SynAF Annotation FrameWork Background, MetaModel and DataCatgeories

►Thierry Declerck (DFKI, Saarbücken)

LIRICS IAG Workshop, Paris 10/05/2007

Brand New:

"Not psycho-linguistics, socio-linguistics or another X-linguistics, but:

ISO-linguistics“
(Thorsten Trippel)

Standards for Language Resources

- Heterogeneous Resources:
  - Linguistics
  - Language Technologies
  - Translation
  - Lexicon development
- Idiosyncratic data formats
- Problems:
  - Data exchange
  - Re-usability

Standards for Language Resources

- Possible Solution:
  - Standardization of Language Resources:
    - Best practice
    - Portability
- ISO TC 37/SC4 (Management of Language Resources)
  - DIN NA 105 Normenausschuss Terminologie (NAT), NA 105-00-06 AA in Germany, and other national bodies involved
  - The LIRICS project, promoting the whole line of standards under development in ISO TC37/SC4

Overview of Activities in ISO TC37/SC4 and DIN NAAT6

- Standards for Language Resources
  - Linguistic Annotation Framework (LAF)
  - Word Segmentation
  - Morpho-Syntactic Annotation Framework (MAF)
  - Lexical Markup Framework (LMF)
  - Feature Structure Representation (FSR)
  - Syntactic Annotation Framework (SynAF)
  - Semantic Annotation Framework (SemAF)
  - Data Categories (DatCats)

Linguistic Annotation Framework (LAF)

- Goal: Unitary base for the annotation of Linguistic Data
  - XML based, incl. Semantic Web representation languages
  - Stressing on higher level of annotation
- Content of LAF: A generic Data Format
  - Based on results of ISLE/EAGLES, TEI
- State: At the beginning of the ISO Procedure (WD)

Word Segmentation

- A must in automated language processing
- Problem:
  - Not always by blanks
  - Treatment of compounds
  - Evaluation of tools/processing strategy missing
- Goal: A Metamodel for Segmentation (for the time being in MAF, but to be extended in a new Work Item)
  - Word property of Multi-Word-Expressions
  - Linguistic rules

Morpho-Syntactic Annotation Framework (MAF)

- Goal: Unitary codification of morpho-syntactic annotation
- Content of MAF:
  - Segmentation
  - Content description of the annotation
- State: At an advanced level in the ISO procedure (DIS)

Lexical Mark-Up Framework (LMF)

- Goal: Exchange Format for lexical databases Stressing on higher level of annotation
  - Similar to ISO 12200 (Martif)
  - Including dictionaries
- Content: Unitary Model for representation of dictionaries and (computational) lexicons
- State: Just before being adopted

Feature Structure Representation (FSR, ISO 24610-1)

- Syntax for defining feature structures
  - e.g. HPSG lexicons
- in XML
- closely related to previous TEI work
- final since last year

Feature system declaration(FSD, ISO 24610-2)

- declaration scheme for feature structures
- with ISO 12620
- and ISO 24610-1
- WD status

Syntactic Annotation Framework (SynAF)

- Goal: Propose a meta model for syntactic annotation
- Content: Constituency and Dependency structures
- State: Submitted as a NewWork Item by the LIRICS Consortium, now at CD stage

Semantic Annotation Framework (SemAF)

- Goal: Propose a meta model for semantic annotation
- Content: Semantic Roles, Reference, Temporal expressions, Dialogues, ...)
- State: Submitted as a NewWork Item by the LIRICS Consortium. Part 1 on Temporal Annotation started in October 2006. Starting point: TimeML and OWL-Time

Data Categories (DatCats)

- Goal: Propose (neutral) definition of data categories that subsume various encodings and formats used in different systems.
- Content: Analog to ISO 12600 for terminology
- Lirics is proposing such lists for Lexicons, MAF, SynAF, SemAF Discussion: open list of DatCats (Control?)

The Background: Linguistic Annotation Framework (slides from Nancy Ide)

- Under development within ISO TC37/ SC4 (Language Resource Management)
- Intended to provide standardized means to represent linguistic data and annotations

LAF Approach

- Develop a common, abstract model that can capture all types of annotation information, regardless of the physical encoding
- Develop a generic, XML instantiation of the model, to and from which specific formats can be mapped
- Define a common set of data categories, for reference and use by annotators

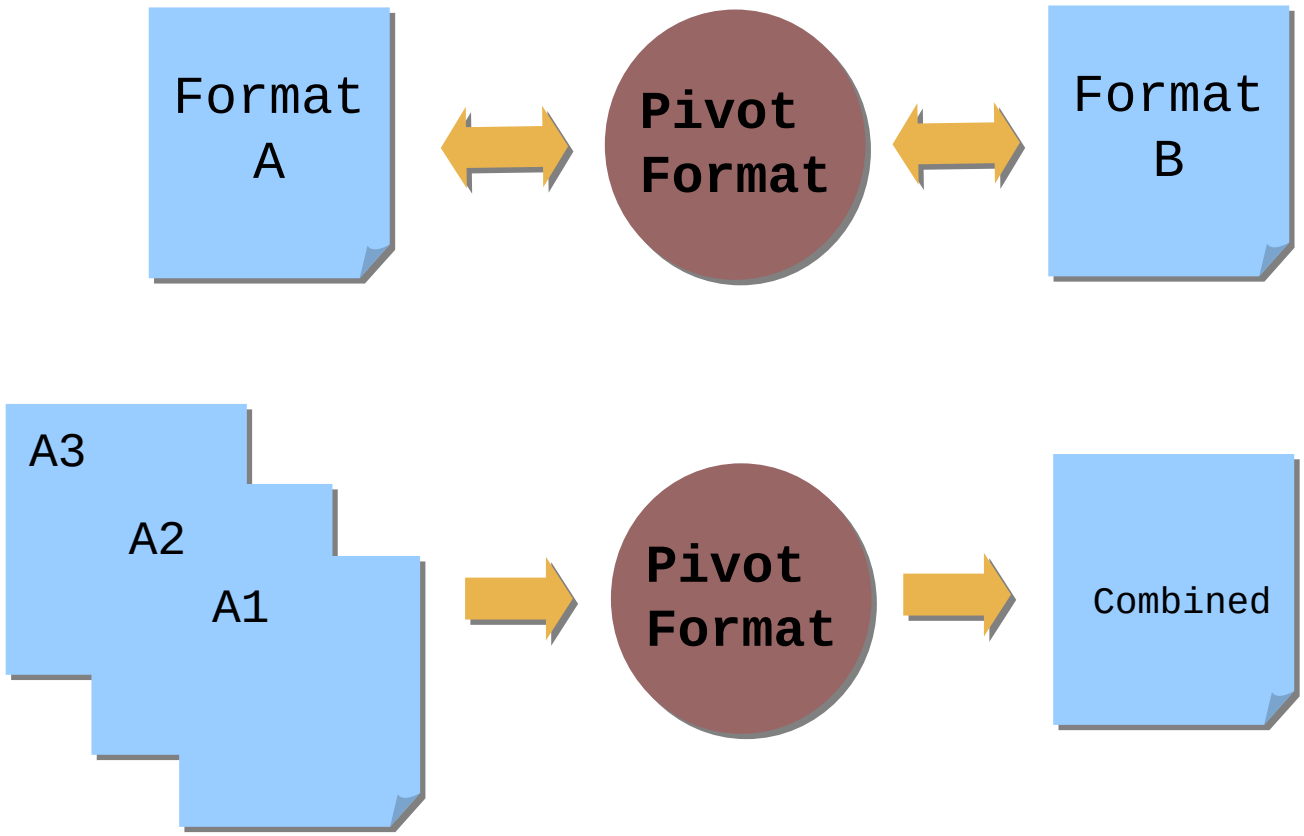
General Principles

- Separation of data and annotations
- Separation of user annotation formats and the exchange (“pivot”) format
- Separation of annotation structure and content in the pivot format

Abstract Model

- Annotations represented as a graph of feature structures
  - Nodes are locations in primary data or other annotations
- Any format instantiating the model can be trivially mapped to another format via the pivot format

Overview

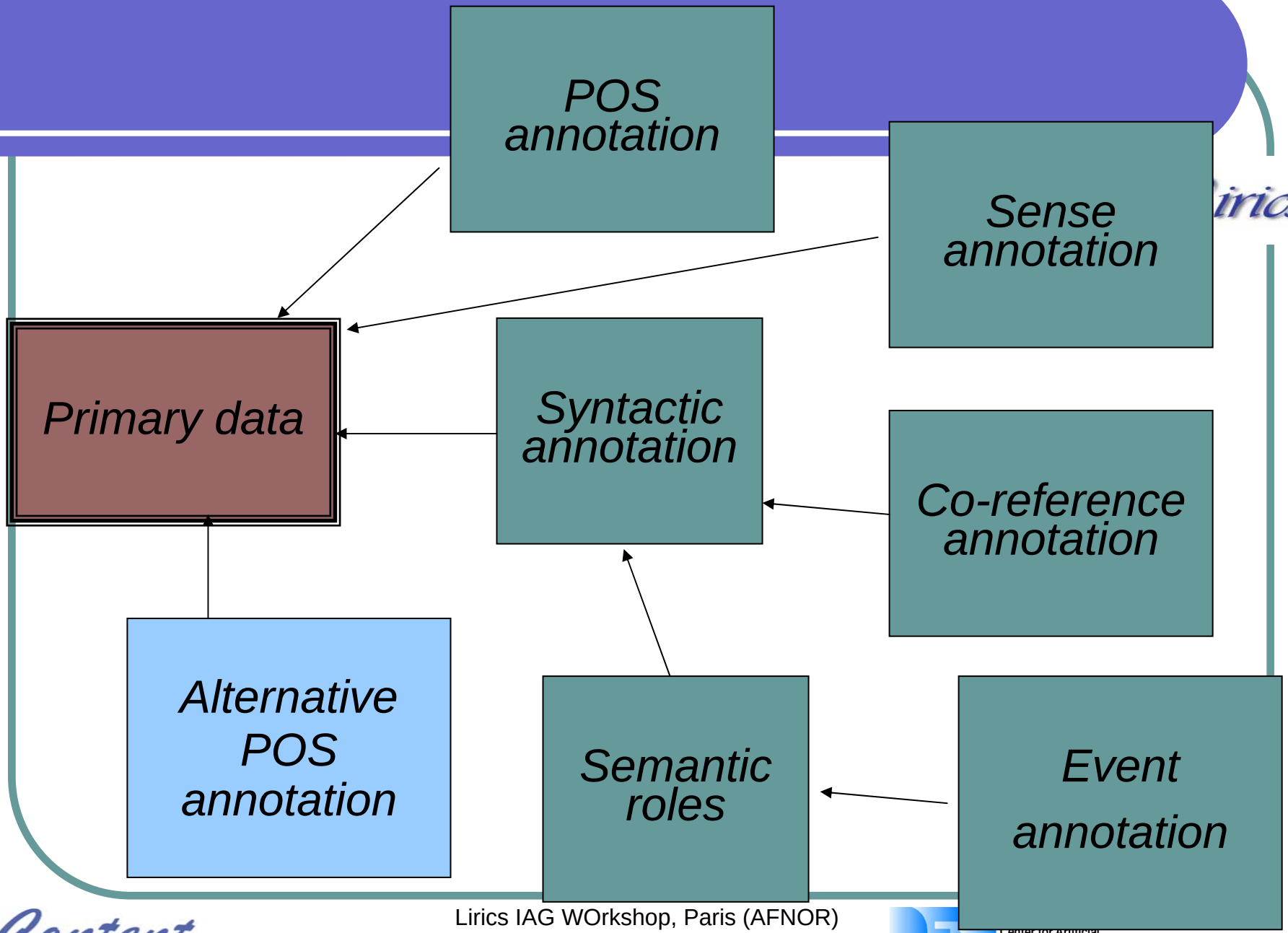

Format A
Pivot Format
Format B
A3
A2
A1
Pivot Format
Combined

Pivot Format

- Need never be seen/used by user
  - In principle, user defines “mapping rules” and pivot is automatically generated (and vice versa)
  - Exchange format
  - Model used to enable mapping, also to inform design of new annotation schemes

Pivot: Stand-off Annotation

- Language data is regarded as “read-only” and contains no annotations
- Annotations are stand-off linked to the primary data or other annotation documents


POS annotation
Sense annotation
Primary data
Syntactic annotation
Co-reference annotation
Alternative POS annotation
Semantic roles
Event annotation

Data Category Registry

- Addresses issue of standardization of annotation content
- Provides a set of reference categories onto which scheme-specific names can be mapped
- Provides a precise semantics for annotation categories
- Provides a point of departure for definition of variant, more precise, or new data categories

Exchange Specification

- Annotator provides a Data Category Specification (DCS)
  - mapping between scheme-specific instantiations and concepts in the DCR
    - Including differences, departures, new categories
  - provides documentation for the user's annotation scheme
- DCS included or referenced in data exchange
  - provides receiver with information to interpret annotation content or map to another instantiation
  - semantic integrity guaranteed by mutual reference to DCR concepts or definition of new categories in DCS

Pivot Format Design

- Primary concerns
  - Maximize processing efficiency and consistency
  - Ensure that processing is unambiguous
  - Instantiate with a simple, minimal set of elements
- Fulfillment of these requirements has repercussions for users
  - Information must be explicitly provided in their representations or made explicit via the mapping
- N.B.:
  - Only requirement is that user format can be mapped to the spec

Segmentation

- Minimal unit of granularity
- Points to virtual nodes characters in primary data
- May have multiple segmentations over the same data
- No associated annotation content (at this level)
- Set of linearly ordered edges

Annotations

- Annotation label
- May be data category in DCR

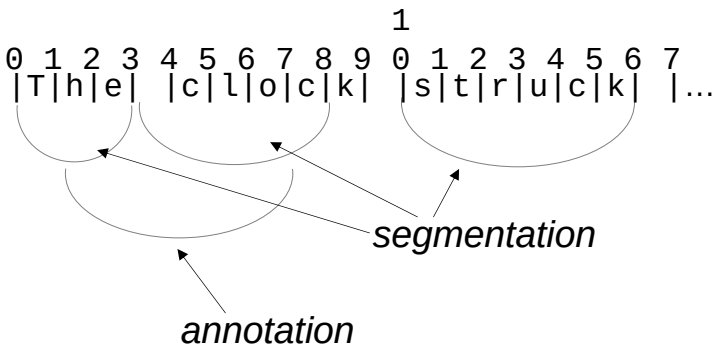
- Link label pointing to object(s) of the annotation (idref)
- Link label may be data category in DCR

- Feature structure content providing annotation information
- Attribute-value pairs
- Recursive
- Can specify alternatives etc.

Annotation Layers

- Conceptual layers of annotation
  - E.g. morpho-syntax, syntax, co-reference…
  - ISO TC37/SC4 defining a set of layers
- Each layer has a schema defining the relevant categories and relations
  - E.g. syntax
    - Category: Sentence
    - Relations: SUBJ (Object: NP), MainVerb (Object: VP), “Constituent” (Object: NP | VP | PP)
- Inter-layer and cross-layer relations

Example



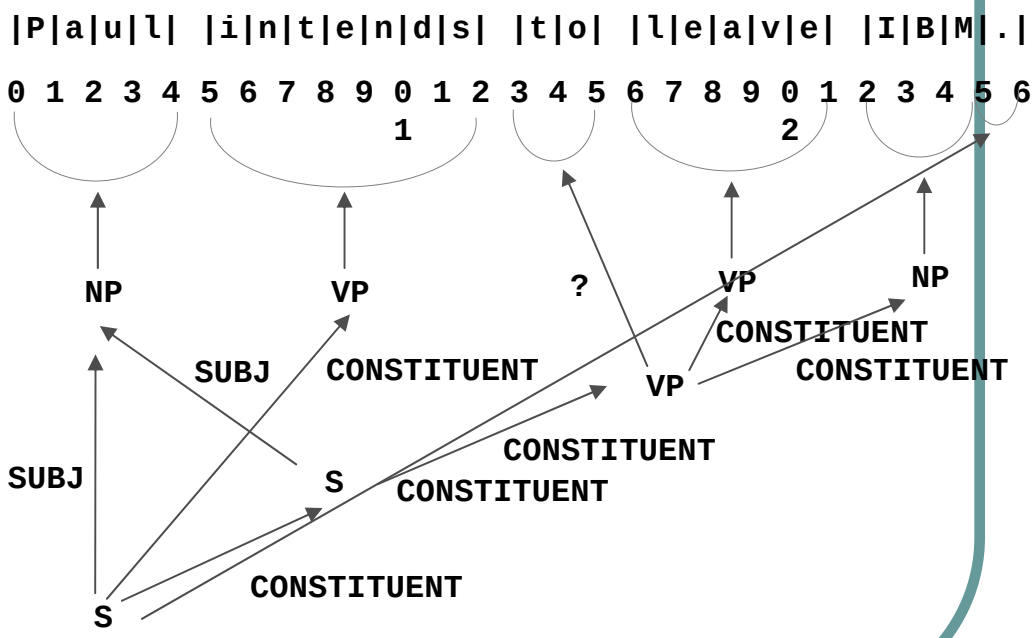

```
<!-- edges over primary data -->
<edge id="e1" from="0" to="3"/>
<edge id="e2" from="4" to="9"/>
<edge id="e2" from="10" to="16"/>
```

```
<!-- msd layer annotation -->
<edge id="t2">
   <cat name="token"/>
   <seg ref="e2"/>
   <fs>
      <f name="lemma" sVal="clock"/>
      <f name="pos" sVal="NN"/>
   </fs>
</edge>
```

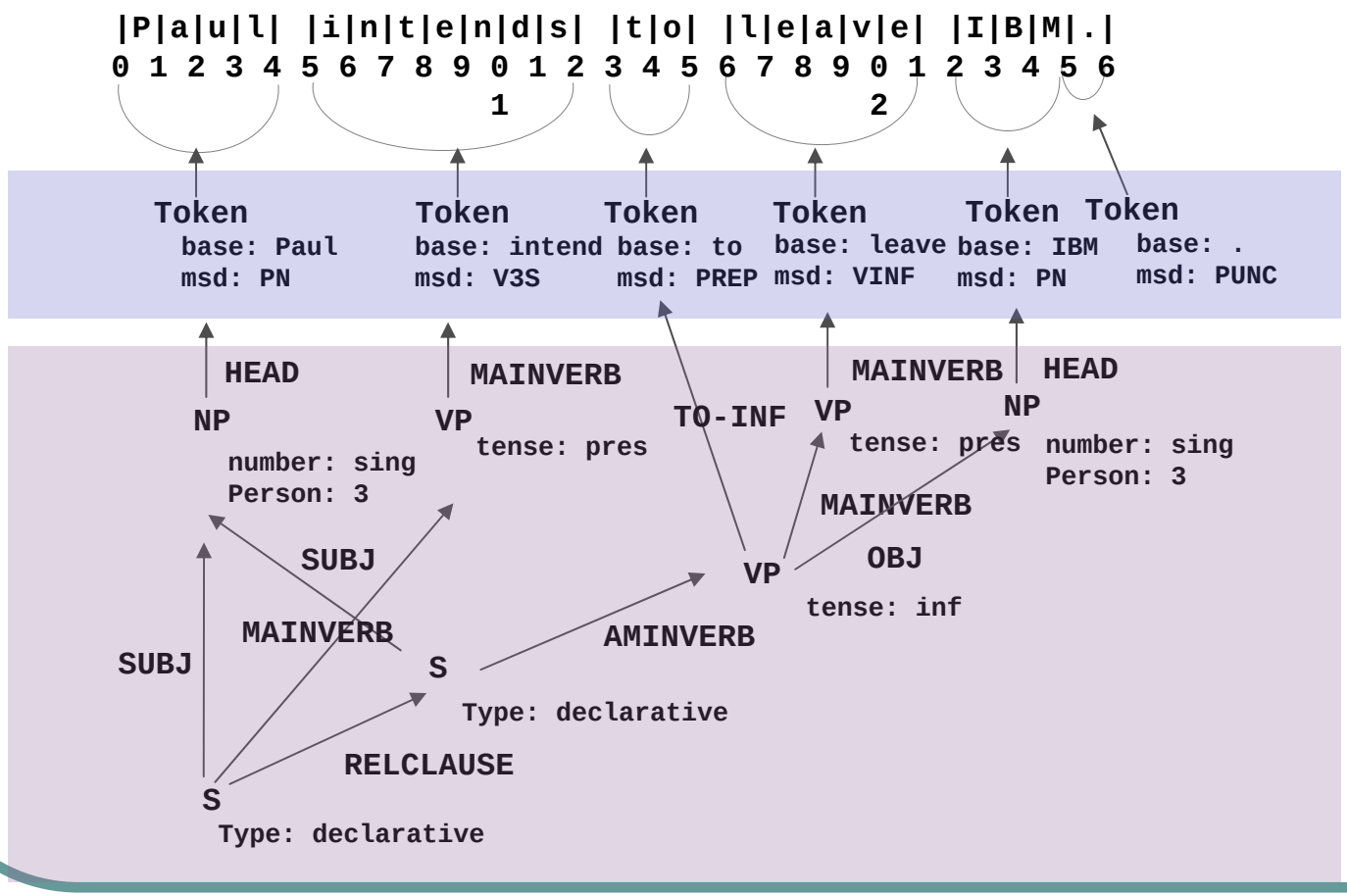
```
<!-- Syntactic layer annotation -->
<edge id="np1">
   <cat name="NP">
   <rel type="det" target="t1"/>
   <rel type=”head" target="t2"/>
   <fs>
      <f name="number" sVal="singular"/>
   </fs>
</edge>
```

Mapping to the Pivot Format

```
((S
    (NP-SBJ-1 Paul)
    (VP intends)
    (S
       (NP-SBJ  -1)
       (VP    to
       (VP leave)
       (NP IBM ) ) ) )
.))
```



Ideal Result?



Segmentation

Morpho-syntact layer

Syntactic layer

Goals

- Reference categories in DCR rather than give cats
- Reference FS fragments and schema layer definitions in on-line libraries
- Annotation schemes designed/modified to conform to the model

Summary (end of slides by Nancy Ide)

- Model still evolving
  - Precise pivot XML not fixed
- Basic principles/ideas already appearing in applications/schemes
- Mapping to pivot will be simple, straightforward

Data categories (ISO/DIS 12620)

- Data categories needed for
  - termbases
  - lexicons
  - corpus description
  - linguistic description
- Problem
  - major portion identical in all areas
  - deviations in detail
  - two researcher, three sets of data categories
  - homonyms

Data categories (ISO/DIS 12620) cont

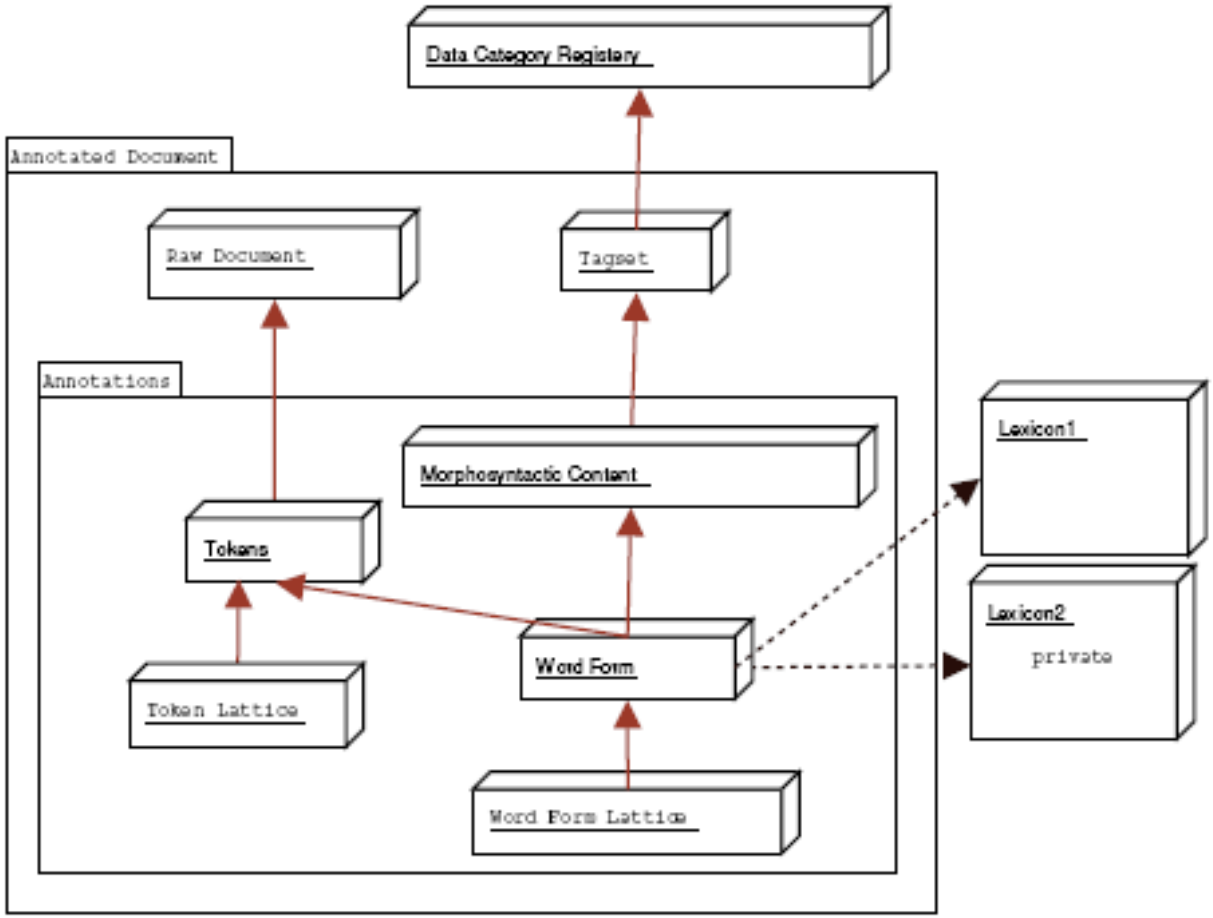
- Idea:
  - no fix of data categories!
  - data category repository
    - need a category, take a category,
    - have a category, give a category
  - central registration repository
    - automatized registration
    - comparison of existing categories
    - if in doubt: "fast" ballot, DCR board

Data category repository

- standard = open
  - not (never!) finished
  - method defined
- requires infrastructure
  - hosted by MPI
  - archiving task

MAF: A brief survey

MAF: MetaModel



MAF: Morpho-Syntax and Content

- <**wordForm** e n t r y ="manger " t o k e n s=" 0 ">
- < **f s** >
- < **f** name="mode">
- <symbol v a l u e=" imp e r a t i v e " / >
- < / **f** >
- < **f** name=" number ">
- <symbol v a l u e=" s i n g u l a r " / >
- < / **f** >
- . . .
- < / **f s** >
- < / **wordForm**>

The SynAF Working Draft

- SynAF (Syntactic Annotation Framework) has been adopted by ISO as a NWI, and is now as a WD close to be submitted as a Committee Draft (CD). For reference:
  - Project number: 24615
  - Project abbreviation: SynAF
  - Project leader: Thierry Declerck, DIN
  - WG: ISO/TC 37/SC 4/WG 2 Representation schemes
- SynAF is partly based on MAF (Morpho-Syntactic Annotation Framework) and will propose a base for future standardisation of (linguistic) semantic annotation.

Topic of SynAF

- SynAF is dealing with the description of a meta-model for syntactic annotation, which means that SynAF will describe elementary linguistic (in fact syntactic) abstractions that support the construction and the interoperability of (syntactic) annotations and resources, as well as the procedure for the creation of data categories for syntactic annotation.
- SynAF is thus not proposing a tagset for syntactic annotation, but is dedicated to proposing a (possibly hierarchical) list of data categories, which is much easier to update and extend, and which will represent a point of reference for particular tagsets used for the syntactic annotation of various languages, also in the context of various application scenarios.

Basis for SynAF

- Corpus (Linguistic) Annotation Frameworks that combine syntactic constituency and syntactic dependency
  - Tiger for Germany
  - ISST for Italian
  - Similar resources for other languages (see D3.1)
- Grammar Resources
  - Parsing output syntactic structures for various languages (HPSG, LS-GRAM Project, LFG parallel grammars, shallow grammars etc.)

The SynAF Proposal

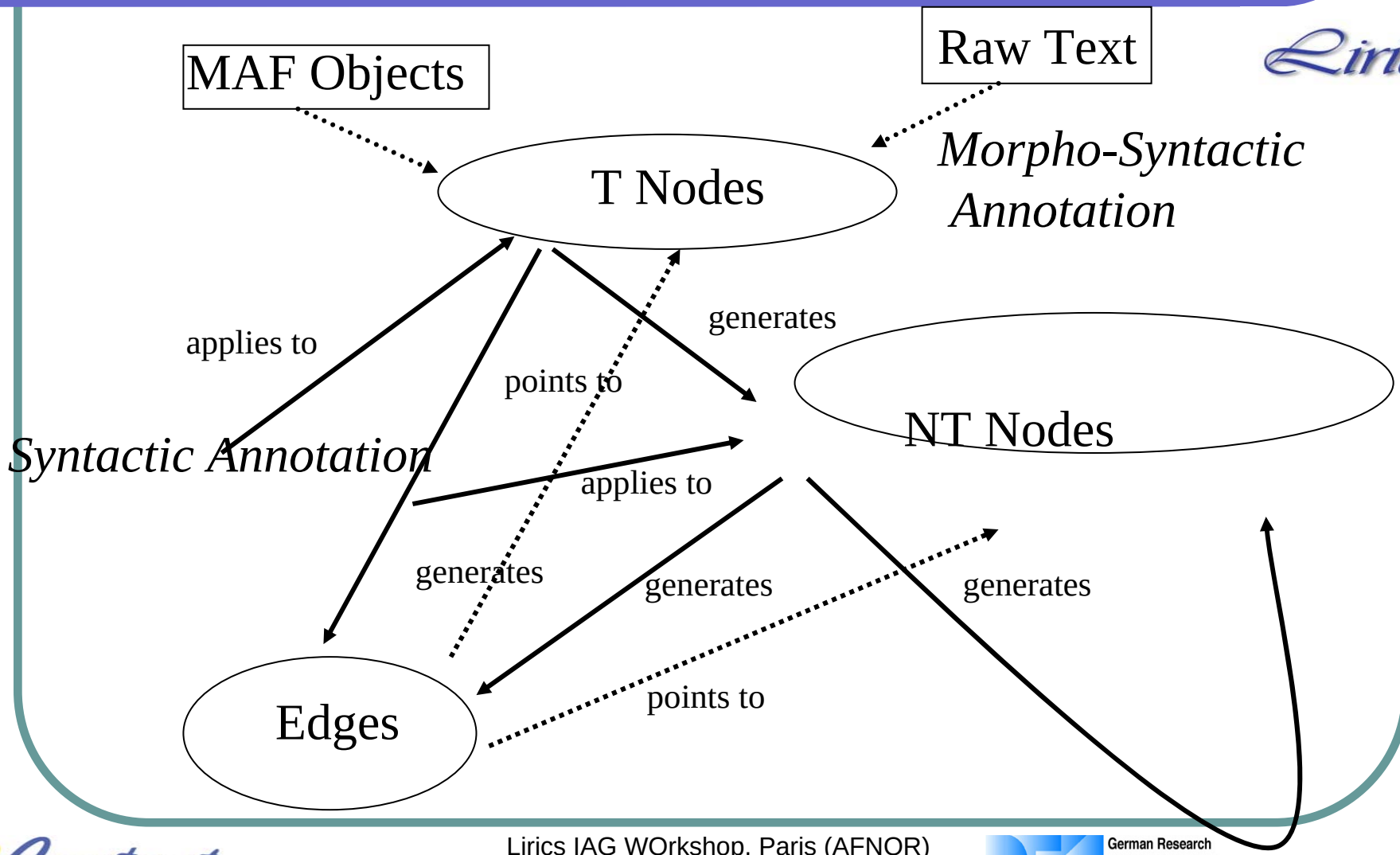
Syntactic Annotation has 2 Functions in NLP

- 1) To represent linguistic constituencies, like Noun Phrases (NP), describing a structured sequence of morpho-syntactically annotated items, where we consider also constituents built from non-contiguous elements, and
- 2) To represent dependency relations dependency information can exist between morpho-syntactically annotated items within a phrase (an adjective is the modifier of the head noun within an NP) or describe a specific relation between syntactic constituents at the clausal and sentential level (i.e. an NP being the "subject" of the main verb of a clause or sentence). In the first case we speak of an *internal dependency* and in the second case we speak of an *external dependency.* But the dependency relation can also be stated including empty elements (like the pro-drop property in romance languages)

The SynAF Proposal (2)

SynAF is concerned thus with a meta-model that covers both dimensions of syntactic *constituency* and *dependency*, and SynAF is proposing a multi-layered annotation framework that allows the combined and interrelated annotation of language data along both lines of consideration. Also the data-categories to be proposed to ISO standardization will be about the basic annotation concerning both dimensions.

The SynAF Model: A first Draft



Some Remarks

- SynAF is about Graphs. *Nodes* in a syntax graph can be both terminals (word-forms) and non-terminals (constituents). Nodes can be interrelated by *Edges* (between source and target nodes). This definition supports discontinuity (a constituent can be described by multiple edges)
- At the representation level: List of edges within the corresponding nodes or separated? If separated then the edge information mentions explicitely source and target of the edges (cleaner from the point of view of algebra).

Comments... (2)

- Also encoding of initial and ending points of a node is supporting underspecification of annotation and the description of empty elements (start point = end point, naming the point where the empty element is belonging to).
- Need to define coherence condition on the paths one can build with edges.

Some Definitions

- Span: a pair of points identifying a segment of the document submitted to syntactic annotation. The first point <= the second point.
- Multiple span: A sequence of spans where the ending point of each span <= the starting point of the subsequent span.
- Category: a feature value providing the content of a node.
- Node: pair consisting of a (possibly multiple) span, a category,
- Edge: a triplet with a source node, a target node, and a label. Non-Terminal nodes have an outgoing constituency edges (to be defined)
- Label: a feature value providing the content of an edge.
- A Terminal node: refers to a single wordFrom/lexical unit or a span with lentgh=0, and the node and the wordForm/lexical unit have identical span.

Putting SynAF in XML

- A Terminal node: refers to a single wordFrom/lexical unit, and the node and the wordForm have identical span T

```
<terminal
    <category name="$CATEGORY_DatCat“ span=„$DIGIT -
$DIGIT“
    </cateogry>
    <edges>
        <edge label="$LABEL_DatCat„
              sourcenode=„$SourceNode“
              targetnode=„$TargetNode“
        </edge>
    </edges>
</terminal>
```

Putting SynAF in XML (2)

- Non-Terminal nodes have at least one outgoing constituency edges (to be defined)

```
<nonterminal
            <category name="$CAT_DatCat“ span=„$DIGIT -
$DIGIT“></category>
        <edges>
            <edge label="$LABEL_DatCat„
                  sourcenode=„$SourceNode
                  targetnode=„$TargetNode“></edge>
            <edge label="$LABEL_DatCat„
                  sourcenode=„$SourceNode
                  targetnode=„$TargetNode></edge>
        </edges>
</nonterminal>
```

Data Categories for SynAF

- We need 2 types of data categories in SynAF
  - Namimg the nodes (constituents), for example: noun phrase (NP), proper noun (PN), adpositional phrase (PP), etc.
  - Naming the labels (dependencies), for example: head (HD), modifier (MOD), accusative object (AO) or subject (SB), etc

Data Categories for SynAF (Consistency)

- Naming the nodes (constituents)

Data Categories for SynAF (Dependency)

- Naming the labels (dependencies)

Issues for SynAF

- Level of complexity: deal only with the intersection of syntactic phenomenons that are present in all (or most) languages vs. an almost complete list of phenomena describing language dependant phenomea in details.
- Closely related: monolingual description vs. multilingual descriptions. Cross-lingual aspects: for example including in the annotation information taht supports translation?)
- Surface syntactic phenomena vs. „deep“ lingusitic phenomena (including transformation, movement, lexical rules)
- Etc...

Conclusions

- A lot of common activities for proposing standards in the domain of language resources, from which we hope that they will facilitate cooperation in Europe and the take up of commercial/industrial activities, on the base of the ISO framework for interoperability of linguistic resources.

And now...

- Which Standard do you need?