# The D-SPIN Text Corpus Format and International Standards

Ulrich Heid, Kerstin Eckart

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

Garching, October 2009: D-SPIN Advisory Board Meeting

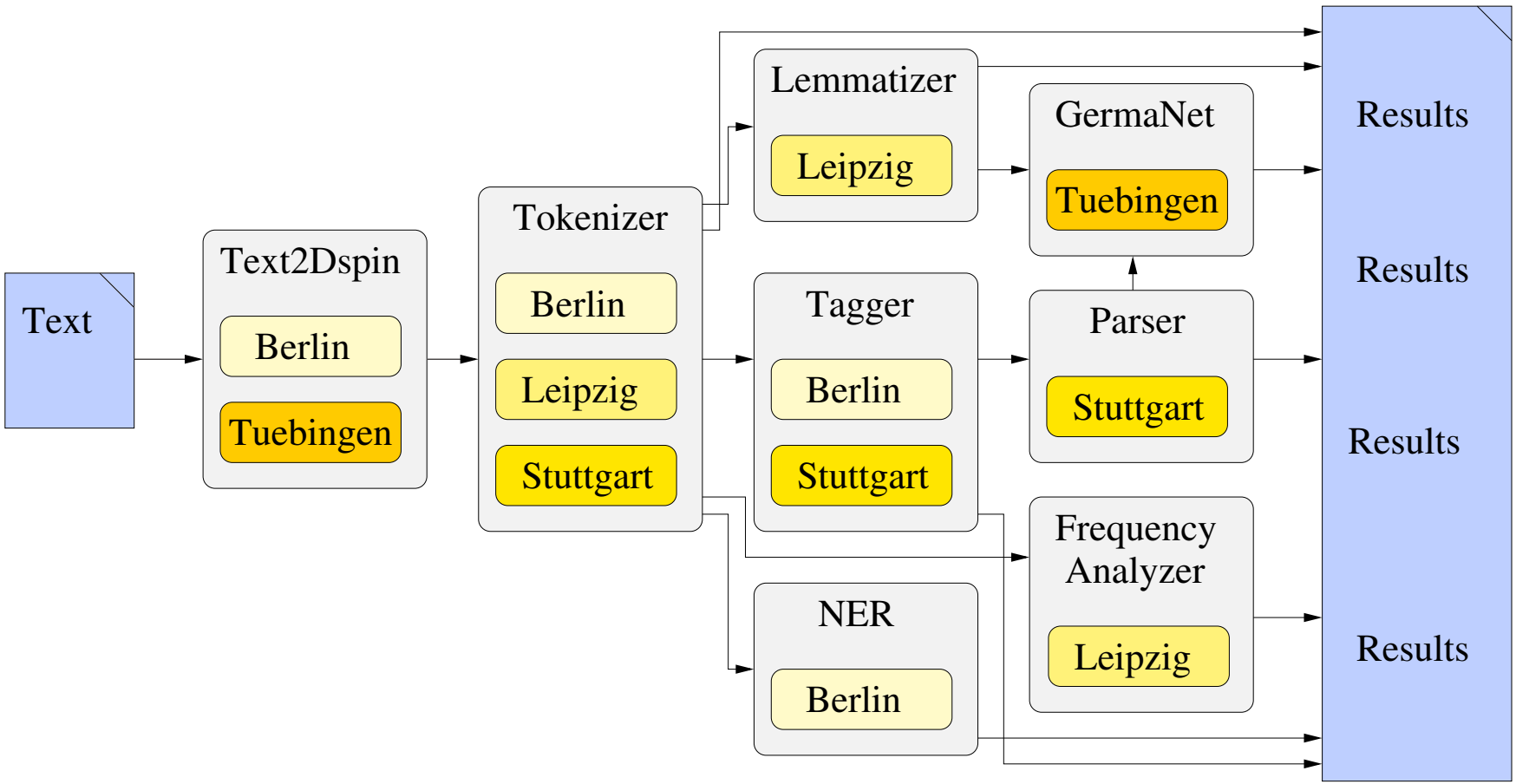
# Overview

- Need for standardisation in D-SPIN:
  scenarios and types of needs
- Current state of the
  international standardisation of Language Resources
  - Areas
  - Vision of an integrated representation
- Ongoing work on standards in D-SPIN
  and open issues
- Conclusion

# Need for standardisation in D-SPIN

## Sample scenarios

Corpus processing in the WebLicht tool chain:



# Need for standardisation in D-SPIN

## Sample scenarios: Corpus processing

(1) Comparing tools of the same functionality:
e.g. Taggers from Berlin and Stuttgart
   - one text
   - different annotations
   - possibly: different tagsets

(2) Combining tools for different functionalities:
e.g. Tokenizer - Tagger - Parser
   - one text
   - annotations from different tools and steps
   - possibly: different tagsets

# Need for standardisation in D-SPIN

## Interpreting the sample scenarios

(1) Tool comparison:
  - Need for common vocabulary to convey results of linguistic analysis
  - cf. "Tagset mapping": common data categories

(2) Tool combination:
  - Need for common structure and representation of annotations
  - cf. "Format conversion"
  - Plus need for common data categories

# International standardisation of linguistic resources

## Philosophy

- Expert groups: open forum for discussion
  Language resources: ISO TC37/SC-4
- Structured set of procedures to prepare international standards, organised by ISO
- National mirror committees:
  in Germany organised by DIN

- Standardisation by consensus:
  Proposals – discussion – improvements – ...
- Different related projects and initiatives:
  CLARIN/D-SPIN, FLaReNet, Language Grid, DARIAH

# International standardisation of linguistic resources

## Fields of technology covered

- Infrastructure for storage and processing:
  - Archiving of language resources: PIDs
  - WebService technologies, e.g. UDDI, ...
- Representational Infrastructure:
  - Unicode
  - Stand-off XML and XML Schemas; RDF, OWL, ...
- Metadata for resources

- Structure and representation of (e.g. corpus) annotations
- Vocabulary for conveying analysis results:
  data categories

# International standardisation of linguistic resources

## Levels of abstraction in language resource standardisation

- Metamodels:
  - how to do annotation
  - how to represent text and annotations
  - using a graph-based formalism with feature structures
- Models for specific linguistic layers:
  Proposals for morphosyntax, syntax
  and (some aspects of) lexical semantics
- Sample instantiations: exemplification (e.g. in annexes to standards)

# International standardisation of linguistic resources

## Standards relevant for D-SPIN corpus work

| Aspect → <br> Abstraction Level ↓ | | Structural aspects <br> Annotation Feat. Structs. | "Vocabulary" |
|---|---|---|---|
| Metamodel | Representation | GrAF | |
| | Annotation | LAF | |
| Methodology | Feature Str. | FSD | |
| | Feature Str. | FSR | |
| Layer Model | Lex. Sem. | SemAF | DCR |
| | Syntax | SynAF | DCR |
| | Morphosynt. | MAF | DCR |

# International standardisation of linguistic resources

## Corpus related standards proposals: a list of titles

- Metamodels: Representation of texts and annotations:
  - GrAF: Graph Annotation Framwork
  - LAF: Linguistic Annotation Framwork
- Methodology for feature structure-based representation:
  - FSD: Feature System Declarations
  - FSR: Feature Structure Representation
- Models for single layers of linguistic description:
  - SemAF: Semantic Annotation Framework
  - SynAF: Syntactic Annotation Framework
  - MAF: Morphosyntactic Annotation Framework

$\rightarrow$ Philosophy: multi-layered annotation

# International standardisation of linguistic resources

## Principles underlying the corpus-related standards

- Graph-based representation: nodes, edges;
  Annotations at nodes and (possibly) at edges
- Feature structures: easy to integrate with graphs

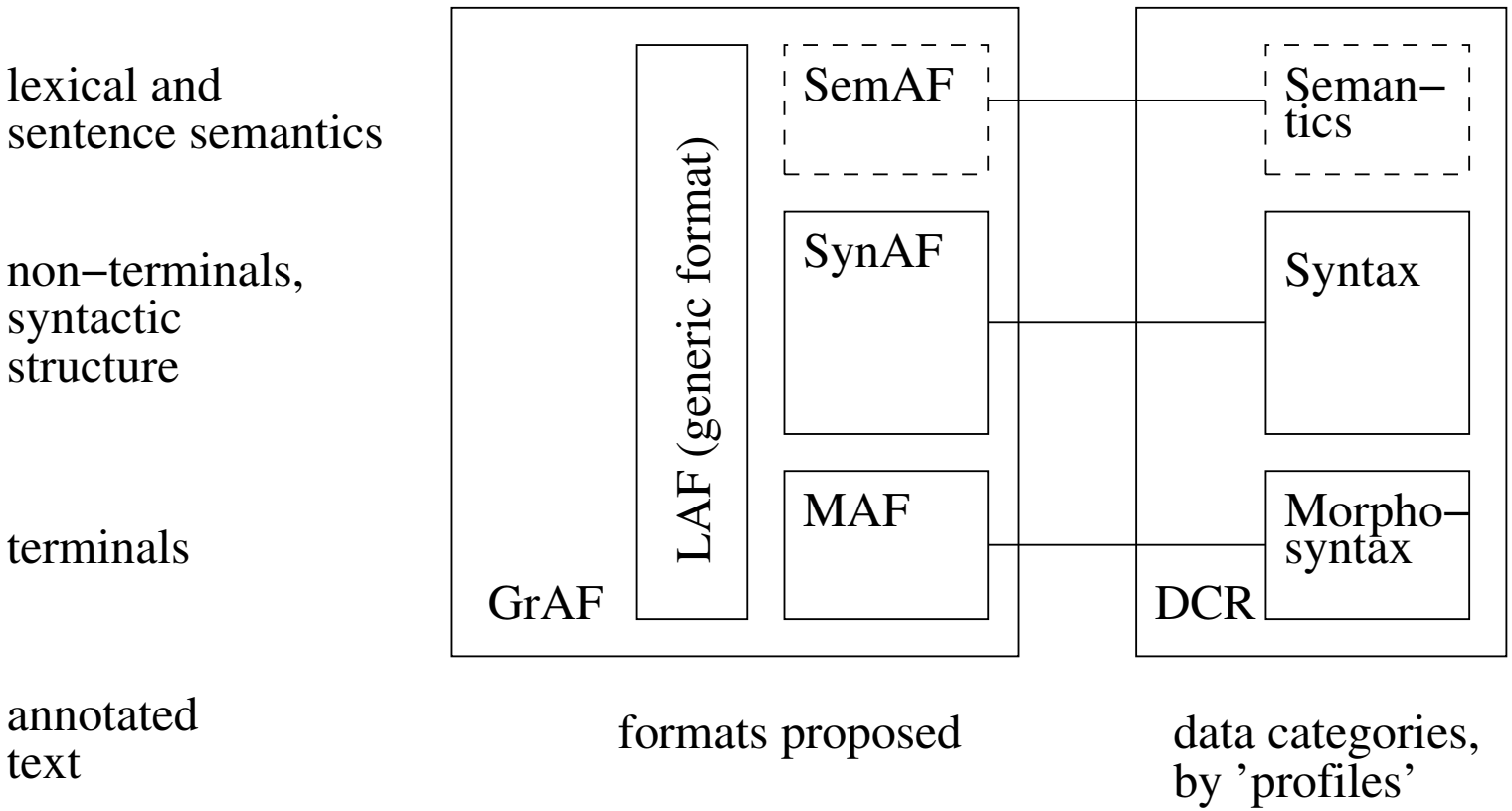
- "Vocabulary" from DCR: Data Category Registry:
  implemented as ISOcat:
  online registry and facility for "tagset mapping"

Vision:

- Integrated format for data exchange:
  multi-layered – mappable data categories – pivot function
- Mapping of data categories: as far as possible

# International standardisation of linguistic resources

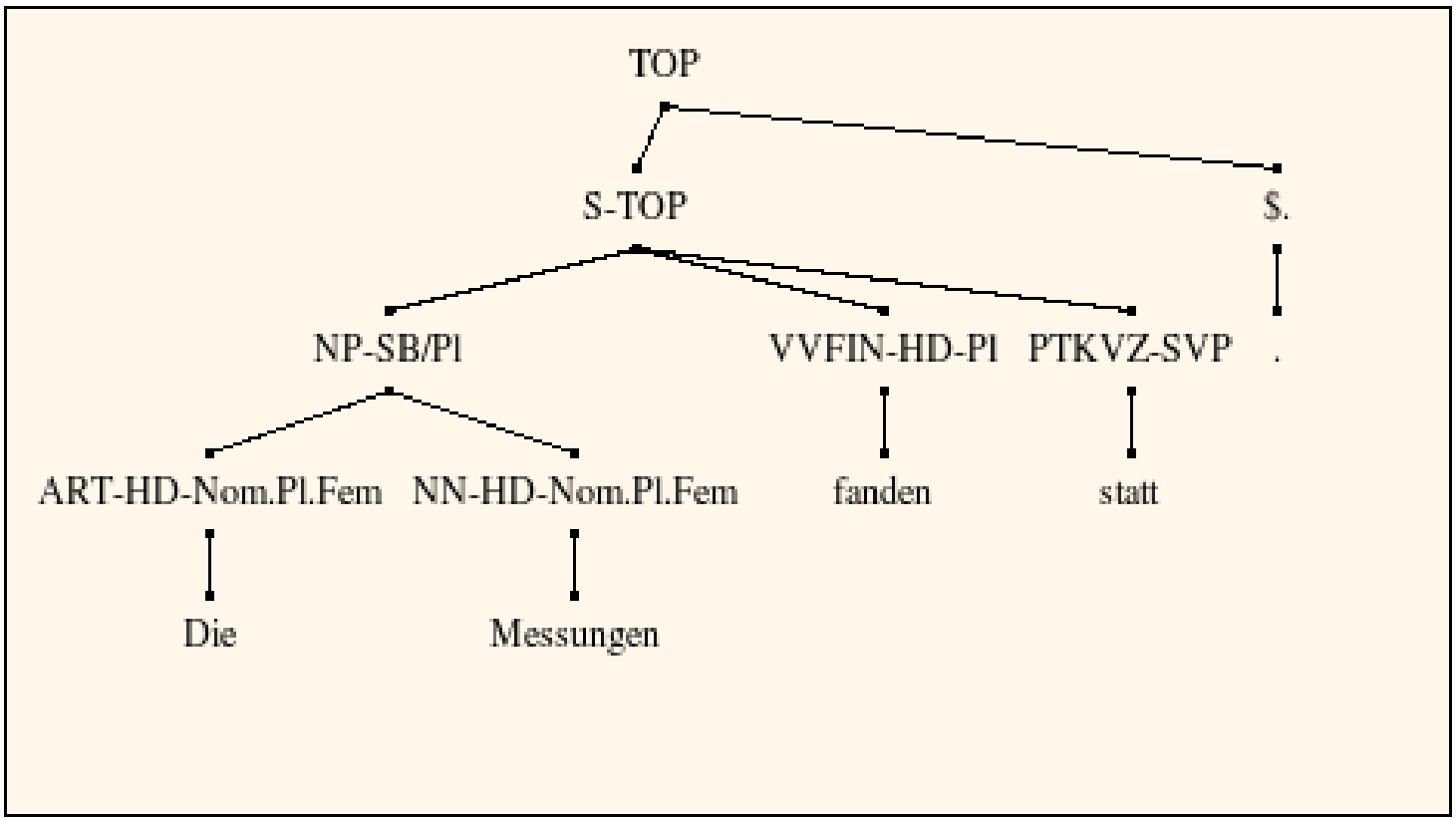
## Corpus-related standards



# International standardisation of linguistic resources

## Corpus-related standards: Experiment: Representing a sample sentence in LAF

Sample sentence: *Die Messungen fanden statt.* (the measurements took place)



# International standardisation of linguistic resources

## Corpus-related standards: Example

References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>    Die
<graf:region id=''r2'' anchors=''4 13''/>   Messungen
<graf:region id=''r3'' anchors=''14 20''/>  fanden
<graf:region id=''r4'' anchors=''21 26''/>  statt
<graf:region id=''r5'' anchors=''26 27''/>  .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example

### References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>     Die
<graf:region id=''r2'' anchors=''4 13''/>    Messungen
<graf:region id=''r3'' anchors=''14 20''/>   fanden
<graf:region id=''r4'' anchors=''21 26''/>   statt
<graf:region id=''r5'' anchors=''26 27''/>   .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example

References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>     Die
<graf:region id=''r2'' anchors=''4 13''/>    Messungen
<graf:region id=''r3'' anchors=''14 20''/>   fanden
<graf:region id=''r4'' anchors=''21 26''/>   statt
<graf:region id=''r5'' anchors=''26 27''/>   .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example

References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>     Die
<graf:region id=''r2'' anchors=''4 13''/>    Messungen
<graf:region id=''r3'' anchors=''14 20''/>   fanden
<graf:region id=''r4'' anchors=''21 26''/>   statt
<graf:region id=''r5'' anchors=''26 27''/>   .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example

References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>     Die
<graf:region id=''r2'' anchors=''4 13''/>    Messungen
<graf:region id=''r3'' anchors=''14 20''/>   fanden
<graf:region id=''r4'' anchors=''21 26''/>   statt
<graf:region id=''r5'' anchors=''26 27''/>   .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example

References to primary data: character-offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                    1                   2
```

LAF/GrAF: **region**

*SynAF-subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>     Die
<graf:region id=''r2'' anchors=''4 13''/>    Messungen
<graf:region id=''r3'' anchors=''14 20''/>   fanden
<graf:region id=''r4'' anchors=''21 26''/>   statt
<graf:region id=''r5'' anchors=''26 27''/>   .
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>                                    node
  <graf:link to=''r1''/>
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

terminal node

reference to region r1: ‘‘Die’’

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>              reference to region r1: ''Die''
  <graf:as type=''BitPar''>           annotation set
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>              reference to region r1: ''Die''
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>            morpho-syntactic description
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>                reference to region r1: ''Die''
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>                          feature structure
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – terminal node

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>                reference to region r1: ``Die''
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>            features
          <symbol value=''ART''/>
        </ns1:f>                        (ART-HD-Nom.Pl.Fem Die)
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>                                    node
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

non-terminal node (no link)

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>        annotation set
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>          constituency description
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>                          feature structure
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – non-terminal node

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>   NP node
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – edges

```
<graf:edge id=''e4'' from=''n7'' to=''n3''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''HD''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

```
<graf:edge id=''e5'' from=''n7'' to=''n4''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''SVP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

# International standardisation of linguistic resources

## Corpus-related standards: Example – edges

```
<graf:edge id=''e4'' from=''n7'' to=''n3''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''HD''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

edge with start node n7 (S)
and end node n3 (VVFIN:‘‘findet’’)

```
<graf:edge id=''e5'' from=''n7'' to=''n4''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''SVP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

edge with start node n7 (S)
and end node n4 (PTKVZ:‘‘statt’’)

# International standardisation of linguistic resources

## Corpus-related standards: Example – edges

```
<graf:edge id=''e4'' from=''n7'' to=''n3''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''HD''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

```
<graf:edge id=''e5'' from=''n7'' to=''n4''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''SVP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

edge with start node n7 (S)
and end node n3 (VVFIN:‘‘findet’’)

edge with start node n7 (S)
and end node n4 (PTKVZ:‘‘statt’’)

nodes and edges are annotated with annotation sets

# D-SPIN and the ongoing standardisation work

- CLARIN has a Standardisation Action Plan
  $\longrightarrow$ also relevant for D-SPIN
- D-SPIN is related with ongoing NLP work on German
  - national de-facto standards
  - existing tools and resources
    $\longrightarrow$ need to accommodate both:
    established work on German *and* international standards
- D-SPIN web services
  $\longrightarrow$ need for processing-oriented formats *and* for exchange

# CLARIN Action Plan for Standardisation

## Double Structure

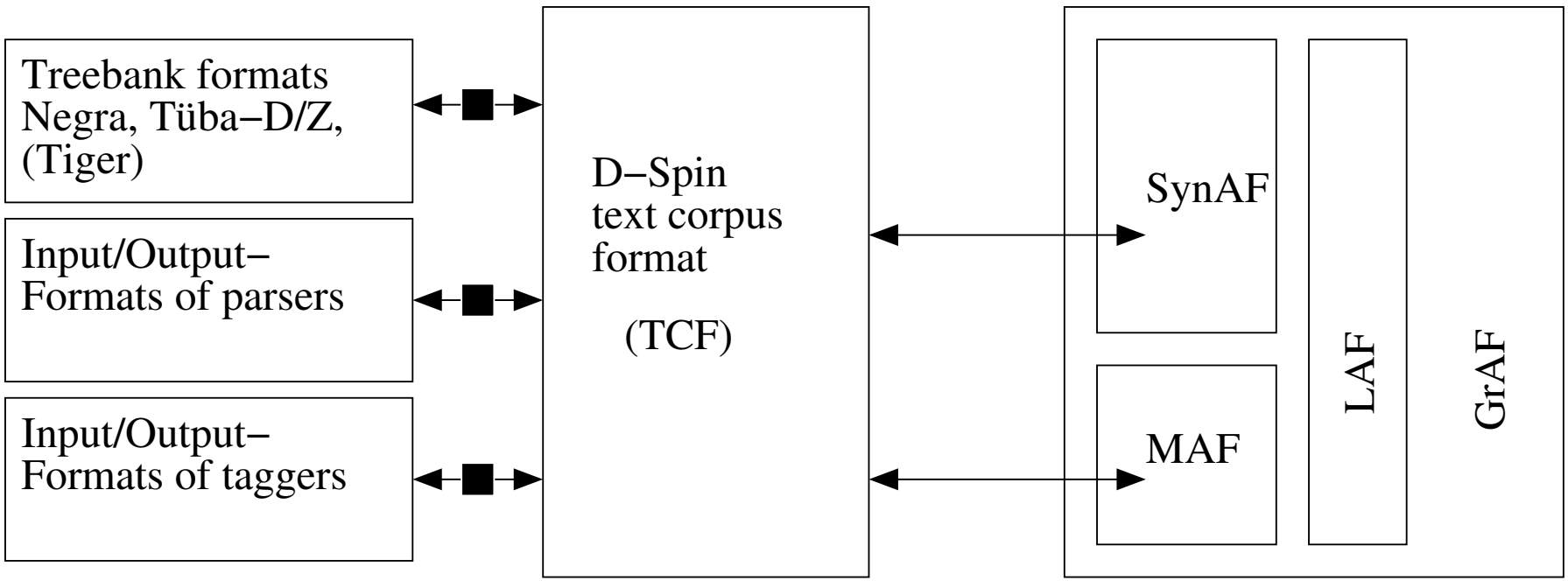
- By topic:
  - storage and processing (PIDs...)
  - representation infrastructure (XML, Unicode...)
  - metadata
  - structure of linguistic representations (LAF, MAF, SynAF)
  - vocabulary for conveying analysis results (DCR)
- By state of completion of the standards
  - official standards:
    recommendation to adopt these standards as a basis for the work
  - ongoing proposals: recommendation for experimental use

# D-SPIN and the CLARIN Action Plan

- Infrastructural Standards
  - PIDs
  - stand-off XML, Schemas...
    $\Rightarrow$ Used as a basis of daily work
- Standards for linguistic representation
  $\Rightarrow$ Experimentation and co-development ongoing
  - Structure of representations: mappings
  - Vocabulary for conveying analysis results:
    comparison and documentation of data categories

# Structure of linguistic representations

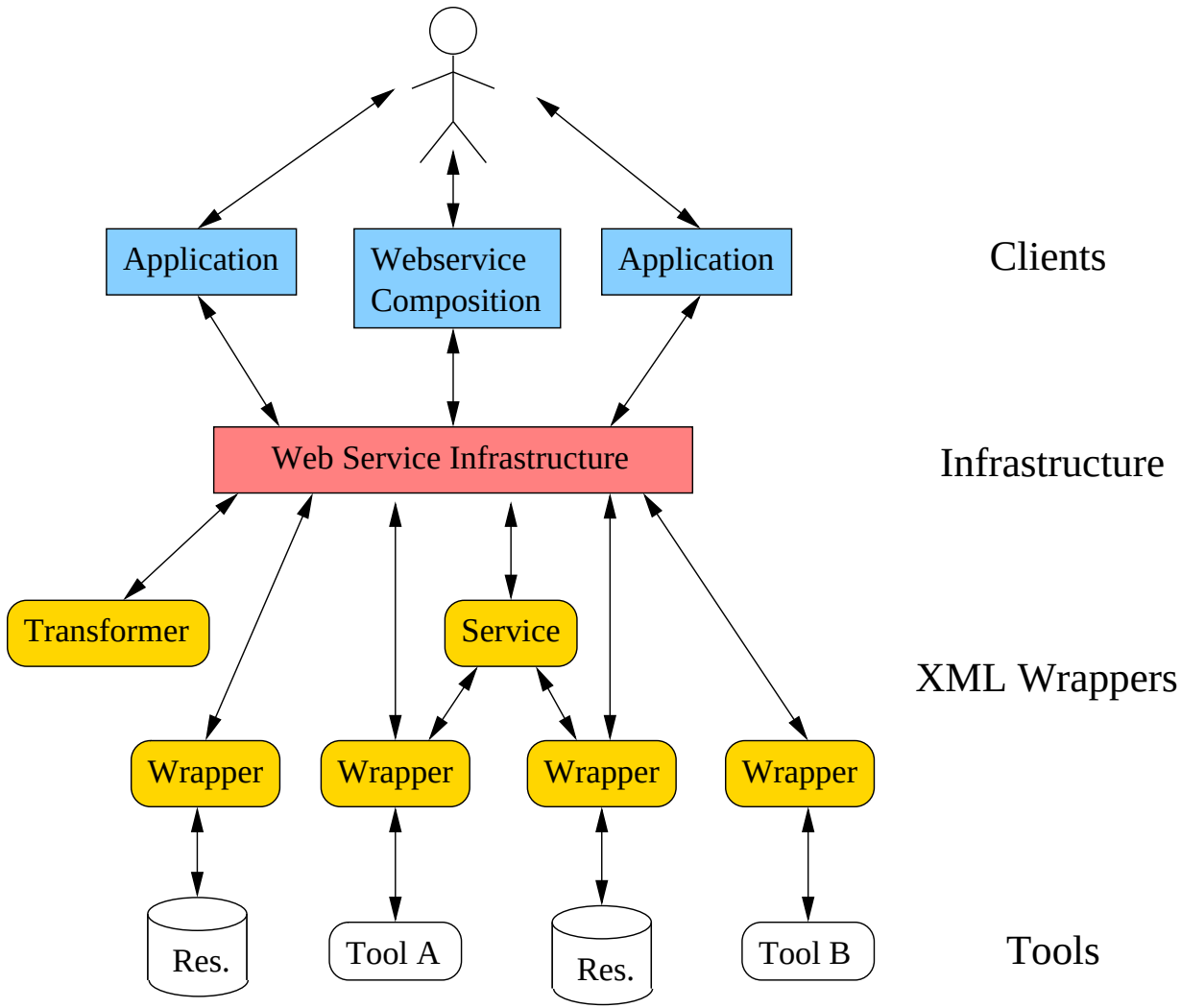
## Annotated corpus data in D-SPIN



# Structure of linguistic representations

## Requirements in terms of formats for annotated text data (1/2)

Recall: layers of the web service architecture



# Structure of linguistic representations

## Requirements in terms of formats for annotated text data (2/2)

Formats in three layers

- tool-internal:
  tools should not be modified
  when being entered into D-SPIN web service chains
- web-service-infrastructural (TCF):
  for exchange between tools of the web service chains
- external:
  for exchange interoperability outside D-SPIN, for archiving,
  combination with other formats, etc.

# D-SPIN Text Corpus Format (TCF)

## Properties: linguistic layers

Developed for efficient processing in web service

- stand-off XML, but "slim"
- processing-related metadata
- layers of linguistic description separated explicitly

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>          information about the document
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

language: German

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...                                    raw text
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>                          text is tokenized
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

part-of-speech annotation

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>                          lemma annotation
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

# D-SPIN Text Corpus Format (TCF)

## Text features for use in the web service chain

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<D-Spin version=''0.3''>
<MetaData/>
  <TextCorpus lang=''de''>
    <text>Charles Perrault,Das Rotkäppchen.
    ...
    </text>
    <tokens>
       <token ...
    </tokens>
    <POStags tagset=''STTS''>
       <tag ...
    </POStags>
    <lemmas>
       <lemma ...
    </lemmas>
    <parsing tagset=''TigerTB''><parse><constituent ...</parse></parsing>
  </TextCorpus>
</D-Spin>
```

(constituent) parse

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

token layer

```
<tokens>
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                          token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...              unique identifier
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                          token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...              token form
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                              token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>                             part of speech
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                              token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>                             part of speech
    <tag tokID=''t2''>NE</tag>
    ...                          token reference
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                          token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>                         part of speech
    <tag tokID=''t2''>NE</tag>
    ...                    POS-Tag (STTS-tagset)
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>
    <lemma tokID=''t2''>Charles</lemma>
    ...
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

token layer

part of speech

lemma annotation

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                          token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>                         part of speech
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>                                          lemma annotation
    <lemma tokID=''t2''>Charles</lemma>
    ...              token reference
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Tokens – POS-tags – Lemmas

```
<tokens>                                          token layer
    <token ID=''t2''>Charles</token>
    <token ID=''t3''>Perrault</token>
    ...
    <token ID=''t1736''>.</token>
</tokens>
<POStags tagset=''STTS''>                         part of speech
    <tag tokID=''t2''>NE</tag>
    ...
    <tag tokID=''t1736''>$.</tag>
</POStags>
<lemmas>                                          lemma annotation
    <lemma tokID=''t2''>Charles</lemma>
    ...              lemma form
    <lemma tokID=''t1736''>.</lemma>
</lemmas>
```

# D-SPIN Text Corpus Format (TCF)

## Sentence layer

```
<parsing tagset=''TigerTB''>
  <parse>
    <constituent cat=''TOP''>
      <constituent cat=''NP-TOP''>
        <constituent cat=''PN-NK-Nom.Sg''>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t2''/></constituent>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t3''/></constituent>
        </constituent>
        <constituent cat=''$,''><tokenRef tokID=t4"/></constituent>
        <constituent cat=''NP-APP''>
          <constituent cat=''ART-NK-Nom.Sg''><tokenRef tokID=''t5''/></constituent>
          <constituent cat=''NE-NK-Nom.Sg''><tokenRef tokID=''t6''/></constituent>
        </constituent>
      </constituent>
      <constituent cat=''$.''><tokenRef tokID=''t7''/></constituent>
    </constituent>
  </parse>
  <parse>
    ...
</parsing>
```

# D-SPIN Text Corpus Format (TCF)

## Sentence layer

```
<parsing tagset=''TigerTB''>
  <parse>                    Charles Perrault,Das Rotkäppchen.
    <constituent cat=''TOP''>
      <constituent cat=''NP-TOP''>
        <constituent cat=''PN-NK-Nom.Sg''>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t2''/></constituent>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t3''/></constituent>
        </constituent>
        <constituent cat=''$,''><tokenRef tokID=t4"/></constituent>
        <constituent cat=''NP-APP''>
          <constituent cat=''ART-NK-Nom.Sg''><tokenRef tokID=''t5''/></constituent>
          <constituent cat=''NE-NK-Nom.Sg''><tokenRef tokID=''t6''/></constituent>
        </constituent>
      </constituent>
      <constituent cat=''$.''><tokenRef tokID=''t7''/></constituent>
    </constituent>
  </parse>
  <parse>
     ...
</parsing>
```

# D-SPIN Text Corpus Format (TCF)

## Sentence layer

```
<parsing tagset=''TigerTB''>
  <parse>
    <constituent cat=''TOP''>                    non-terminal nodes
      <constituent cat=''NP-TOP''>
        <constituent cat=''PN-NK-Nom.Sg''>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t2''/></constituent>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t3''/></constituent>
        </constituent>
        <constituent cat=''$,''><tokenRef tokID=t4"/></constituent>
        <constituent cat=''NP-APP''>
          <constituent cat=''ART-NK-Nom.Sg''><tokenRef tokID=''t5''/></constituent>
          <constituent cat=''NE-NK-Nom.Sg''><tokenRef tokID=''t6''/></constituent>
        </constituent>
      </constituent>
      <constituent cat=''$.''><tokenRef tokID=''t7''/></constituent>
    </constituent>
  </parse>
  <parse>
     ...
</parsing>
```

# D-SPIN Text Corpus Format (TCF)

## Sentence layer

```
<parsing tagset=''TigerTB''>
  <parse>
    <constituent cat=''TOP''>
      <constituent cat=''NP-TOP''>
        <constituent cat=''PN-NK-Nom.Sg''>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t2''/></constituent>
          <constituent cat=''NE-PNC-Nom.Sg''><tokenRef tokID=''t3''/></constituent>
        </constituent>
        <constituent cat=''$,''><tokenRef tokID=t4"/></constituent>
        <constituent cat=''NP-APP''>
          <constituent cat=''ART-NK-Nom.Sg''><tokenRef tokID=''t5''/></constituent>
          <constituent cat=''NE-NK-Nom.Sg''><tokenRef tokID=''t6''/></constituent>
        </constituent>
      </constituent>
      <constituent cat=''$.''><tokenRef tokID=''t7''/></constituent>
    </constituent>
  </parse>
  <parse>
    ...
</parsing>
```

terminal nodes

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

- Converters from/to existing formats
  - tokenizers, taggers, NER, (Berlin, Leipzig, Stuttgart)
  - treebank formats:
    NeGra, Tüba-D/Z (Tübingen)
    TiGer: to be done yet
  - general formats: PAULA (Tübingen)
- Converters from/to upcoming standards
  - MAF (Stuttgart)
  - experiments with LAF (DFKI/Stuttgart)

$\Rightarrow$ TCF and MAF/SynAF are structurally close

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

**MAF**

```
<token id=''t2''>The</token>

<wordForm lemma=''the''
tag=''pos.det'' tokens=''t2''/>
```

**TCF**

```
<token ID=''t2''>The</token>

<tag tokID=''t2''>det</tag>

<lemma tokID=''t2''>the</lemma>
```

part-of-speech

lemma

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

| **MAF** | | **TCF** |
|---|---|---|
| `<token id=''t2''>The</token>` | | `<token ID=''t2''>The</token>` |
| | part-of-speech | `<tag tokID=''t2''>det</tag>` |
| `<wordForm lemma=''the'' tag=''pos.det'' tokens=''t2''/>` | | |
| | lemma | `<lemma tokID=''t2''>the</lemma>` |

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

| **MAF** | | **TCF** |
|---|---|---|
| `<token id=''t2''>The</token>` | | `<token ID=''t2''>The</token>` |
| | part-of-speech | `<tag tokID=''t2''>det</tag>` |
| `<wordForm lemma=''the'' tag=''pos.det'' tokens=''t2''/>` | | |
| | lemma | `<lemma tokID=''t2''>the</lemma>` |

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

**MAF**

```
<token id=''t2''>The</token>

<wordForm lemma=''the''
tag=''pos.det'' tokens=''t2''/>
```

**TCF**

```
<token ID=''t2''>The</token>

<tag tokID=''t2''>det</tag>

<lemma tokID=''t2''>the</lemma>
```

part-of-speech

lemma

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

**MAF**

```
<token id=''t2''>The</token>

<wordForm lemma=''the''
tag=''pos.det'' tokens=''t2''/>
```

**TCF**

```
<token ID=''t2''>The</token>

<tag tokID=''t2''>det</tag>

<lemma tokID=''t2''>the</lemma>
```

part-of-speech

lemma

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

Example: MAF → TCF mapping

- Extract information about the words from MAF

| **MAF** | | **TCF** |
|---|---|---|
| `<token id=''t2''>The</token>` | | `<token ID=''t2''>The</token>` |
| | part-of-speech | `<tag tokID=''t2''>det</tag>` |
| `<wordForm lemma=''the'' tag=''pos.det'' tokens=''t2''/>` | | |
| | lemma | `<lemma tokID=''t2''>the</lemma>` |

⇒ Mapping TCF information about token, pos and lemma to MAF is also possible, but converter is not existing yet

# D-SPIN Text Corpus Format (TCF)

## Relating TCF with existing formats and with upcoming standards

MAF representations not (yet) covered

- n-to-m relations between wordForms and tokens
  → could be replaced
  by TCF 1-to-1 relations by redefining the tokens in the converter
- discontinuous / nested wordForms:
  *setzt den Hut ab / Geburtstagsgeschenkpapier*
- ambiguous wordForms / tokens: *sichere/ADJ/V*
- POS tags represented as feature structures (TCF: atomic POS tags)
- lexicon references
- normal forms, phonetic forms:
  `<token form=''...'' phonetic=''...'' ...`

⇒ Extend TCF, create new elements like `<phonetic/>`
if needed in WebServices

# Convergence on descriptive vocabularies

## D-SPIN work on tagsets

- Need to reinterpret existing tagsets
  in a way to map them onto other tagsets
  - by analysis in terms of feature structures
  - by trying to define relationships:
    - identity
    - overlap
    - disjointness
- Recently: start of mapping of STTS
  (Stuttgart-Tübingen TagSet) onto ISOcat/DCR
  $\Rightarrow$ More such work is needed and planned

# Convergence on descriptive vocabularies

## D-SPIN work on tagsets: an example

- ISOcat POS tagset, for nouns:
  - commonNoun, properNoun
  - diminutiveNoun, ...
- STTS POS tagset:
  - NN (normales Nomen), NE (Eigenname)
- ISOcat definitions, by classes and values:
  - nounClass [+ definition from ISO-12620]
    - commonNoun: (example: continent)
      [Def.: ... denoting a class of objects ...]
    - properNoun: (example: Europe)
      [Def.: ... denoting a single object...]

$\Rightarrow$ Intellectual work on mapping

# Conclusion

- International standards for linguistic representation are evolving: D-SPIN contributes to their development
- D-SPIN formats on three levels:
  - tool specific
  - web service-related (TCF):
    - mappings to/from de facto standards for German
    - mapping to/from international standards
  - LAF/MAF/SynAF as exchange format: Then TCF/LAF...-Mapping has a double function