A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts

Martin Scholz

Friedrich-Alexander-University Erlangen-Nürnberg

Digital Humanities Research Group

Outline

- Motivation: information extraction from free text
- GermaNet — a German wordnet
- Simple mapping CRM events → GermaNet
- Fine-grained mapping
- Statistics & evaluation
- Difficulties with modelling event mentions

Semantic Annotation of Free Text

- In CH documentation, information is often encoded as free text.
- Poorly machine-processable
  - Full text search
- Extract information as structured data
  - Entities, events and relations
  - Poor quality with automatic extraction

Free text — Example

Die Tafel zeigt die legendäre, drei Tage und zwei Nächte dauernde Schlacht.

Ein Engel kam zu Hilfe und führte den Sieg herbei.

Zum Dank stiftete Karl das Schottenkloster in Regensburg.

1514 entbrannte um dessen Fortbestand ein Streit zwischen Bischof, Kaiser und Stadtrat.

The panel shows the legendary, three days long battle.

An angel came to their aid and led them to victory.

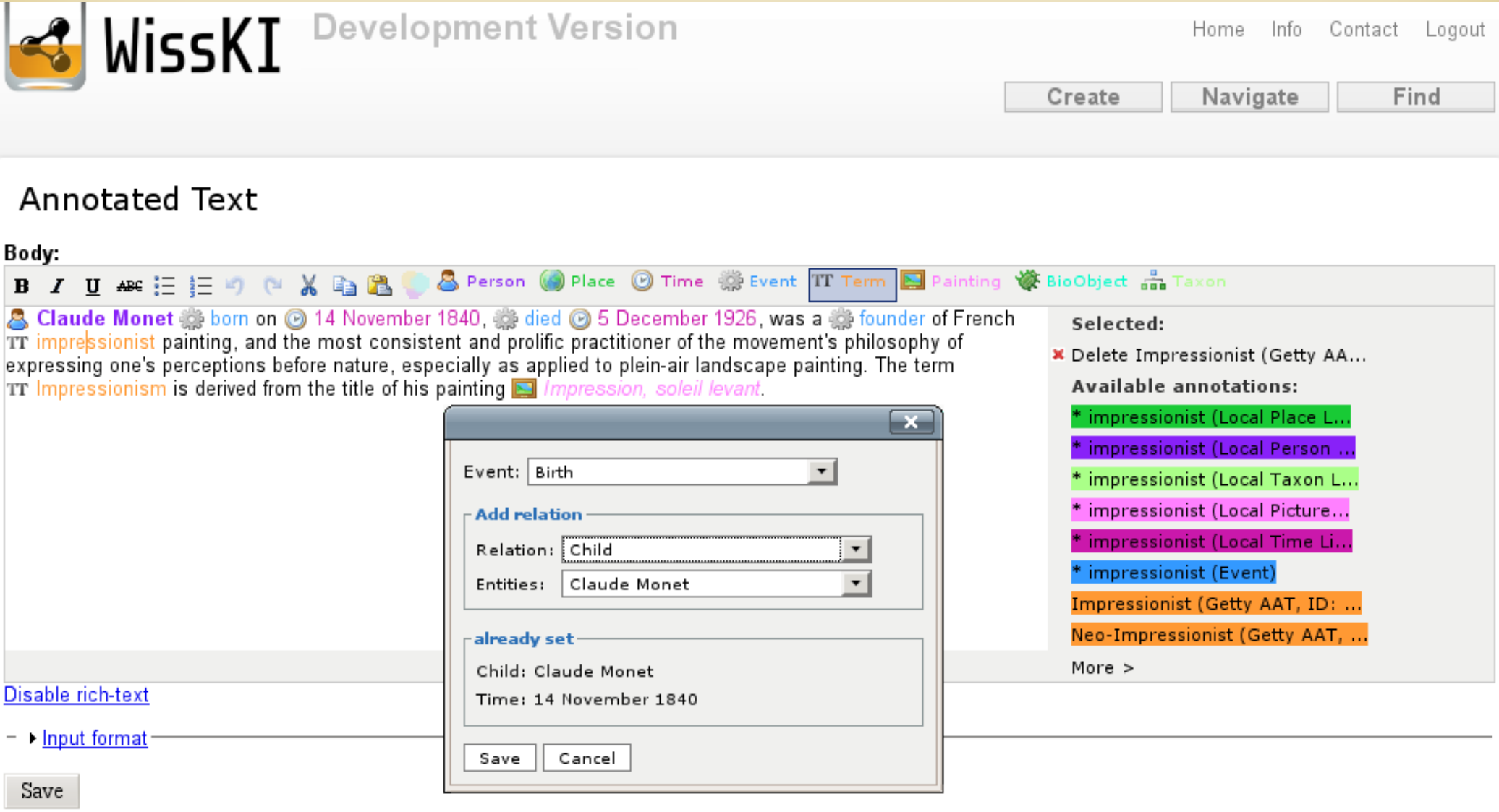
In return, Karl founded the Schottenkloster in Regensburg.

In 1514, the quesion about its continued existence gave rise to a conflict between the bishop, emperor and the city council.

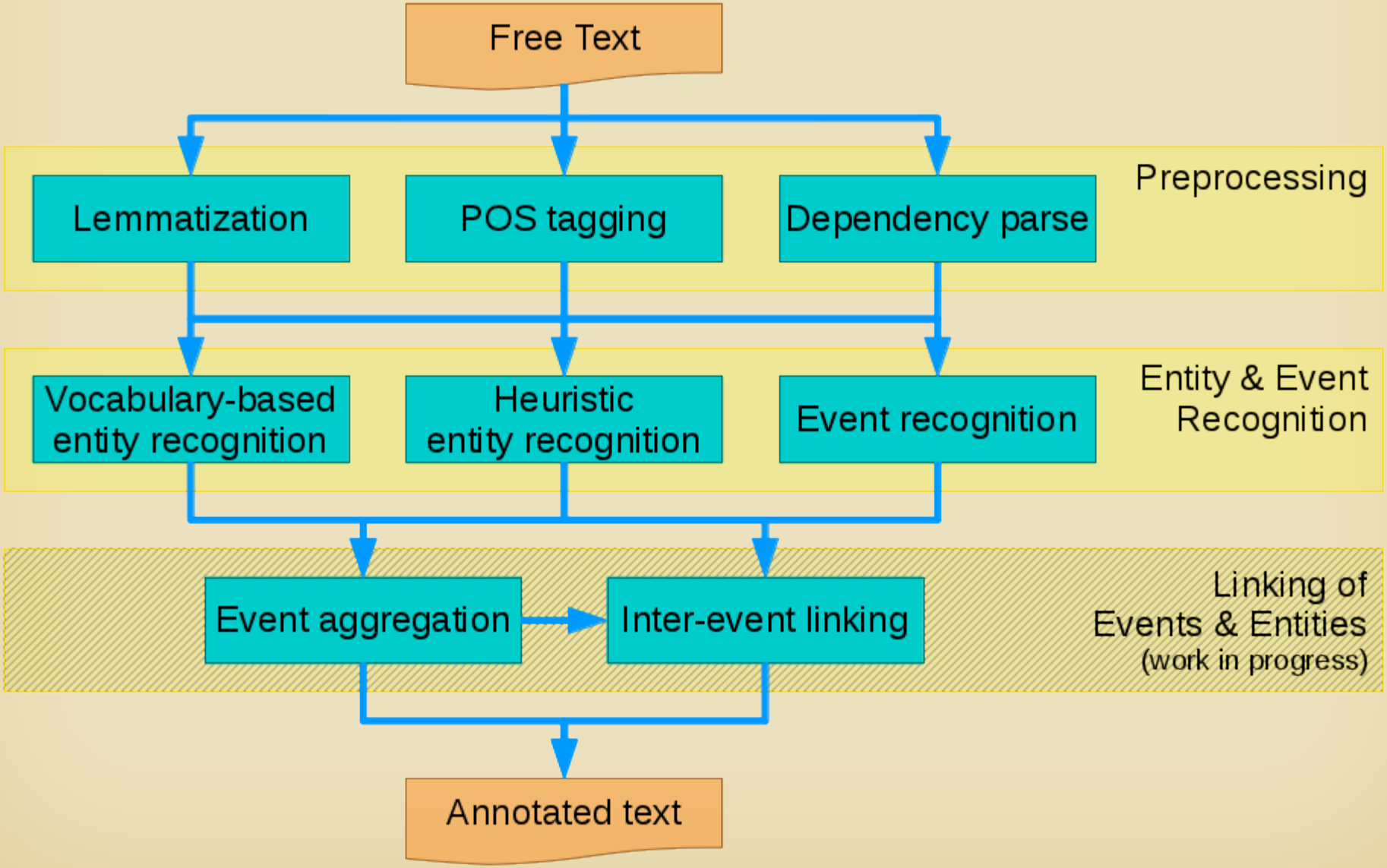
Use case: WissKI

- Semi-automatic information extraction
  - Not unsupervised
  - User annotates text
    - From annotations structured data is generated
  - Machine gives annotation proposals
    - Not only one best solution, but
    - Choice between possible/likely annotations

WissKI — Text annotation



Text analysis pipe



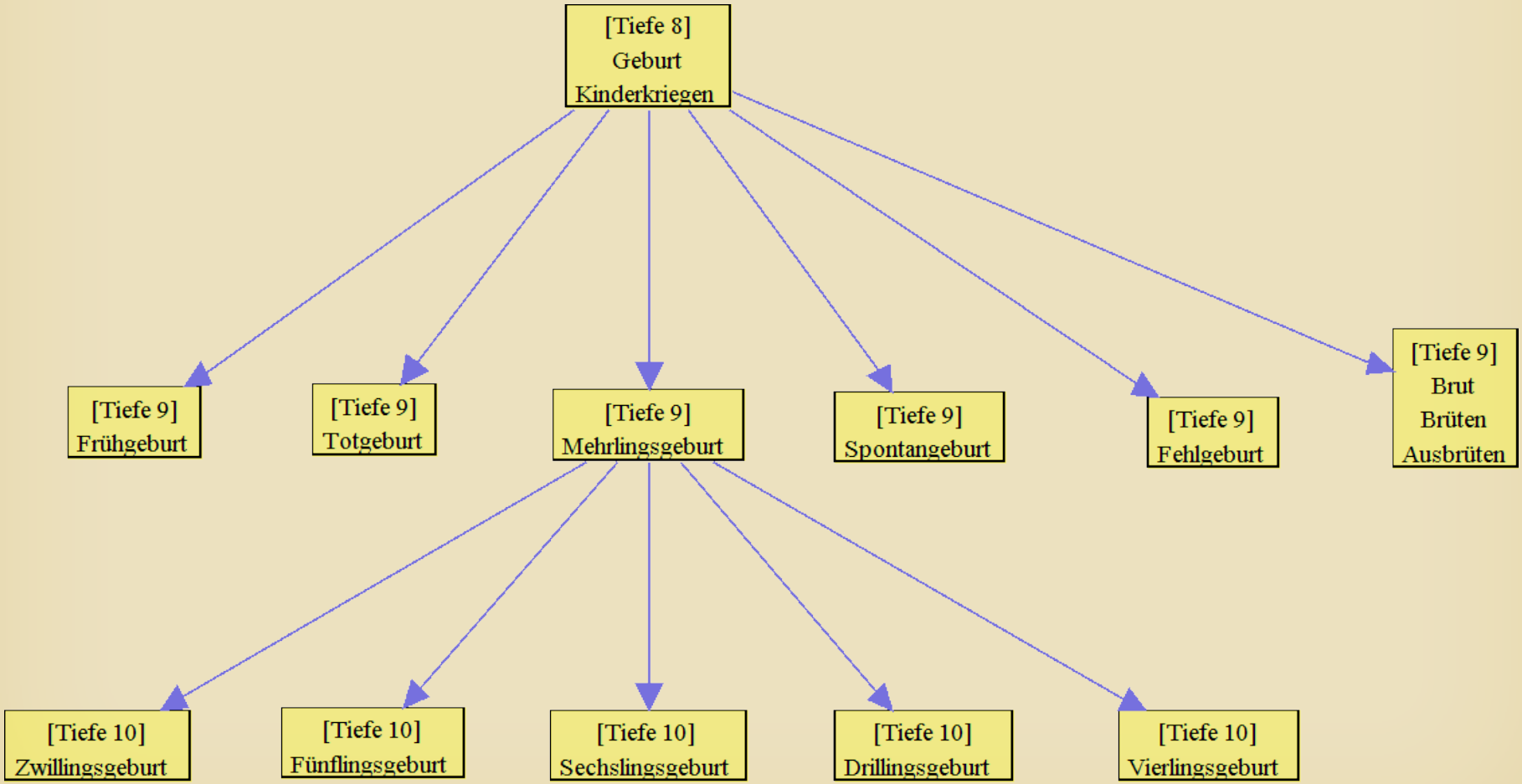
Building a lexicon

- Lexicon-based detection of events (in its core)
  - Lexicon of words that support an event class
- Lexicon creation
  - Manual creation is time-consuming
  - No corpus to build from
  - Idea: use an intermediate lexicon for bootstrapping

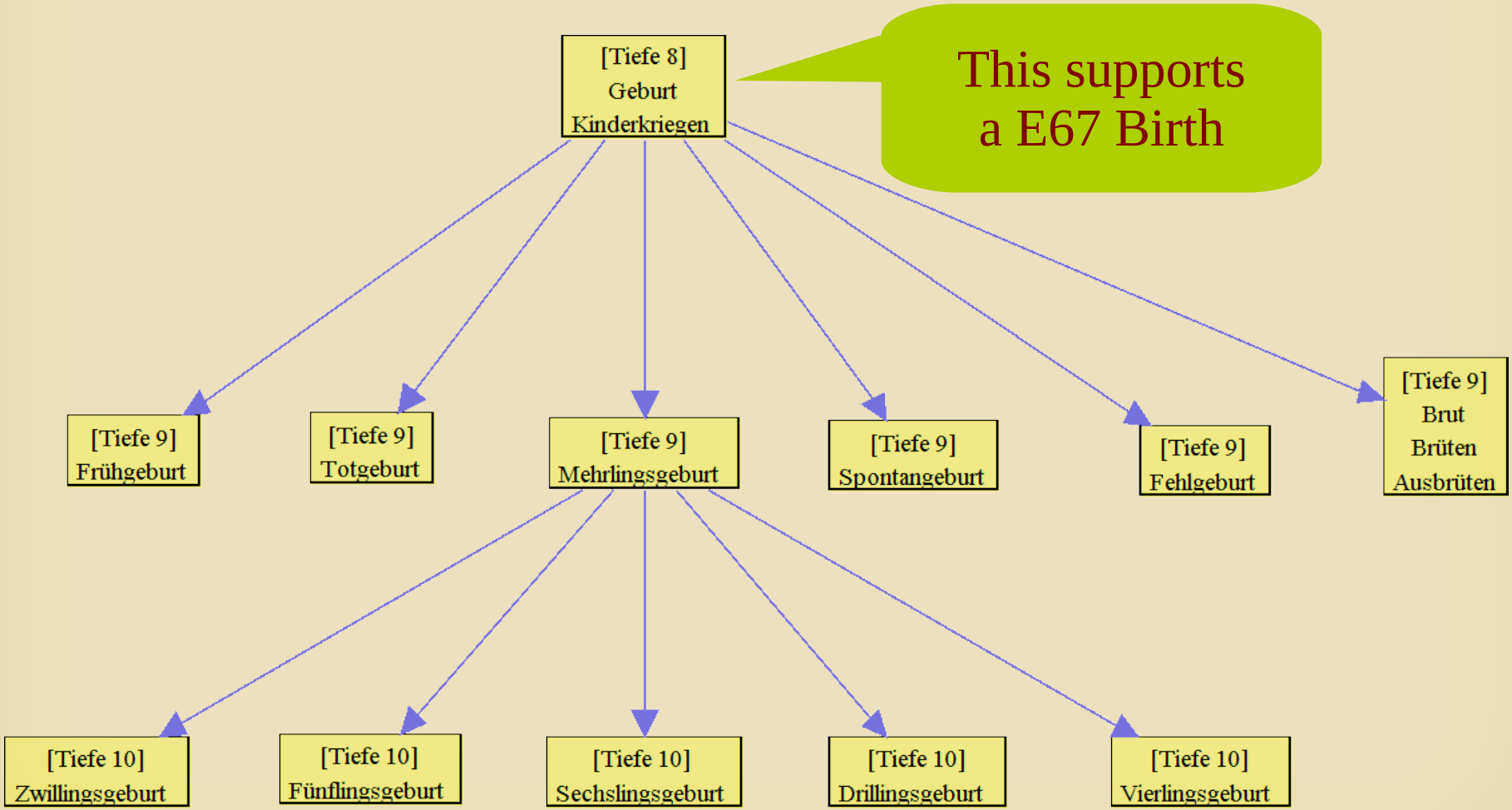
GermaNet

- A word net for German
  - Similar to Princeton Wordnet
- Word (sense) ↔ Synset
- Relations between synsets
  - Hyponymy, Meronymy, ...
- ~ 75k synsets, ~ 100k word senses
- This work uses version 7 (latest version: 8)

GermaNet — Example



GermaNet — Example



A simple mapping

```
<class name="ecrm:E67_Birth">
  <synset pos="v" word="gebären" sense="1" />
  <synset pos="n" word="Geburt" sense="1" />
  <synset pos="n" word="Geburt" sense="2" />
  <synset pos="n" word="Geburt" sense="3" />
</class>
```

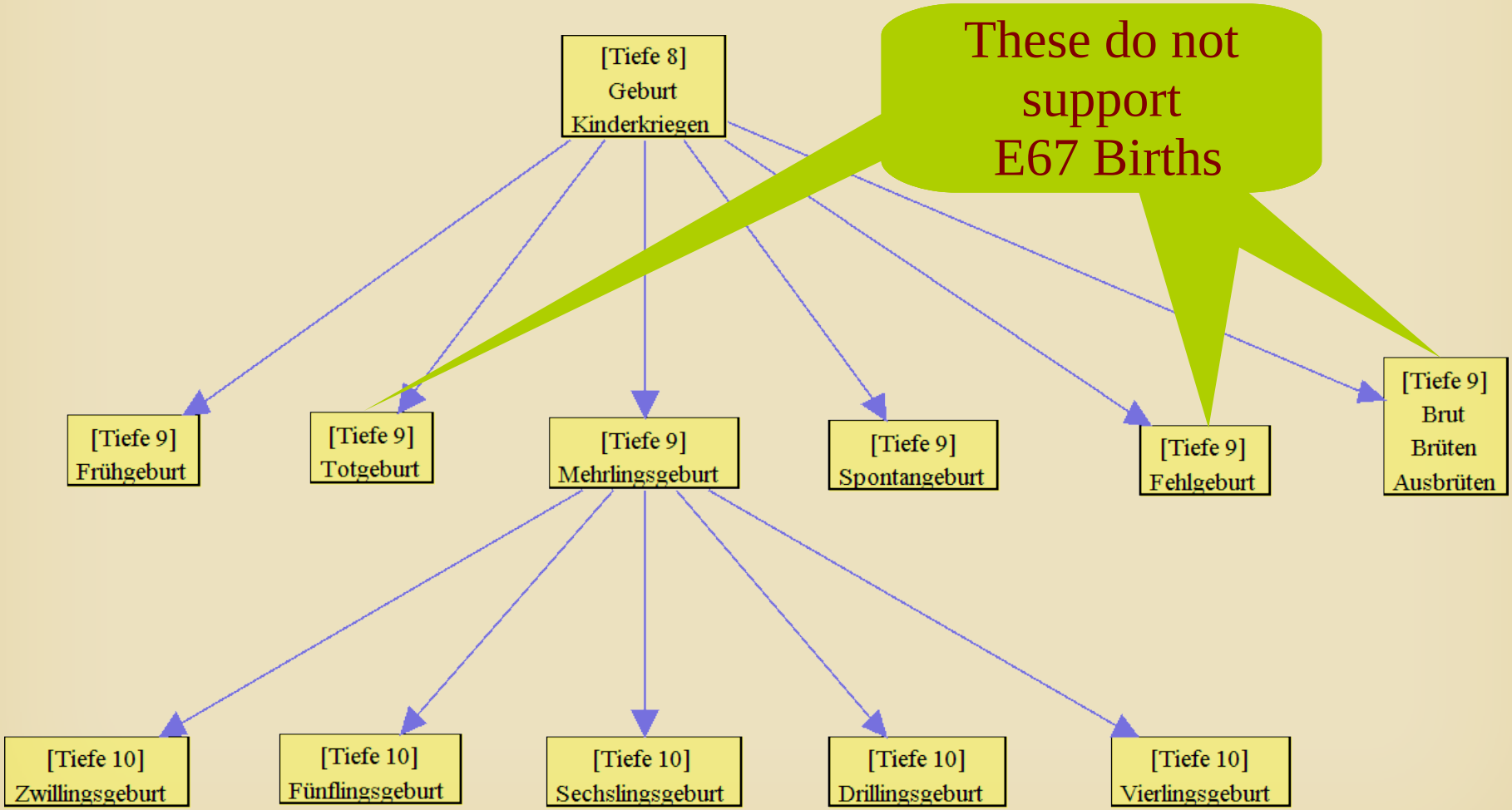
A simple mapping — Compiled

```
<class name="ecrm:E67_Birth">
    <word lemma="gebären" pos="v"/>
    <word lemma="entbinden" pos="v"/>
    <word lemma="laichen" pos="v"/>
    ...
    <word lemma="Geburt" pos="n"/>
    <word lemma="Drillingsgeburt" pos="n"/>
    <word lemma="Brut" pos="n"/>
    ...
</class>
```

A simple mapping — Compiled

```
<class name="ecrm:E67_Birth">
    <word lemma="gebären" pos="v"/>
    <word lemma="entbinden" pos="v"/>
    <word lemma="laichen" pos="v"/>
    ...
    <word lemma="Geburt" pos="n"/>
    <word lemma="Drillingsgeburt" pos="n"/>
    <word lemma="Brut" pos="n"/>
    ...
</class>
```

GermaNet — Example



A simple mapping — Problem

- Meanings of GermaNet synsets don't match CRM events exactly
  - A synset supports an event class, but some or even all of its hyponymic synsets may not
- Scope of CRM event is not reflected in language
  - E67 Birth / E69 Death only applies to humans
    - Personalized „actors“ / figurative meanings
  - E12 Production, E65 Creation, etc. distinguish material ↔ immaterial

Fine-grained mapping

- Sort out obvious false positive synsets/words
  - Exclude a synset and its descendants or
  - Exclude all descendants of a synset

- Drawbacks
  - Creation and maintenance more complex and time-consuming
  - Hyponymic synsets can get sorted out wrongly

Fine-grained mapping

```
<class name="ecrm:E67_Birth">
  <synset pos="n" word="Geburt" sense="1">
    <exclude_synset word="Brut" sense="1" />
    ...
  </synset>

<class name="ecrm:E66_Formation">
  <synset pos="n" word="Heirat" sense="1" descend="false" />
  <synset pos="n" word="Liebesheirat" sense="1" />
  ...
```

Mapping statistics

	synsets	excludes	words (1)	words (2)	Δwords	% reduction
E5 Event	2	19	22817	20005	2812	12.3
E6 Destruction	6	12	268	239	29	10.8
E7 Activity	12	7	17310	16065	1245	7.2
E8 Acquisition	15	42	1357	839	518	38.2
E9 Move	9	21	1880	1823	57	3.0
E10 Transfer of Custody	15	17	1357	1016	341	25.1
E11 Modification	6	212	5321	3933	1388	26.1
E12 Production	33	27	1540	1158	382	24.8
E63 Beginning of Existence	5	0	5939	4848	1091	18.4
E64 End of Existence	1	0	4070	3160	910	22.4
E65 Creation	37	12	1121	1096	25	2.2
E66 Formation	24	9	281	269	12	4.3
E67 Birth	7	3	46	33	13	28.3
E68 Dissolution	8	0	16	16	0	0.0
E69 Death	10	13	157	129	28	17.8
E79 Part Addition	2	4	127	94	33	26.0
E80 Part Removal	1	5	235	183	52	22.1
E81 Transformation	5	164	3849	2944	905	23.5
E85 Joining	7	3	29	18	11	37.9
E86 Leaving	6	0	105	105	0	0.0

words (1): simple mapping, words (2): fine-grained mapping

Reminder: this is work in progress, not a complete mapping!

Event detection evaluation

- Lexicon-based
- Preprocessor:
  - Detect separable verb particles
- Postprocessor:
  - Detect light verbs
- No word sense disambiguation (currently)

Event detection evaluation

- Hand-made corpus
  - 50 short texts from art history
  - 3000 tokens, 500 event class annotations

	Precision	Recall
Simple mapping	48%	74%
Fine-grained mapping	59%	72%

- Observations
  - Synsets wrongly excluded mapping → Recall
  - Polysemy / no word sense disamb. → Precision

Difficulties with modelling event mentions

- Context influences how to model an event mention in text
  - Which CRM class
  - How many instances

Modelling event mentions: Co-triggered events

- Words primarily denoting objects may also support events

  Gemälde (painting) → E12 Production

  Geschenk (present / gift) → E8 Acquisition, E10 Transfer of Custody

- Influences interpretation:

  John's painting vs. John's house

- Lexicalize to what extend?
  - GermaNet synsets and hierarchy often not suitable

Modelling event mentions: CRM event or Symbolic Object

- An event in the text may not necessarily be modelled as CRM event
- Past events → E5 Event and subclasses
- Future/hypothetical events
  → E55 Type or E29 Design or Procedure

Ein Engel kam zu Hilfe und führte den Sieg herbei.

An angel came to their aid and led them to victory.

Modelling event mentions: Reclassification of events

- Context may reclassify an CRM event supported by a word
  - Negation, interruption, alteration, lack of knowledge
  - More general CRM class
- Document foreseen purpose as E55 Type

Der römische Feldherr Dentatus weist die Geschenke [...] zurück.

Der Beginn mit der Anfertigung [...] erfolgte [...] nach der Anmeldung.

The Roman commander Dentatus rejects the presents [...].

The production [...] started [...] after the application.

Modelling event mentions: Instance(s) or class

- How many event instances are referred to?
  - Individual
  - Collection → E5 or subclass
  - Class → E29 or E55

  } border?
- Can often only determined by event's arguments

Die Einzelteile der zerlegten Retabel gelangten in den Kunsthandel.

Durch den Besitz dieser Güter stellten wohlhabende Bürger [...] zur Schau.

The parts of the disassembled retable ended up in art trade.

By the possession of these goods wealthy people showed [...].

Thank you!

Mapping & Compiler:

http://wiss-ki.eu/node/167

wiss-ki.eu