Standards for language encoding: ISO

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute

ESSLLI 2011

Overview of the lecture

1. How ISO works
2. ISO TC 37
3. Dates, times & languages
4. ISO Annotation Frameworks

International Organization for Standardization

http://www.iso.org/

- The world's **largest developer** and publisher of **International Standards**, founded in 1947.
- A **network** of the national standards institutes of **162 countries**
- Central Secretariat in Geneva coordinates the system
- **non-governmental organization** but strong links to goverments via national institutes
- bridge between the public and private sectors.
- ISO enables a **consensus** to be reached on solutions that meet both the requirements of business and **the broader needs of society**.
- Somewhat problematic business model:
  standards are published on paper and have to be bought
  - however, many standards (or at least the data category registries) are also freely available on the Web via other organisations.

National standard bodies

- National standards bodies: ANSI, DIN, SIST, …
- Adopt national standards
- National standards and ISO:
  - adoption
  - translating the title
  - translating the full standard
- Propose delegates for membership in particular working groups of ISO
  - Voting member
  - Observer
  - not free (for national bodies) and not paid (to delegates)

The birth (and death) of an ISO standard

- Standards are proposed, commented, voted on, and adopted (or not…); also withdrawn

In short:

1. ISO/PWI: proposed work item
2. ISO/AWI: approved work item
3. ISO/CD: committee draft
4. ISO/DIS: draft international standard
5. ISO/FDIS: final draft international standard
6. ISO: international standard

- Official stages (number codes)
- Show example of an ISO standard

Structure of ISO

- ISO **Technical Committees** (TC) are composed of members from participating countries, who then propose, develop, comment, and approve standards from their field
- ISO TCs are further composed of
  - **Sub-Committees** (SC)
  - and these contain **Working Group**s (WG)
  - e.g. ISO TC 37 SC 3 WG 4

ISO TC 37

- Technical Committee on Terminology
- Important for all other standards: each standard must contain a section on terminology
- In 2001 name and scope of TC 37 changed to: Technical Committee on Terminology **and other language and content resources**
- Subcommittees:
  - TC 37/SC 1 Principles and methods
  - TC 37/SC 2 Terminographical and lexicographical working methods
  - TC 37/SC 3 Systems to manage terminology, knowledge and content
  - **TC 37/SC 4 Language resource management**

II. Standards discussed @ this lecture

- Dates and times
- ISO TC 37 standards for
  - Language codes
  - Feature structures
- ISO TC 37 SC4 standards and proposals for Annotation Frameworks
  - LAF, LMF, MAF, SynAF, SemAF
  - Data category registries

Dates and Times

- ISO 8601 *Data elements and interchange formats – Information interchange – Representation of dates and times*
- Dates: *1984, 1984-04, 1984-04-04*
- Times: *13:00:00, 1984-04-04T13:00*
- UTC - Coordinated Universal Time (GMT)
- Timezones:
  - GMT: *1984-04-04T13:00Z*
  - GMT+1: *1984-04-04T13:00+01*
- ISO 8601 also gives formats for durations and intervals

Language codes

- ISO 639 is the set of international standards that lists short codes of two to four letters for language names.
- When referring to languages in a computational setting, one should always use language codes defined in ISO 639

Not just one, but many..

Standard	Name (*Codes for the representation of names of languages – ...*)	First edition	Current	No. in list
ISO 639-1	*Part 1: Alpha-2 code*	1967 (as ISO 639)	2002	184
ISO 639-2	*Part 2: Alpha-3 code*	1998	1998	>450
ISO 639-3	*Part 3: Alpha-3 code for comprehensive coverage of languages*	2007	2007	7704 + local range
ISO 639-4	*Part 4: Implementation guidelines and general principles for language coding*	2010-07-16	2010-07-16	(not a list)
ISO 639-5	*Part 5: Alpha-3 code for language families and groups*	2008-05-15	2008-05-15	114
ISO 639-6	*Part 6: Alpha-4 representation for comprehensive coverage of language variants*	2009-11-17	2009-11-17	?

ISO 639-1 (Alpha 2)

- 184 languages
  - English is represented by en
  - French is represented by fr
  - Italian is represented by it
  - Portuguese is represented by pt
  - German is represented by de (from the endonym *Deutsch*)
  - Spanish is represented by es (*español*)
  - Swedish is represented by sv (*Svenska*)
  - Japanese is represented by ja (but endonym is *Nihongo*)
- N.B. do not confuse language codes with country codes (ISO 3166)
  - e.g. Slovene language is “sl”, Slovenia is “si”

SIL database

- You can buy the text of ISO 639..
- But the lists of language codes are freely accessible:
  - SIL (also fonts, software etc.)
  - Ethnologue (much info on languages)
  - Wikipedia

Standards for Language resource management

- ISO TC 37/SC 4: Subcommittee for Language resource management
- http://www.tc37sc4.org/
- Aim: Prepare standards that specify principles and methods for creating, coding, processing and managing language resources. These standards will also cover the information produced by natural language processing components.
- Problems:
  - “Proactive” – for many areas there is no consensus on the theoretical level
  - Standard first, then applications (validation)

TC 37 / SC4 Working Groups

WG 1: Basic descriptors and mechanisms for language resources

- Terminology used in LRs, basic mechanisms & data structures, meta-data representation scheme for linguistic information

WG 2: Representation schemes

- annotation/representation schemes for morpho-syntax, syntax, semantic content, discourse

WG 3: Multilingual text representation

- Translation memory and alignment of parallel corpora, segmentation and counting algorithms, meta-markup for internationalization and localization

WG 4: Lexical databases

- Lexical representation formats for the various types of NLP applications

WG 5: Workflow of language resource management

- Guidelines for language validation and net-based distributed cooperative work

Feature structures

- Record-field representation of (linguistic) information
- Recursive, co-indexing, (types), constraints
- FSs are a basic datatype in unification-based grammars
- Taken from TEI:
  - ISO 24610-1:2006 Language resource management -- Feature structures -- Part 1: Feature structure representation
  - ISO/FDIS 24610-2 Language resource management -- Feature structures -- Part 2: Feature system declaration

AVM and ISO-FS XML representations

$$\begin{bmatrix} \textit{syn-cat} & \\ \text{ORTH} & \text{'walks'} \\ \text{HEAD} & \begin{bmatrix} \textit{verb\_st} \\ \text{FORM finite} \end{bmatrix} \\ \text{VALENCE} & \begin{bmatrix} \textit{val-cat} \\ \text{SPECIFIER} \langle \boxed{1}\ \text{NP} \rangle \\ \text{COMPS} \langle\ \rangle \end{bmatrix} \\ \text{ARG-ST} & \langle \boxed{1} \rangle \end{bmatrix}$$

N.B. AVM is not the same as XML!

```
<fs>
  <f name="orth">
    <string>love</string>
  </f>
   <f name="syntax">
     <fs>
        <f name="pos">
          <symbol value="verb"/>
        </f>
        <f name="valence">
          <symbol value="transitive"/>
        </f>
      </fs>
     </f>
  </fs>
```

LAF – overall concept of *AF

- ISO/DIS 24612 Language resource management -- Linguistic annotation framework (LAF)
- Basic principles of linguistic annotation
- Types of Annotation:
  - **segmentation**: delimits linguistic elements that appear in the primary data (continuous, super-, sub-, discontinuous, landmarks)
  - **linguistic annotation**: provides linguistic information about the segments in the primary data
- Use of stand-off annotation

LAF data model

- The abstract data model
  - feature-structure graph with some operators (disjunction)
  - written in UML (Unified Modeling Language)
- The Dump-format(s)
  - a concrete instantiation of the data model
  - usu. expressed in XML

DCR: Data Category Registries

- The idea that there should be public and stable repositories of (linguistic) data categories
- GOLD: General Ontology for Linguistic Description
  "encompasses linguistic concepts, definitions of these concepts and relationships between them in a freely available ontology"
- MULTEXT-East: LRs for (Slavic) languages
  - Morphosyntactic specifications

ISO DCR & ISOcat

- ISO 12620:2009 Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources
- ISOcat is the reference implementation of ISO 12620:2009
- Most Annotation Framework standards require that the categories used are registered at ISOcat

ISOcat

ISOcat - Web interface - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Course: Standards ... | Program | ISO/AWI 24617-4 -... | ISO - About ISO | Working Group | ISOcat - Data Cate... | ISOcat - ISO 12620 | ISOcat - Web interf...

http://www.isocat.org/interface/index.html

ISO 12620

Google Koledar | ESSLLI 2011 | Program | Reviewer's Center: Wel... | IntelliText Corpus Que... | CLDR - Unicode Com... | ISO Standards Develo... | IMPACT@JSI | ESSLLI 2011 | tomaz erjavec

ISOcat

Welcome Guest | Help

enter keywords here

- My Workspace
  - Public
    - Thematic Views
      - Metadata
      - Morphosyntax
        - Morphosyntax
        - Basics
        - Cases
        - FormRelated
        - MorphologicalFeaturesExcludingCas
        - Operations
        - PartOfSpeech
        - RegisterDatingFrequency
      - Semantic Content Representation
      - Syntax
      - Language Resource Ontology
      - Lexicography
      - Language Codes
      - Terminology
      - Multilingual Information Management
      - Lexical Resources
      - Lexical Semantics
      - Translation
      - Sign language
      - Audio
    - Athens Core
    - CLARIN-NL/VL

MorphologicalFeaturesExcludingCases

#	Name	Version	Administration stat	Registration status	Check	Type	Owned by	Scope
1227	active voice	1:0	private	private		simple	Francopoulo, Gil	public
3844	adjutative voice	1:0	private	private	✓	simple	Francopoulo, Gil	public
1902	animacy	1:0	private	private		closed	Francopoulo, Gil	public
1911	animate	1:0	private	private		simple	Francopoulo, Gil	public
3845	antipassive voice	1:0	private	private	✓	simple	Francopoulo, Gil	public
1240	**aorist**	1:0	private	private		open	Francopoulo, Gil	public
3843	apocope mood	1:0	private	private	✓	simple	Francopoulo, Gil	public
3846	applicative voice	1:0	private	private	✓	simple	Francopoulo, Gil	public
1242	aspect	1:0	private	private		closed	Francopoulo, Gil	public
1933	bound	1:0	private	private		simple	Francopoulo, Gil	public
2218	broken plural	1:0	private	private		simple	Francopoulo, Gil	public

aorist - 1:0

[-] 2.1 English Language Section	
Language	English (en)
2.1.1 Name Section	
Name	aorist
Name Status	standardized name
2.1.2 Definition Section	
Definition	Simple past tense that is predominantly used for narration. Both the perfective and the imperfective forms can be used in the aorist without any restrictions.
Source	www.helsinki.fi/~bontchev/grammar/index.html
[-] 2.2 French Language Section	
Language	French (fr)
2.2.1 Name Section	
Name	aoriste

Find: | Match case

ISO Annotation Frameworks

- LAF
- LMF: lexica
- MAF: tokenisation and word-level linguistic annotation
- SynAF: syntactic annotation
- SemAF: various types of semantic annotation: time & space, named entities, discourse
- Also some others, not mentioned here..

LMF

- ISO-24613:2008 – **Lexical Markup Framework** (LMF)
  - „an abstract metamodel that provides a common, standardized framework for the construction of computational lexicons. … of lexical objects, including morphological, syntactic, and semantic aspects.“
- Core specifications and extensions: morph., synt., sem.; multilingual
- As other MAFs, normative reference is in UML & informative XML dump DTDs
- http://www.lexicalmarkupframework.org/

Basic structure and simple morphology extension

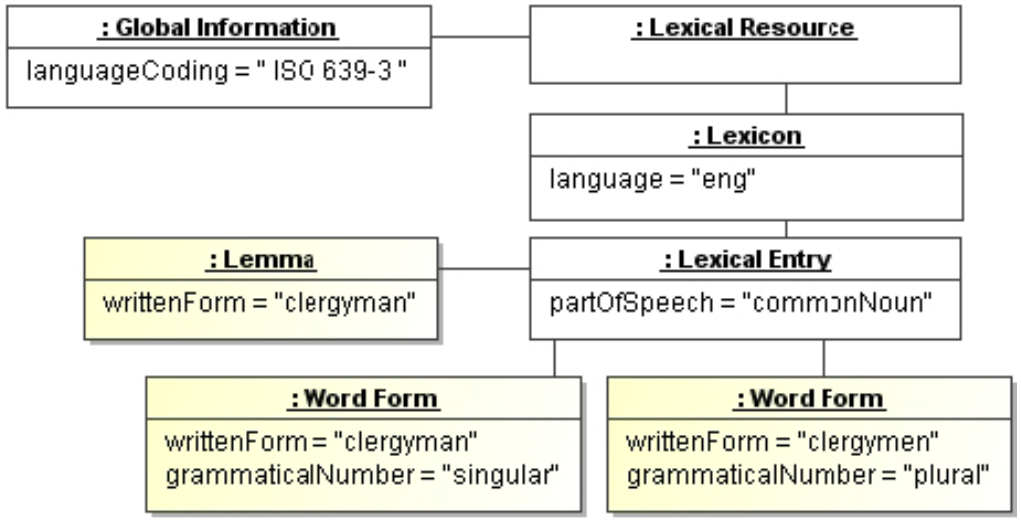
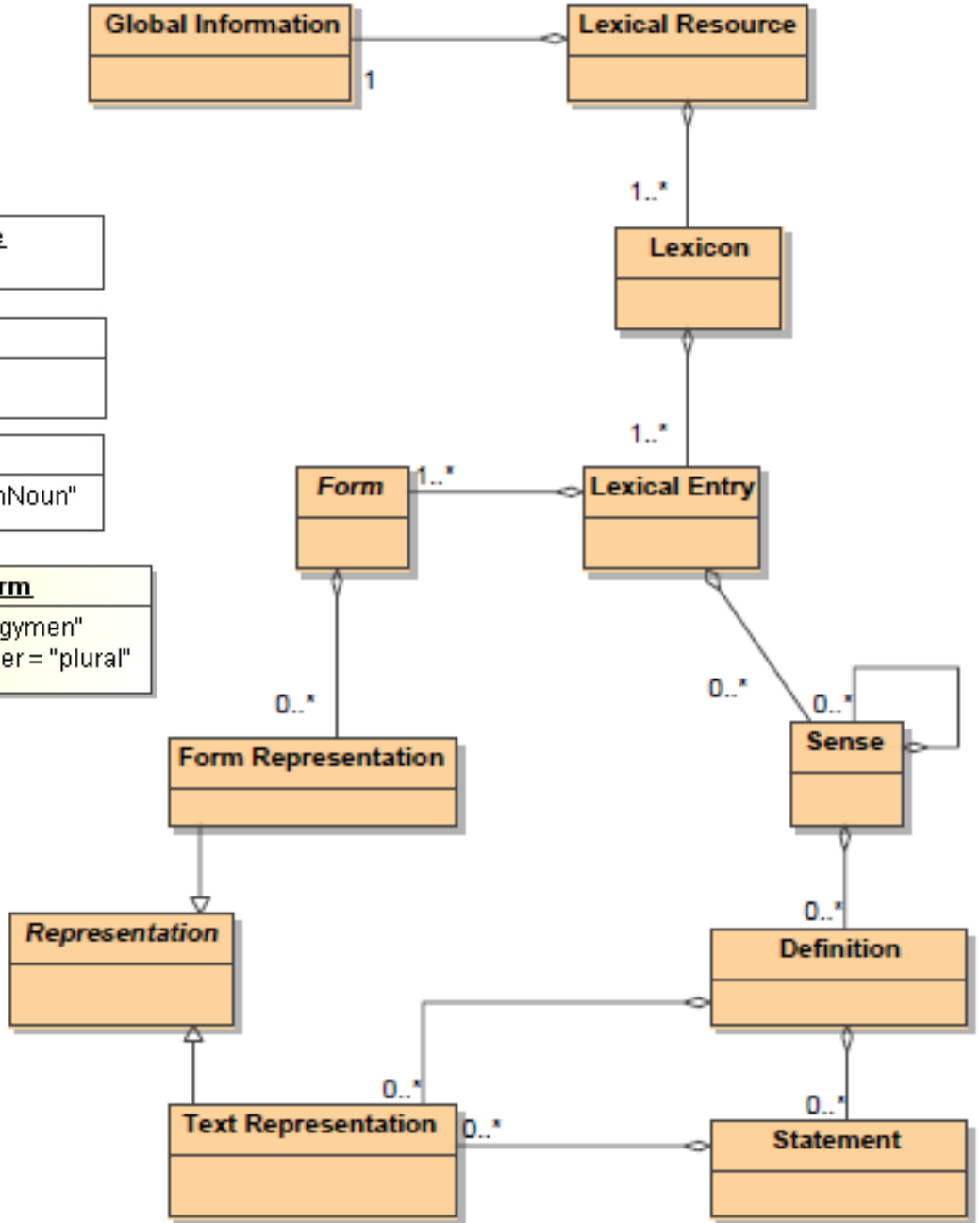


Figure B.1 – Instance diagram for a simple example



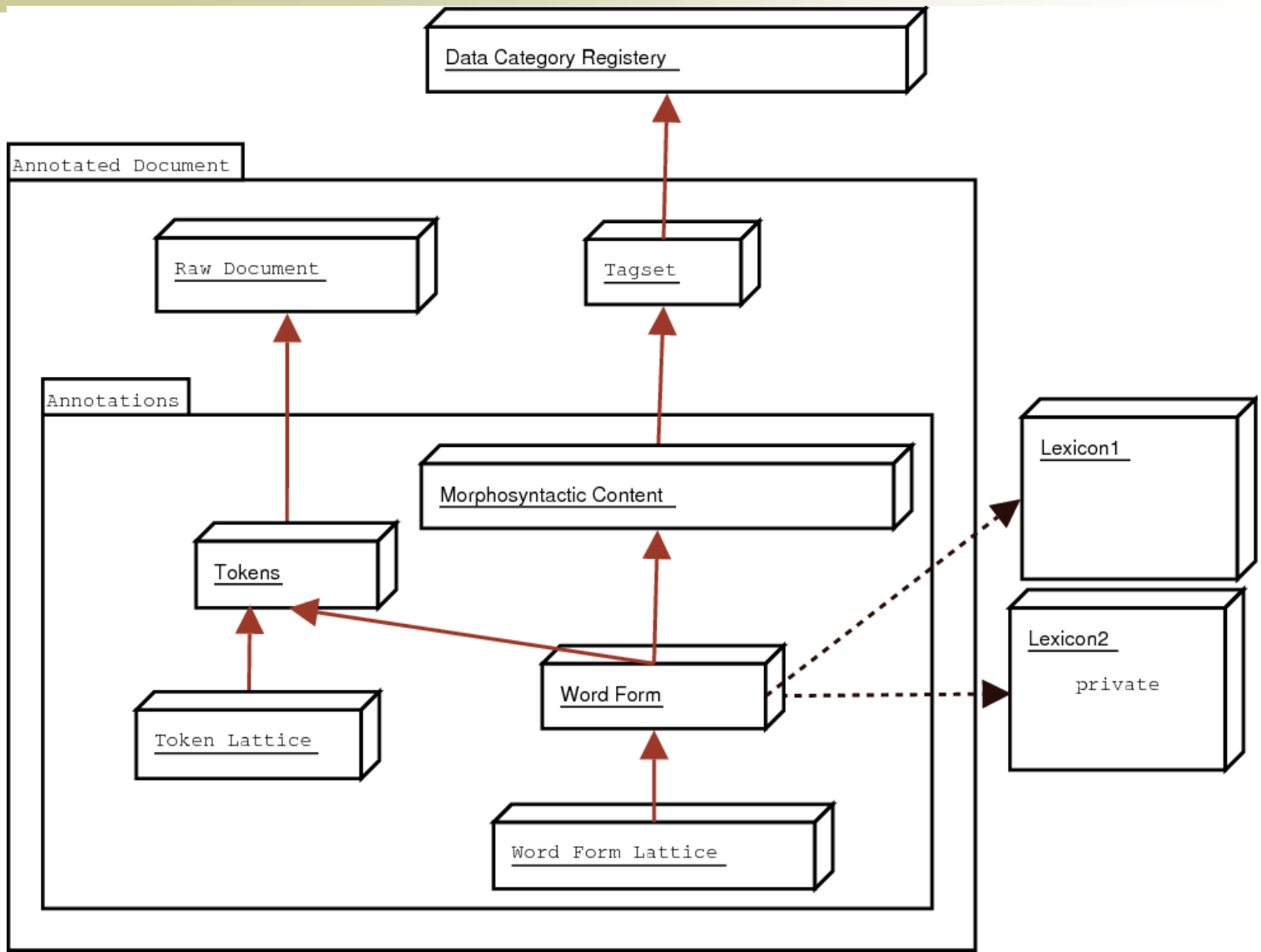
LMF Examples

- English inflected entry
- English syntactic data
- English phonetic data
- French morphological pattern and constraint setting
- Italian syntax/semantic mapping
- Spanish lemmas and wordforms

MAF

- ISO/DIS 24611 Language resource management -- **Morpho-syntactic annotation framework** (MAF)
- Last version: 2008-12-07, Eric de la Clergerie, INRIA
- Tokenisation and token morphosyntactic annotations
- Syntax given in UML / RELAX NG
- http://atoll.inria.fr/~clerger/MAF/

MAF metamodel



MAF, con‘t

- MAF advises stand-off annotation for all levels of analysis
- it does also support in-line annotation; in fact, most examples are of this sort, e.g.

```
<token id="t1">The</token>
<token id="t2">victim</token>
<token id="t3">'s</token>
<token id="t4">friends</token>
<token id="t5">told</token>
…
```

Token attributes

Informative attributes

```
<token form="et cetera" id="t1">etc.</token>
<token form="tzar" id="t2">csar</token>
<token phonetic="/platto/" id="t5">plateau</token>
```

Joining tokens

```
<token id="t1">L'</token>
<token id="t2" join="left">on</token>
<token id="t3">dit</token>
```

Overlapping tokens

```
<token form="et cetera" id="t1">etc.</token>
<token form="#dot#" id="t2" join="overlap"/>
```

Wordforms

Wordforms are separate elements:

```
<token id="t0">apple</token>
<wordForm lemma="apple" tag="pos.noun" tokens="t0"/>
```

Compound wordforms:

```
<token id="t0">prime</token>
<token id="t1">minister</token>
<wordForm lemma="prime_minister" tokens="t0 t1"/>
```

Split wordforms:

```
<token id="t0">auquel</token>
<wordForm lemma="à" tokens="t0"/>
<wordForm lemma="lequel" tokens="t0"/>
```

Morpho-syntactic content

- **Using feature-structures:**

```
<token id="t0">belle</token>
<wordForm lemma="beau" tokens="t0">
  <fs>
    <f name="pos"><symbol value="adjective"/></f>
    <f name="adj_type"><symbol value="qualifier"/></f>
    <f name="gender"><symbol value="feminine"/></f>
    <f name="number"><symbol value="singular"/></f>
  </fs>
</wordForm>
```

- **Using compact tags:**

```
<wordForm tokens="t0"
  tag="pos.adj adj_type.qual gender.fem num.sing"/>
```

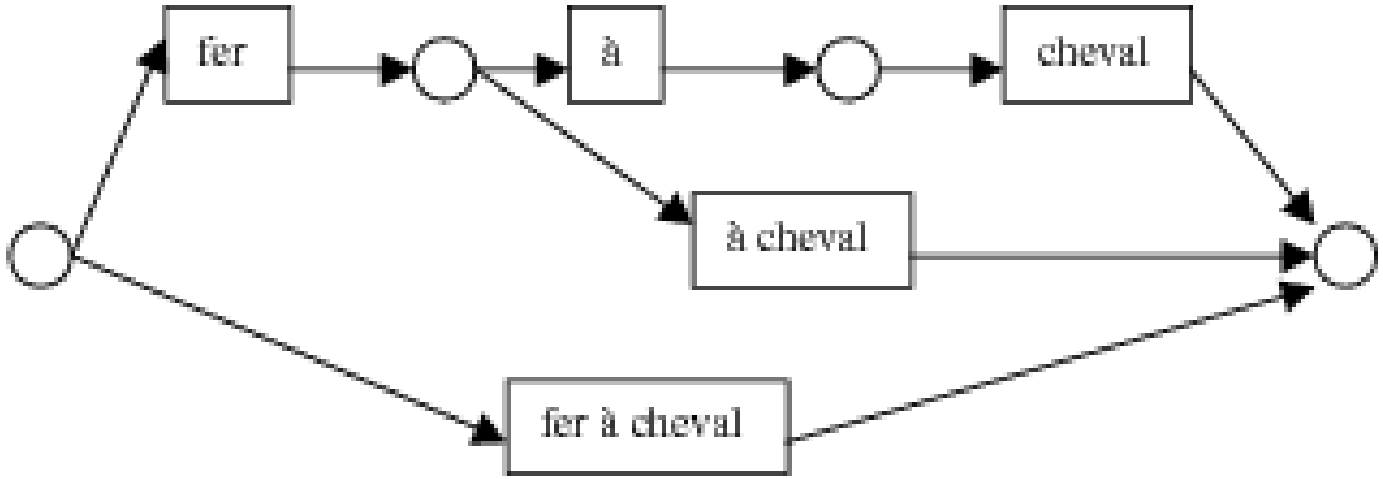
Connecting to the lexicon

- Use of URN (Uniform Resource Name)
- A URN is a URI that uses the *urn* scheme;
  does not imply availability of the identified resource.
- e.g.

```
<token id="t1">Prime</token>
<token id="t2">minister</token>
<wordForm
  entry="urn:lexicon:en:prime_minister"
  tokens="t1 t2"/>
```

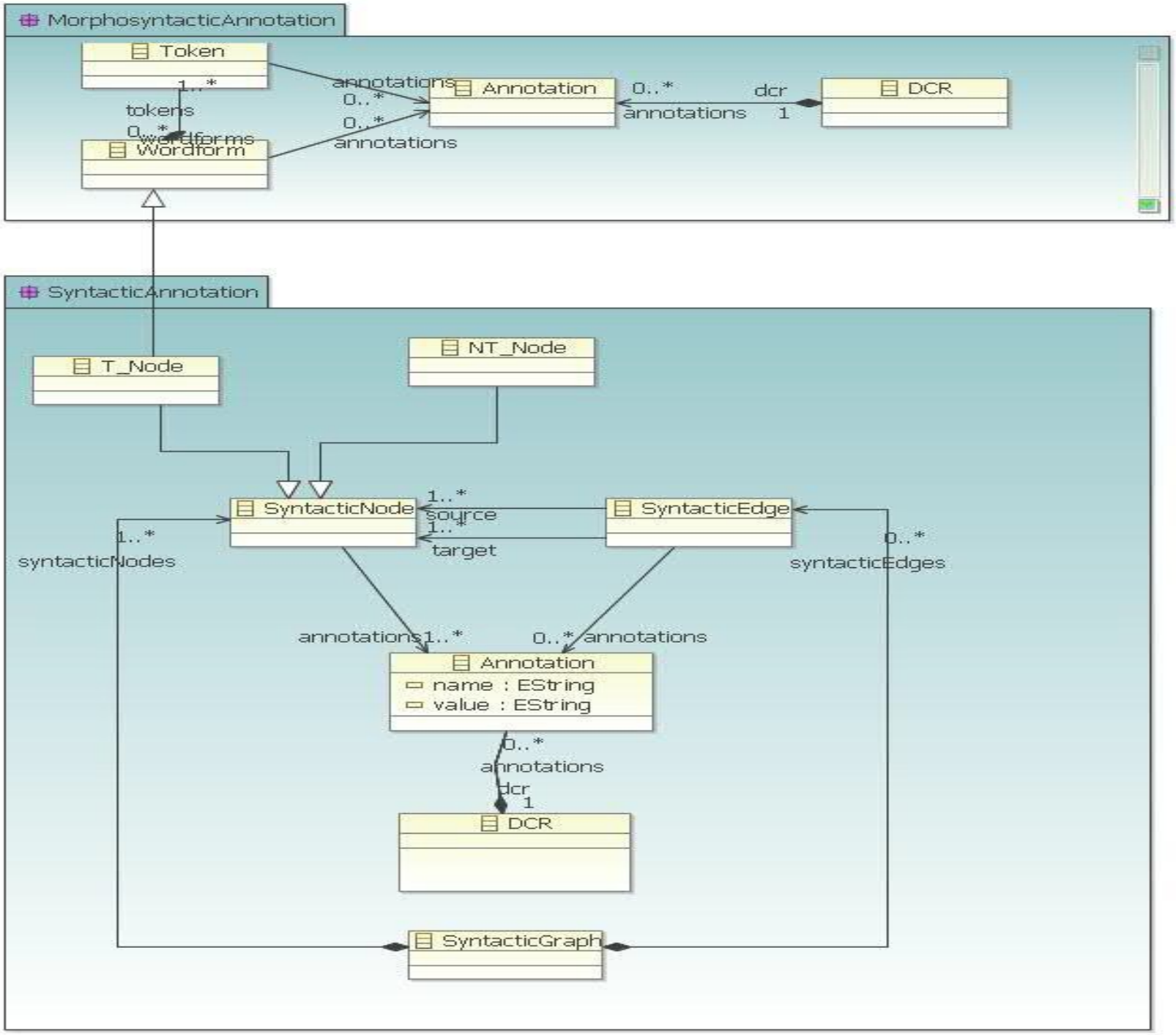
Ambiguities

- Supports word-form and lexical ambiguities
- Also structural (token) ambiguities, using DAGs / FSMs

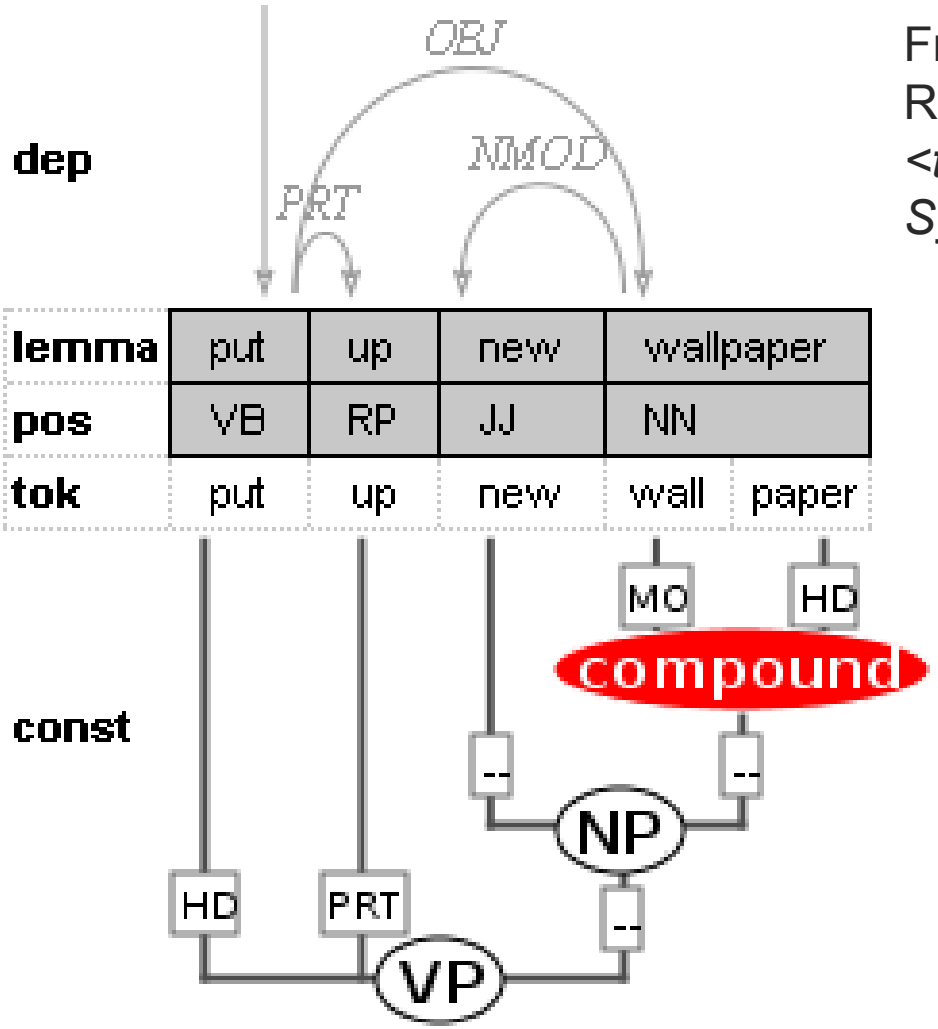


SyNAF

- ISO 24615:2010 Language resource management -- **Syntactic annotation framework** (SynAF)
- Based on the TIGER schema and results of the EU LIRICS project
- ISO defines only the meta-model, does not give a concrete dump
- Stand-off annotation, DAG, DCR
- Classes:
  - Syntactic node class: T Node class + NT Node class
  - Syntactic Edge class
  - Annotation class


MorphosyntacticAnnotation
Token
1..*
tokens
0..*
wordforms
Wordform
annotations
0..*
0..*
annotations
Annotation
0..*
annotations
dcr
1
DCR
SyntacticAnnotation
T_Node
NT_Node
SyntacticNode
1..*
source
1..*
target
SyntacticEdge
1..*
syntacticNodes
0..*
syntacticEdges
annotations
1..*
0..*
annotations
Annotation
name : EString
value : EString
0..*
annotations
dcr
1
DCR
SyntacticGraph

Graphic representation of a syntactic fragment annotated in multiple layers



From:
Romary, Zeldes, Zipser. (2011)
<tiger2/> - Serialising the ISO SynAF Syntactic Object Model.

Semantic annotation framework (SemAF)

- Many parts:
  - ISO/DIS 24617-1 Language resource management -- Semantic annotation framework (SemAF) -- Part 1: **Time and events** (SemAF-Time, ISO-TimeML)
  - ISO/DIS 24617-2 SemAF -- Part 2: **Dialogue acts**
  - ISO/PWI 24617-3 SemAF -- Part 3: **Named entities**
  - ISO/AWI 24617-4 SemAF -- Part 4: **Semantic roles**
  - ISO/AWI 24617-5 SemAF-- Part 5: **Discourse structure**

ISO - TimeML example

John taught 20 minutes every Monday.

John

<EVENT eid="e1" class="OCCURRENCE">taught</EVENT> <MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB" tense="PAST" aspect="NONE" polarity="POS"/>

<TIMEX3 tid="t1" type="DURATION" value="P20TM">20 minutes</TIMEX3>

<TIMEX3 tid="t2" type="SET" value="xxxx-wxx-1" quant="EVERY">every Monday</TIMEX3>

<TLINK timeID="t1" relatedToTime="t2" relType="IS_INCLUDED"/>

<TLINK eventInstanceID="ei1" relatedToTime="t1" relType="DURING"/>

MLIF

- ISO/FDIS 24616 Language resources management -- Multilingual information framework (MLIF)
- MLIF provides a platform for modeling and managing multilingual information: localization, translation, multimedia annotation, document management, …
- Provides a metamodel and DCR
- Also provides strategies for the interoperability and/or linking of models including, but not limited to XLIFF, **TMX**, SMILText, and ITS.

TMX – Translation Memory Exchange

- A standard of LISA (now defunct)
- e.g.

```
<tmx version="1.4">
 <header adminlang="en" creationdate="20040731T164933Z“
creationtool="Heartsome TM Server“ … />
 <body>
  <tu creationdate="20020930T004233Z" tuid="1091303313515">
   <tuv xml:lang="fr">
    <seg>Le processus de <hi xml:id="X3" type="term">contrôle de qualité</hi>
en dix étapes qu'il a créé il y a plus de 1300 ans est beaucoup plus complet et
précis que ceux existant aujourd'hui.</seg>
   </tuv>
   <tuv xml:lang="en">
    <seg>His 10-stage <hi corresp="#X3" type="term">quality control</hi>
process initiated more than 1300 years ago is far more thorough and exacting than
any existing today.</seg>
   </tuv>
…
```

TMX in MLIF

```
<MultiC>
 <creationIdentifier>1091303313515</creationIdentifier>
 <creationDate>20020930T004233Z</creationDate>
 <MonoC xml:lang="fr">
  <SegC>Le processus de <SegC xml:id="X3" type="term">contrôle
        de qualité</SegC> en dix étapes qu'il a créé il y a
      plus de 1300 ans est beaucoup plus complet et précis que
      ceux existant aujourd'hui.</SegC>
 </MonoC>
 <MonoC xml:lang="en">
  <SegC>His 10-stage <SegC corresp="#X3" type="term">quality
        control</SegC> process initiated more than 1300
      years ago is far more thorough and exacting than any
      existing today.</SegC>
 </MonoC>
</MultiC>
```