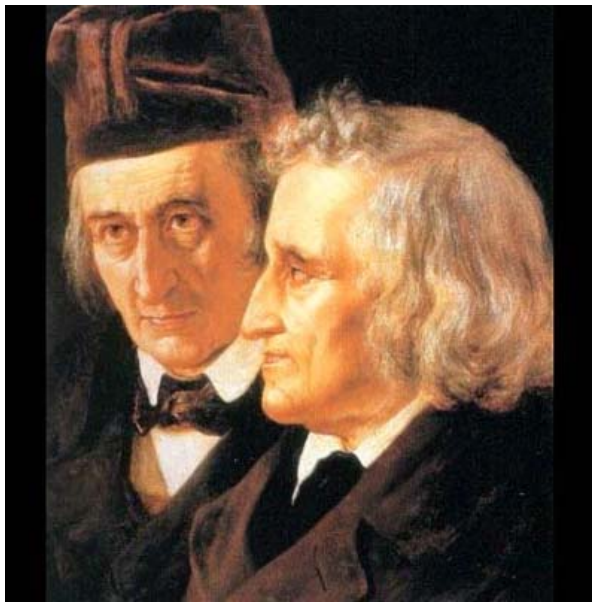


Network Analysis of Morpheme Decompositions



Network of Digitized Print Dictionaries

- German dictionaries (old to present day language including varieties like regional dialects) are annotated in TEI P5
- the fine grain annotation makes detailed additional analyses possible
- data sources:
  - Lexer
  - Grimm
  - Adelung
  - Campe
  - Luxemb., Lothr.
  - WDG



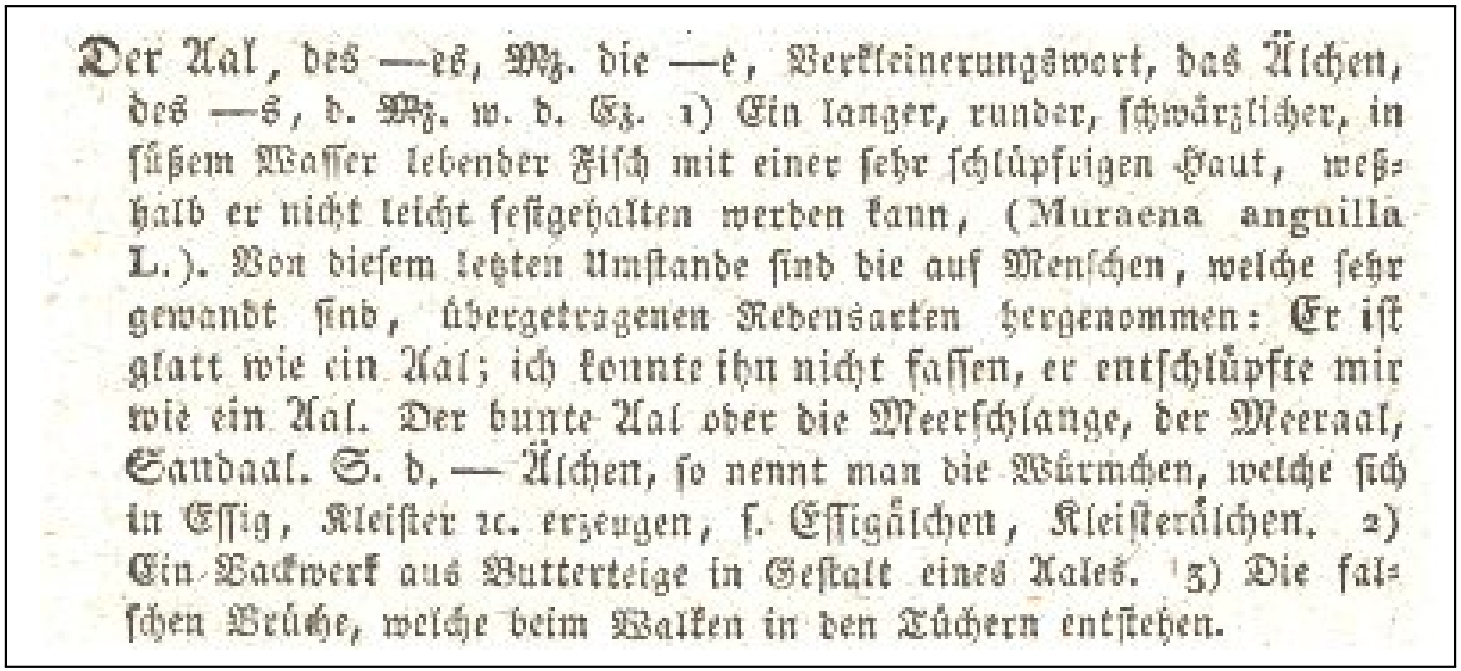
Network of Digitized Print Dictionaries – Trier

Die Wörterbücher (Mit * gekennzeichnete Wörterbücher sind externe Angebote.)

- DRW — Deutsches Rechtswörterbuch*
- DWB — Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm
- ElsWB — Elsässisches Wörterbuch
- FindeB — Findebuch zum Mittelhochdeutschen Wortschatz
- GWB — Goethe-Wörterbuch
- Adelung — Grammatisch-Kritisches Wörterbuch der Hochdeutschen Mundart
- LEI — Lessico Etimologico Italiano
- LLU — Lexikon der Luxemburger Umgangssprache*
- LWB — Luxemburger Wörterbuch*
- Meyers — Meyers Großes Konversationslexikon
- MHDBDB — Mittelhochdeutsche Begriffsdatenbank*
- Lexer — Mittelhochdeutsches Handwörterbuch von Matthias Lexer
- BMZ — Mittelhochdeutsches Wörterbuch
- MWB — Mittelhochdeutsches Wörterbuch*
- NLexer — Nachträge zum Mittelhochdeutschen Handwörterbuch von Matthias Lexer
- NRhWB — Nachträge zum Rheinischen Wörterbuch
- Kruenitz — Oekonomische Encyklopaedie von Johann Georg Krünitz*
- PfWB — Pfälzisches Wörterbuch
- RhWB — Rheinisches Wörterbuch
- LothWB — Wörterbuch der deutsch-lothringischen Mundarten
- WLM — Wörterbuch der Luxemburgischen Mundart*

Home | Impressum | Kontakt

Entry of the Adelung Dictionary



Der Aal, des —es, Mz. die —e, Verkleinerungswort, das Älchen, des —s, d. Mz. w. d. Ez. 1) Ein langer, runder, schwärzlicher, in süßem Wasser lebender Fisch mit einer sehr schlüpfrigen Haut, weßhalb er nicht leicht festgehalten werden kann, (Muraena anguilla L.). Von diesem letzten Umstande sind die auf Menschen, welche sehr gewandt sind, übergetragenen Redensarten hergenommen: Er ist glatt wie ein Aal; ich konnte ihn nicht fassen, er entschlüpfte mir wie ein Aal. Der bunte Aal oder die Meerschlange, der Meeraal, Sandaal. S. d. — Älchen, so nennt man die Würmchen, welche sich in Essig, Kleister ꝛc. erzeugen, s. Essigälchen, Kleisterälchen. 2) Ein Backwerk aus Butterteige in Gestalt eines Aales. 3) Die falschen Brüche, welche beim Walken in den Tüchern entstehen.

Fine Grain Structuring of the Entry

Der Aal, | *des –es,* | *Mz. die –e,*

Verkleinerungswort, | *das* | *Älchen,*

des –s, | *b. Mz. w. b. Ez.*

1) Ein langer, runder ... Fisch ...

2) Ein Backwerk aus Butterteig ...

3) Die fal=schen Brüche, ...

Annotation in TEI P5 (Text Encoding Initiative)

Der Aal, ...

```
<entry xml:id="cwds1_00005_aal">
   <form type="lemma">
      <gramGrp>
         <pos value="noun"/>
         <gen value="m"/>
      </gramGrp>
      <form type="determiner">Der</form>
      <form type="headword">Aal</form>
      <pc>,</pc>
   </form> ...
   <sense> ... </sense>
</entry>
```

Extended Definite Clause Grammars

```
entry ===>
    form:[type:lemma],
    ...,
    sense.
form:[type:lemma] ===>
    sequence(*, form:[type:determiner]),
    form:[type:headword].
sense ===> ...
```

The call `sequence(*, form:[type:determiner])` generates a sequence of zero or more `form` elements.

Techniques from Computer Science

Grammars

- higher precision compared to regular expressions and statistical parsers
- we use a DCG (definite clause grammar) extension, which is even more compact and directly generates XML

XML is a common data format for modelling, managing, and exchanging semi–structured data.

- There exist powerful query, transformation and update languages for XML.

Declarative Languages

Examples

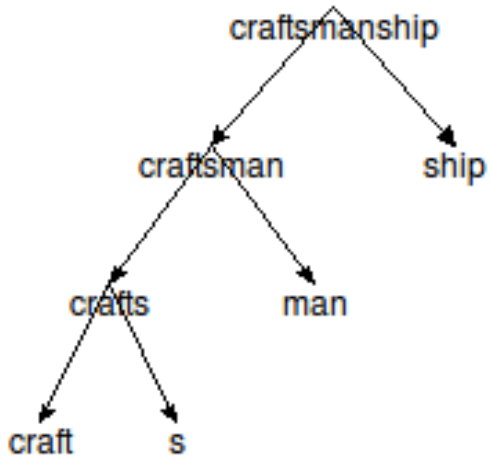
- SQL (relational databases)
- XQUERY, XSLT (XML processing)
- PROLOG (programming)
- rules (decision support systems, grammars)

Advantages

- compakt, rapidly programmable
- clear, less error–prone
- flexibly extensible

Annotating Morpheme Decompositions

- . . . based on the Whole Word Morphology
- extension by alignment methods
- morpheme decomposition:



morpheme term: `((craft + s) + man) + ship`

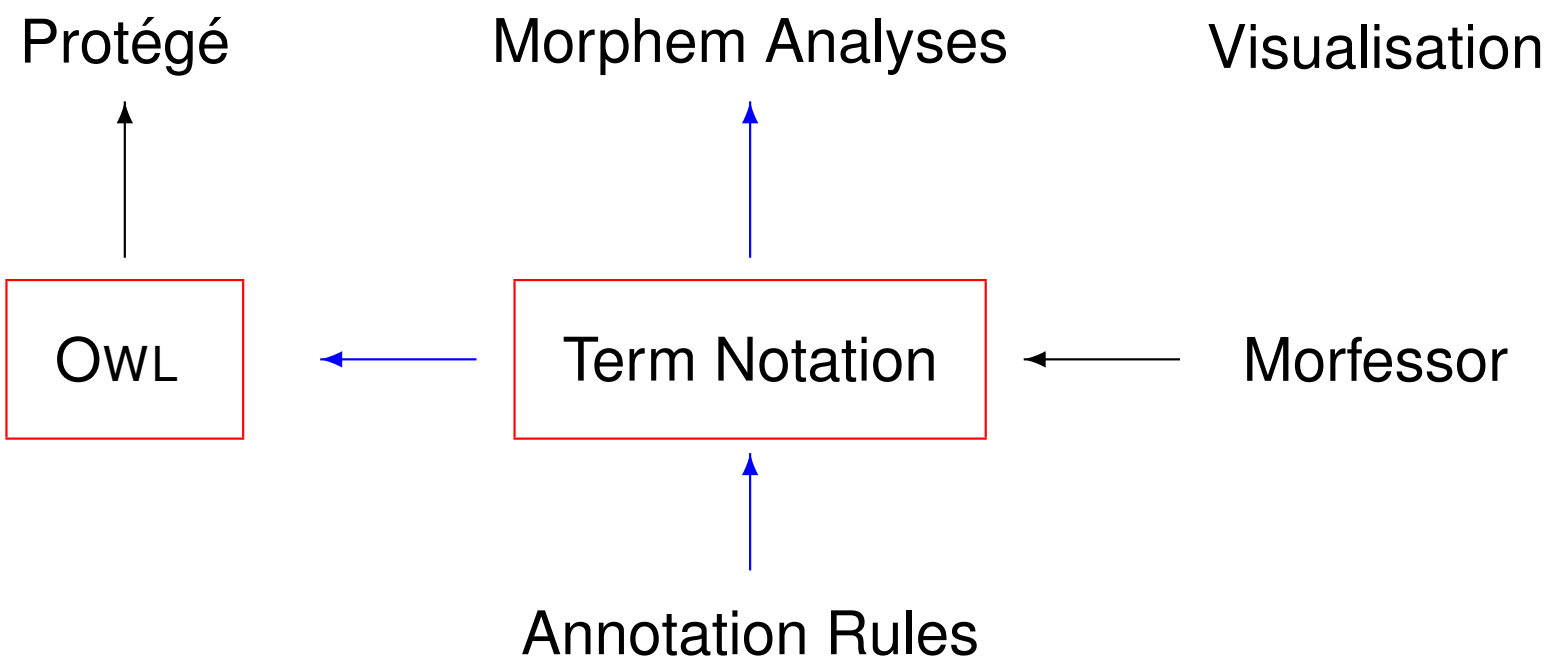
System Architecture

For decomposing and annotating the large number of entries of a dictionary (which can exceed 100.000), one needs

- linguistic knowledge and
- suitable tools from computer science:
  - morpheme decomposer,
  - suitable, compact knowledge representation,
  - inference methods,
  - graphical user interface.

Fine grain annotated dictionaries are the basis for the decomposition.

System Architecture



Annotation Rules

With the annotation rule (in logic)

```
has_word_class(X, noun) :-
    mc(X, A, B),
    has_word_class(A, noun),
    has_text_form(B, [ship, ...]).
```

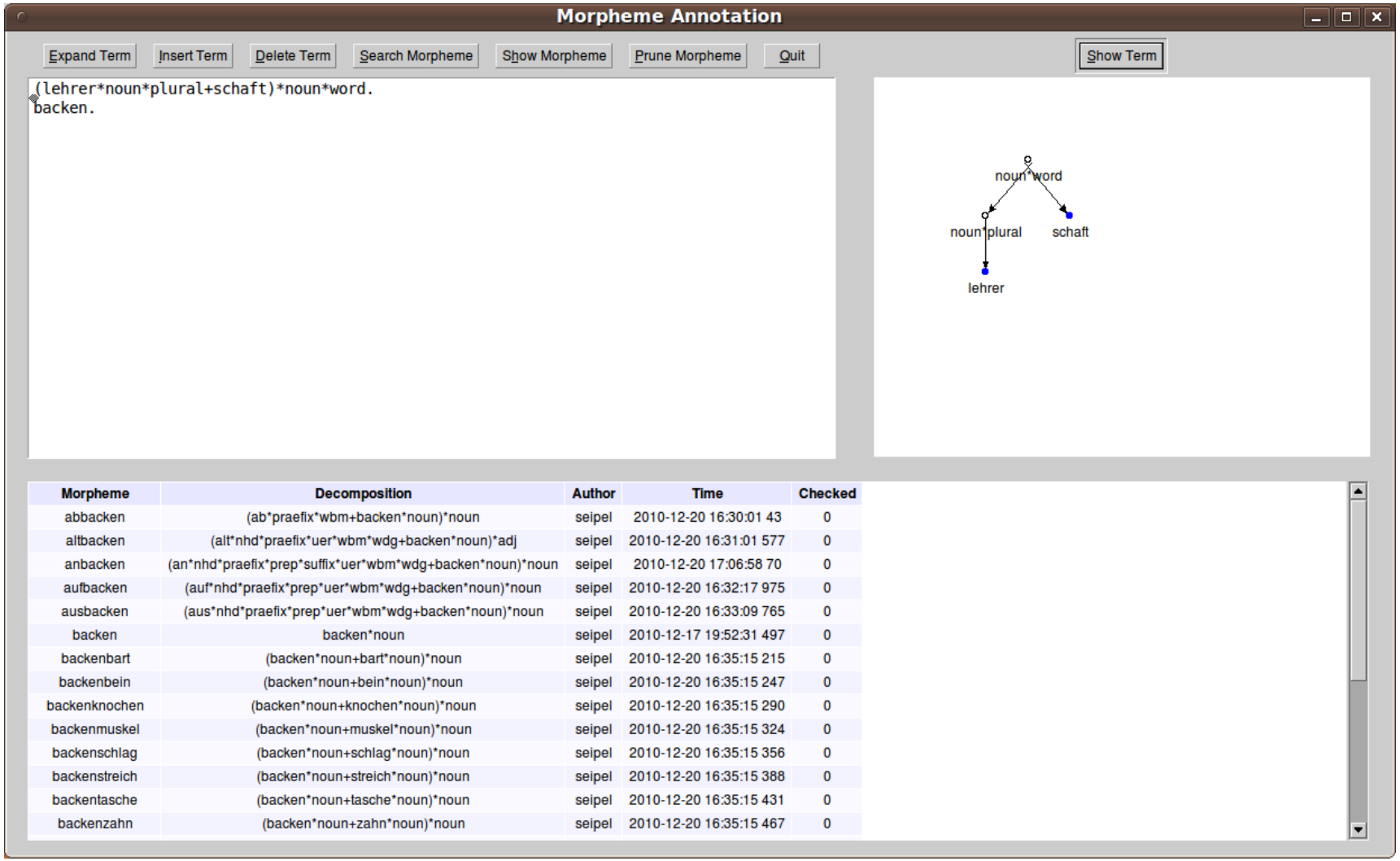
the partially annotated term

```
((craft*bm + s*ge) + man)*noun + ship
```

can be further annotated to

```
(((craft*bm + s*ge) + man)*noun + ship)*noun
```

The Morpheme Annotation Tool



Morpheme	Decomposition	Author	Time	Checked
abbacken	(ab*praefix*wbm+backen*noun)*noun	seipel	2010-12-20 16:30:01 43	0
altbacken	(alt*nhd*praefix*uer*wbm*wdg+backen*noun)*adj	seipel	2010-12-20 16:31:01 577	0
anbacken	(an*nhd*praefix*prep*suffix*uer*wbm*wdg+backen*noun)*noun	seipel	2010-12-20 17:06:58 70	0
aufbacken	(auf*nhd*praefix*prep*uer*wbm*wdg+backen*noun)*noun	seipel	2010-12-20 16:32:17 975	0
ausbacken	(aus*nhd*praefix*prep*uer*wbm*wdg+backen*noun)*noun	seipel	2010-12-20 16:33:09 765	0
backen	backen*noun	seipel	2010-12-17 19:52:31 497	0
backenbart	(backen*noun+bart*noun)*noun	seipel	2010-12-20 16:35:15 215	0
backenbein	(backen*noun+bein*noun)*noun	seipel	2010-12-20 16:35:15 247	0
backenknochen	(backen*noun+knochen*noun)*noun	seipel	2010-12-20 16:35:15 290	0
backenmuskel	(backen*noun+muskel*noun)*noun	seipel	2010-12-20 16:35:15 324	0
backenschlag	(backen*noun+schlag*noun)*noun	seipel	2010-12-20 16:35:15 356	0
backenstreich	(backen*noun+streich*noun)*noun	seipel	2010-12-20 16:35:15 388	0
backentasche	(backen*noun+tasche*noun)*noun	seipel	2010-12-20 16:35:15 431	0
backenzahn	(backen*noun+zahn*noun)*noun	seipel	2010-12-20 16:35:15 467	0

Conclusions

The metaDictionary forms the core part of a generic e–infrastructure:

- derived from analysis of a network of dictionaries
- annotated morpheme decompositions
  yield a more precise alignment for the metaDictionary

The next step will be to test the data using text corpora:

- basic morphemes
- combinations of basic morphemes

Culturomics (Michel et al., Science 2011): *52% of the English lexicon – the majority of the words used in English books – consists of lexical dark matter undocumented in standard references*.