



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

LT resources – lecture 8

Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg
<lars.borin@svenska.gu.se>
<<http://spraakbanken.gu.se/personal/lars/>>

LT Resources
MLT, 6th October, 2011



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

overview

1. why standards?
2. levels of standards
3. some important standardization efforts
4. why these formats and frameworks?
5. some LT infrastructure initiatives



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

why standards?

Bird & Simons (2003) list the following problem areas for digital language data portability:

- | | |
|--------------|-----------------|
| 1. content | 5. citation |
| 2. format | 6. preservation |
| 3. discovery | 7. rights |
| 4. access | |

And they conclude:

“We need (...) open source tools based on agreed data models (...) connected to portable data formats” (Bird & Simons 2003: 580)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

why standards?

It has been argued that attempting standardization for language resources and surrounding information is premature, and the evolving nature of the domain and technology certainly speaks to that claim. But the growth of the web and the explosion in the number of electronic documents to be handled and maintained within the industrial sector has created an immediate and urgent need for generic language processing components for document indexing and classifying, information extraction, summarization, topic detection, etc., in both mono- and multi-lingual environments, together with robust machine translation and facilities for man-machine multimodal communication. While progress will continue, the field has nonetheless reached a point where we can see clear to a reasonable representation and processing model that should fulfill the needs of HLT for at least the foreseeable future. Indeed, commonality that can enable flexible use and reuse of communicative data is essential for the next generation of language processing applications, if we are to build a global information environment. It is therefore critical at this time to move toward standardization, and in particular, to do this in an internationally accepted framework.

(Nancy Ide & Laurent Romary 2007:
Towards international standards for language resources)





GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

standards are often overwhelming...

Universal Serial Bus Specification Revision 2.0

Table 7-12. Cable Characteristics (Note 14)

Parameter	Symbol	Conditions	Min	Max	Units
Vbus Voltage drop for detachable cables	VBUSD	Section 7.2.2		125	mV
GND Voltage drop (for all cables)	VGND	Section 7.2.2		125	mV
Differential Cable Impedance (full-/high-speed)	Z _o	(90 Ω ±15%);	76.5	103.5	Ω
Common mode cable impedance (full-/high-speed)	Z _{cm}	(30 Ω ±30%);	21.0	39.0	Ω
Cable Delay (one way)		Section 7.1.16			
Full-/high-speed	T _{FSCBL}			26	ns
Low-speed	T _{LSCBL}			18	ns
Cable Skew	T _{SKEW}	Section 7.1.3		100	ps
Unmated Contact Capacitance	C _{UC}	Section 6.7		2	pF
Cable loss		Specified by table and graph in Section 7.1.17			

Note 1: Measured at A plug.
 Note 2: Measured at A receptacle.
 Note 3: Measured at B receptacle.

(p. 185 of the USB 2.0 specification [xxviii+622 pp.]



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

...but they let you focus on the fun parts



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

levels of standards

wrt the data:

- ▶ basic data formats
- ▶ metadata standards
- ▶ data model frameworks
- ▶ information content and structure

wrt status of 'standard':

- ▶ widely used formats
- ▶ de-facto standards
- ▶ international (ISO, W3C, etc.) standards



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

data formats vs. information content

(Why is it not enough to say "It's XML"?)

Eventually, the opener departed for 24, Eoin Morgan taking his second catch in the slips. Wickets continued to tumble, despite the best efforts of Kamran Akmal (27), and Johnston's captaincy was spotted on as he brought back Boyd for some extra pace.

The bowler dug a couple in, and both Akmal and Azhar Mahmood spooned catches to Johnston at mid-wicket.

After Mohammad Sami and Ifikhar had added a gutsy 25 for the ninth wicket, spinner McCallan took the last two wickets as wild slogs were held in the deep. Pakistan had been bowled out for 132 in the 46th over.

The rate and manner of vacuum decay are calculated in an explicit flux compactification, including all thick-wall and gravitational effects. For landscapes built of many units of a single flux, the fastest decay is usually to discharge just one unit. By contrast, for landscapes built of a single unit each of many different fluxes, the fastest decay is usually to discharge all the flux at once, which destabilizes the radion and begets a bubble of nothing. By constructing the bubble of nothing as the limit in which ever more flux is removed, we gain new insight into the bubble's appearance. Finally, we describe a new instanton that mediates simultaneous flux tunneling and decompactification. Our model is the thin-brane approximation to six-dimensional Einstein-Maxwell theory.

Somewhere, parently, in the ginnandgo gap between antediluvian and annadominant the copyist must have fled with his scroll. The billy flood rose or an elk charged him or the sultrup worldwright from the excelsissimost empyrean (bolt, in sum) earthspake or the Dannamen gallous banged pan the bliddy duran. A scribicide then and there is led off under old's code with some fine covered by six marks or ninepins in metalmen for the sake of his labour's dross while it will be only now and again in our rear of o'er era, as an upshoot of military and civil engagements, that a gynecure was let on to the scuffold for taking that same fine sum covertly by meddlement with the drawers of his neighbour's safe.





GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

basic format standards

- ▶ e.g.:
 - ▶ **Unicode** <<http://www.unicode.org>>
(≡ ISO/IEC 10646)
 - ▶ **XML** (eXtensible Markup Language)
<<http://www.w3c.org/XML/>> (W3C)
 - ▶ **JSON** (Javascript Object Notation)
<<http://www.json.org/>>
 - ▶ **RDF** (Resource Description Framework)
<<http://www.w3.org/RDF/>> (W3C)
 - ▶ ...
- ▶ (plus media formats. . .)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

metadata standards

- ▶ e.g.:
 - ▶ **Dublin Core** <<http://dublincore.org/>>
 - ▶ **ISO 639(-1-3)** <<http://www.sil.org/iso639-3/>>
 - ▶ **OLAC** (Open Language Archives Community)
<<http://www.language-archives.org/>>
 - ▶ **CMDI** (Component Metadata Infrastructure)
<<http://www.clarin.eu/cmdi>>



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

data model frameworks

- ▶ e.g.:
 - ▶ the **ISO TC37/SC4** standards and standard proposals <<http://www.tc37sc4.org/>>
 - ▶ the **TEI/(X)CES** text/corpus formats
<<http://www.tei-c.org/>>
<<http://www.xces.org/>>
 - ▶ **OWL** (Web Ontology Language)
<<http://www.w3.org/2004/OWL/>>



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

information content and structure

- ▶ e.g.:
 - ▶ the **EAGLES** morphosyntactic tagset standard
 - ▶ **GOLD** (General Ontology for Linguistic Description)
<<http://www.linguistics-ontology.org/>>
 - ▶ **ISO TC37/SC4 DCR** <<http://www.isocat.org/>>



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

TC37/SC4: Data Category Registry (DCR)

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

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

Mest besökta Language Log Lingvistbloggen Tg3b.bmp (BMP-bil...) The LINGUIST Lis... Institutionen för s...

ISOcat

enter keywords here search

My Workspace

- public
- Category Administration
- DCR MetaModel
- Terminology
- Terminology MetaModel
- MetaData
- Morpho Syntax
- Semantic
- Dialog
- Language Description
- Syntax

MorphoSyntax

#	Name	Version	Administration status	Registration status	Type	Owned by	Scope
1223	abessive case	0.0.0	private	candidate	simple	Gil Francopoulo	public
1224	ablative case	0.0.0	private	candidate	simple	Gil Francopoulo	public
1225	absolute case	0.0.0	private	candidate	simple	Gil Francopoulo	public
1226	accusative case	0.0.0	private	candidate	simple	Gil Francopoulo	public
1227	active voice	0.0.0	private	candidate	simple	Gil Francopoulo	public
1228	adessive case	0.0.0	private	candidate	simple	Gil Francopoulo	public
1229	aditive case	0.0.0	private	candidate	simple	Gil Francopoulo	public

absolute case - 0.0.0

2 descriptionSection

profile MorphoSyntax

2.1 languageSection

language en

2.1.1 definitionClass

definition Case for nouns in ergative-absolute languages that would generally be the subjects of intransitive verbs or the objects of transitive verbs in the translational equivalents of nominative-accusative languages such as English.

source IJL

2.1.2 expansionClass

new save new as export

#	Name	Version	Administration status	Registration status	Type	Owned by	Scope
---	------	---------	-----------------------	---------------------	------	----------	-------

Klar



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

TEI

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.toc.html

Mest besökta Language Log Lingvistbloggen Tg3b.bmp (BMP-bil...) The LINGUIST Lis... Institutionen för s...

Front Matter

- I Releases of the TEI Guidelines
- ii Dedication
- iii Preface and Acknowledgments
- iv About These Guidelines
- v A Gentle Introduction to XML
- vi Languages and Character Sets

Back Matter

- Appendix A Model Classes
- Appendix B Attribute Classes
- Appendix C Elements
- Appendix D Attributes
- Appendix E Datatypes and Other Macros
- Appendix F Bibliography
- Appendix G Errata/Notes
- Appendix H Colophon

Text body

- 1 The TEI Infrastructure
- 2 The TEI Header
- 3 Elements Available in All TEI Documents
- 4 Default Text Structure
- 5 Representation of Non-standard Characters and Glyphs
- 6 Verse
- 7 Performance Texts
- 8 Transcriptions of Speech
- 9 Dictionaries
- 10 Manuscript Description
- 11 Representation of Primary Sources
- 12 Critical Apparatus
- 13 Names, Dates, People, and Places
- 14 Tables, Formulas, and Graphics
- 15 Language Corpora
- 16 Linking, Segmentation, and Alignment
- 17 Simple Analytic Mechanisms
- 18 Feature Structures

Klar



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

(X)CES

Arkiv Bedigera Visa Historik Bokmärken Verktyg Hjälp

http://www.xces.org/

Mest besökta Language Log Lingvistbloggen Tg3b.bmp (BMP-bil...) The LINGUIST Lis... Institutionen för s...

XML Corpus Encoding Standard Document XCES 1.0.4. Last Modified 20 June 2008

Vassar College

Department of Computer Science
Vassar College
Poughkeepsie NY
USA

LORIA

Equipe Langue et Dialogue
LORIA/GNRS
Vandœuvre-lès-Nancy
FRANCE

EAGLES

XCES

Corpus Encoding Standard for XML

NEW XCES RESOURCES

Klar



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

TC37/SC4: Feature structures

INTERNATIONAL
STANDARD

ISO
24610-1

First edition
2006-04-15

Language resource management —
Feature structures —

Part 1:
Feature structure representation

Gestion des ressources linguistiques — Structures de traits —
Partie 1: Représentation de structures de traits



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

TC37/SC4: Lexical Markup Framework

INTERNATIONAL
STANDARD

ISO
24613

First edition
2008-11-15

**Language resource management —
Lexical markup framework (LMF)**

Gestion de ressources langagières — Cadre de balisage lexical (LMF)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

some other ISO TC37/SC4 standards

- ▶ word segmentation
- ▶ linguistic annotation framework
- ▶ morpho-syntactic annotation framework
- ▶ syntactic annotation framework
- ▶ semantic annotation framework



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

OWL



OWL Web Ontology Language
Overview

W3C Recommendation 10 February 2004

This version:

<http://www.w3.org/TR/2004/REC-owl-features-20040210/>

Latest version:

<http://www.w3.org/TR/owl-features/>

Previous version:

<http://www.w3.org/TR/2003/PR-owl-features-20031215/>

Editors:

Deborah L. McGuinness (Knowledge Systems Laboratory, Stanford University) dlm@kel.stanford.edu

Frank van Harmelen (Vrije Universiteit, Amsterdam) Frank.van.Harmelen@cs.vu.nl



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

why these formats and frameworks?

First, there are good *extrinsic* reasons for using them:

- ▶ **Unicode** (an ISO standard) and **XML** (a W3C standard) have a large and growing tool base and can be expected – because they are international standards – to be both portable and long-lived
- ▶ There is excellent support for converting into and out of XML from/to other data formats



why these formats and frameworks?, 2

There are also some *intrinsic* reasons for using them:

- ▶ **Unicode** and **XML** are basic; the others build on them, directly or indirectly:
 - ▶ **RDF** (a W3C standard) and the **ISO TC37/SC4** standards/ proposals have XML bindings (i.e., agreed ways of expressing them using XML)
 - ▶ **OWL** (a W3C standard) is expressed using RDF
 - ▶ **metadata** are expressed either directly in XML or in RDF/OWL

why OWL as linguistic description framework?

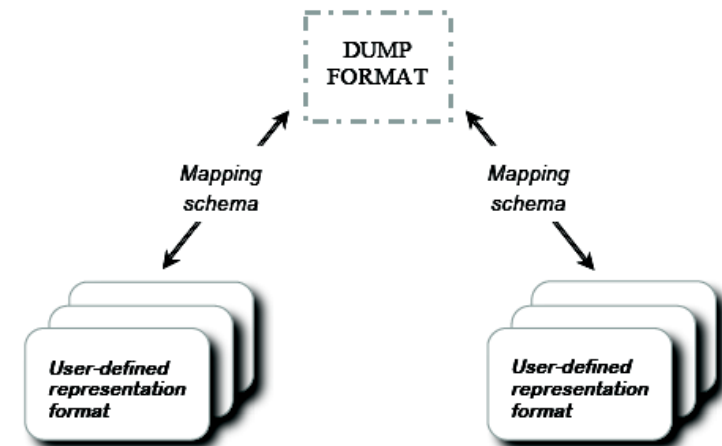
- ▶ A modern W3C format, OWL has a growing tool base
- ▶ OWL promises principled and sound automatic merging of information from distributed information sources (in fact, this is its *raison d'être*)
- ▶ Such information sources – called *ontologies* in the terminology of the OWL community – grow in number, in all domains
- ▶ There is great interest in providing OWL versions of several ISO TC37/SC4 standards
- ▶ Choosing OWL as linguistic representation framework can potentially ensure interoperability between language resources in the narrower sense, multimedia components, and domain knowledge bases

why these formats and frameworks?, 3

Summing up, we could say that

- ▶ **Unicode** provides a standardized way of *expressing* (written) linguistic content
- ▶ **XML** provides a standardized way of *processing form*
- ▶ **OWL** provides a standardized way of *processing content*

when in Rome...



(Nancy Ide & Laurent Romary 2007:
Towards international standards for language resources)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

some LT infrastructure initiatives

- ▶ ELRA <<http://www.ela.info>>
- ▶ LDC <<http://www ldc.upenn.edu>>
- ▶ CLARIN <<http://www.clarin.eu>>
- ▶ FlareNet <<http://www.flarenet.eu>>
- ▶ META-NET <<http://www.meta-net.eu>>
(META-NORD <<http://www.meta-nord.eu>>, CAESAR, METANET4U)
- ▶ SILT <<http://www.anc.org/SILT>>
- ▶ Språkbanken <<http://spraakbanken.gu.se>>



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

SILT

SILT
Sustainable Interoperability for Language Technology

Home Activities Events Collaborations People Become Involved Contact

The Sustainable Interoperability for Language Technology (SILT) project's goal is to turn existing, fragmented technology and resources developed to support language processing technology into accessible, stable, and interoperable resources that can be readily reused across several fields.

Project funded by the National Science Foundation

Community-based Data Interoperability Networks (INTEROP)
NSF INT-0753069

One of today's greatest challenges is the development of language processing capabilities that will enable easy and natural access to computing facilities and information. Because natural language processing (NLP) research relies heavily on such resources to provide training data to develop language models and optimize statistical algorithms, language resources—including (usually large) collections of language data and linguistic descriptions in machine readable form, together with tools and systems (lemmatizers, parsers, summarizers, information extractors, speech recognizers, annotation development software, etc.)—are critical to this development.



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

META-NORD

META-NORD
Baltic and Nordic Parts of the European Open Linguistic Infrastructure

Project

- Home
- About
- Partners
- Links
- Contacts
- Members login
- For Community**
- Flyers, Posters

Latest news

EUROLAN 2011 Summer school

On August 24 - September 4 the **EUROLAN 2011 Summer school**, the venue was Cluj-Napoca, Romania, in the heart of Transylvania, provided one week of intensive study of the natural language processing technologies currently under development to support industrial applications. Internationally known scholars, researchers (with the particular involvement of scientists from the Multilingual Europe Technology Alliance - META), as well as industrials involved in leading-edge work in innovative areas of natural language processing gave **lectures** at the school (tutorials, hands-on labs and demos) to share with students in-depth understanding and experience.

The META-NORD project was one of the silver **sponsors** of the event and the project coordinator Andrejs Vasiljevs gave a lecture "META-NET, related projects and industry-research collaboration"



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Språkbanken

Språk BANKEN UNIVERSITY OF GOTHENBURG

About us Resources Publications Research PhD program Staff SEARCH

View Edit Revisions Track Translate Clone

Corpora

Språkbanken (the Swedish Language Bank)

In recognition of the groundbreaking corpus linguistic work initiated by Sture Allén at the University of Gothenburg in the 1960s (which had resulted in the creation of one of the first large electronic text corpora in another language than English, Press-65, one million words of newstext), Språkbanken (the Swedish Language Bank) was established in 1975 as a national center with a remit to collect, process and store (Swedish) text corpora, and to make linguistic data extracted from the corpora available to researchers and to the public. Since then, Språkbanken has developed into a nationally and internationally acknowledged research unit whose work focuses on the development of linguistic resources and tools, and methodologies for using the resources in research in language technology and a number of other disciplines.

corpora count	56
tokens (total)	770 453 479
sentences (total)	48 471 880

Lexical resources

SBLEX
Search in the lexical resources

lexica count	15
entries (total)	517 517

SveFN++
The Swedish framenet project

SALDO
Lexical resource for Swedish language technology

Söderwall och Schlyter
Dictionary of Old Swedish