Standards for digital encoding

Tomaž Erjavec

Institut für Informationsverarbeitung
Geisteswissenschaftliche Fakultät
Karl-Franzens-Universität Graz

24.11.2006

Overview

1. **what are corpora**
2. **structure of a TEI corpus**
3. **linguistic annotation of corpora**
4. **alignment**
5. **TEI (corpus) header**

Practicum:

- XSLT: making an index and ToC
- TEI encoding of a corpus

I. What is a corpus?

The Collins English Dictionary (1986):

1. a collection or body of writings, esp. by a single author or topic.

Guidelines of the Expert Advisory Group on Language Engineering Standards, EAGLES:

***Corpus** : A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.*

***Computer corpus** : a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.*

Using corpora

- Research on *actual* language: descriptive approach, study of performance, empirical linguistics.
- Applied linguistics:
  - *Lexicography*: mono-lingual dictionaries, terminological, bi-lingual
  - *Language studies*: hypothesis verification, knowledge discovery (lexis, morphology, syntax, ...)
  - *Translation studies*: a source translation equivalents and their contexts
    translation memories, machine aided translations
  - *Language learning*: real-life examples
    "idiomatic teaching", curriculum development
- *Language technology*:
  - testing set for developed methods;
  - *training set* for inductive learning
  - (statistical Natural Language Processing)

Characteristics of a corpus

- *Quantity*:
  - the bigger, the better
  - but processing / sampling problems
- *Quality* :
  - the texts are authentic (character sets, spacing, form);
  - the mark-up is validated (XML)
- *Simplicity*:
  - computer representation is understandable
  - markup easily separated from the text (XML)
- *Documented*:
  - corpus contains bibliographic
  - and other meta-data (teiHeader / Dublin Core / …)

Typology of corpora

- Corpora of written language, spoken and speech corpora
  - authenticity vs. price
  - e.g. the agency ELRA catalog
- Reference corpora and specialised (sub-language) corpora
  - representativeness and balance
  - e.g. BNC, ICE, COLT
- Corpora with integral texts or of text samples
  - historical and legal reasons, e.g. Brown
  - but also balance: the whelk problem
- Static and monitor corpora (language change)
- Monolingual and multilingual parallel and comparable corpora
  e.g. Hansard, Europarl
- Plain text and annotated corpora
  - the “original” should be archived, ideally linked to the corpus

The history of computer corpora:

- First milestones: Brown (AE, 1 million words) 1964
- LOB et al. (~ Brown but BE) 1974
- 100M reference corpora: Cobuild Bank of English (monitor) 1980; BNC 1995; Czech CNC 1998; Slovene FIDA, Nova Beseda 1998; Croatian HNK, Hungarian HNC, ...
- EU corpus oriented projects in the '90: EAGLES, MULTEXT, SQEL, MULTEXT-East,...
- Language resources brokers: LDC 1992, ELRA 1995

Steps in the preparation of a corpus

1. Which texts? Sampling "the universe of discourse"
   - linguistic and non-linguistic criteria; availability; simplicity; size
2. Legalities: ©
   - sensitivity of source (privacy / publishing considerations);
   - agreement with providers; usage, publication
3. How to get them? Acquiring digital originals
   - typing, OCR, get CD-ROM, download from provider
   - Web, Google, BootCat, ...?
4. Up-translation. Conversion to standard format (+documentation)
   - consistency; character set encodings
5. (Linguistic annotation) Segmentation and classification
   - errors, language dependency
6. Documentation Metadata
   - TEI header; Dublin Core; Open Archives etc.
7. Using it yourself, or letting others too?
   - (Web-based) concordancers for linguists
   - download needed for HLT use
   - licences for use

What annotation can be added to the text of the corpus?

- Annotation = interpretation
- Documentation about the corpus and corpus component (<teiHeader>s)
- Document structure (, <p>, etc.) usu. not a priority!
- Basic linguistic markup: sentences, words, punctuation (<s>, <w>)
- Lemmas and morphosyntactic descriptions
- Syntax (treebanks)
- Multilinguality (sentence and word alignment)
- Terms, semantics, anaphora, pragmatics, intonation,...

Markup Methods

- *hand annotation*: documentation, first steps
  generic (XML, spreadsheet) editors or specialised editors
- *semi-automatic*: morphosyntactic and other linguistic annotation
  cyclic approach: machine, hand, validate, correct, machine, ...
- *machine, with hand-written rules*: tokenisation
  regular expression
- *machine, with inductively built models from annotated data*:
  "supervised learning"; HMMs, decision trees, inductive logic programming,...
- *machine, with inductively built models from un-annotated data*:
  "unsupervised leaning"; clustering techniques; text mining
- overview of the field

The future of corpus and data-driven linguistics

- Size:
  - Larger quantities of readily accessible data (Web as corpus)
  - Larger storage and processing power (Moore law)
- Complexity:
  - Deeper analysis:
    syntax, deixis, semantic roles, dialogue acts, ...
  - Multimodal corpora:
    speech, film, transcriptions,...
  - Annotation levels and linking:
    co-existence and linking of varied types of annotations; ambiguity
  - Development of tools and platforms:
    precision, robustness, unsupervised learning, meta-learning

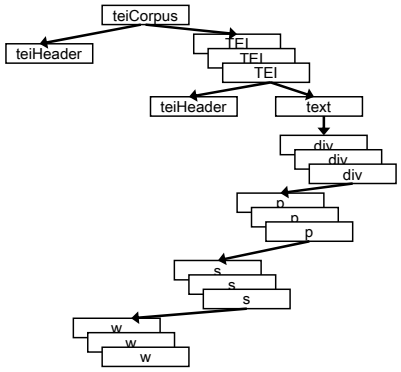
II. Types of annotation

- Segmentation
  - paragraphs, sentences, words, phonemes, …
- Categorisation
  - word-level morphosyntactic information (PoS)
  - word lemmas or stems
  - syntactic structures
- Alignment and correspondence
  - translation equivalence
  - anaphoric reference
- Metadata
  - text source: title, date, author, …
  - text type: register, publ. type, …

Segmentation

- Corpora (and texts) are generally organized hierarchically
- Segmentation and labelling allow their components to be identified and accessed at any level
  - for reference purposes
    - e.g. *this* occurs at ....
  - for scoping purposes
    - e.g. find *this* within a *that*
  - for analytic purposes
    - e.g. 90% of *these* of type *that* contain a *the other*

Typical TEI corpus structure



Example

```
<text id="Oen." lang="en">
  <body>
    <div type="part" id="Oen.1">
      <div type="chapter" id="Oen.1.1">
        <p id="Oen.1.1.1">
          <s id="Oen.1.1.1.1">
            <w>It</w>
            <w>was</w>
            <w>a</w>
            <w>bright</w>
            <w>cold</w>
            <w>day</w>
            <w>in</w>
            <w>April</w>
            <c>,</c>
            <w>and</w>
            <w>the</w>
            <w>clocks</w>
            <w>were</w>
            <w>striking</w>
            <w>thirteen</w>
            <c>.</c>
          </s>
```

Representation of trees

- Example: (NC (N the method) (C to be used))
- in TEI:

```
<seg type=”NC”>
    <seg type=”N”>the method</seg>
    <seg type=”C”>to be used</seg>
 </seg>
```

- <seg> can be used for any type of linguistic segment
- more specific elements: <s>, <cl>, <phr> <w>, <m>, <c>
- all defined in module for linguistic analysis

Alignment and correspondence

- The spans to be annotated are not always structural
- Discontinuities are commonplace
- Parallel structures are the norm
- Solutions:
  - pointers
  - milestone (empty) elements
  - stand-off markup

Alignment of multiple speakers

A: Have you heard the		It’s a disaster!
B:	the election results?	It’s a miracle!

```
<u id="A1" who="A">Have you heard the</u>
<u id="B1" who="B" trans="latching">the election results? </u>
<u id="A2" who="A" trans="pause">its a disaster</u>
<u id="B2" who="B" trans="overlap">its a miracle </u>
```

Synchronization using pointers

- of whole elements

```
<u synch='#B2'>its a disaster</u>
<u xml:id='B2'>its a miracle</u>
```

- of points in time

```
<u xml:id='A1' who='A'>Have you heard
<anchor xml:id='AO1'/>the</u>
<u id='xml:B1' who='B' synch='#A01'>
the election results?</u>
```

Discontinuity: using pointers

- "You put it," Quill reminded him, "in the safe."
- Encoding:

```
<s xml:id="s1" next="#s3">"You put it,"</s>
<s xml:id="s2">Quill reminded him,</s>
<s xml:id="s3" prev="#s1">"in the safe."</s>
```

- can also use PART attribute to indicate that segments are incomplete

Translation equivalence using pointers

```
<s xml:id="s1" corresp="#s2" xml:lang="EN">
For a long time I used to go to bed early</s>
…
<s xml:id="s2" corresp="#s1" xml:lang="FR">
Longtemps je me couchais de bonne heure</s>
```

Translation equivalence using stand-off markup

```
<s xml:id="s1" xml:lang="EN">
For a long time I used to go to bed early</s>
<s xml:id="s2" xml:lang="FR">
Longtemps je me couchais de bonne heure</s>
```

```
<linkGrp type="transEquiv">
  <link targets="#s1"/>
  <link targets="#s2"/>
</linkGrp>
```

Anaphoric reference

- With pointers:

<title xml:id='shirl'>Shirley</title>, which made its Friday night debut only a month ago, was not listed on **<name xml:id='nbc'>NBC</name>**'s new schedule, although **<rs xml:id='nwk' corresp='nbc'>the network</rs>** says **<rs xml:id='#show' corresp='#shirl'>the show</seg>** still is being considered.

- Stand-off:

```
<linkGrp type="anaphor">
   <link targets="#shirl"/>
   <link targets="#show"/>
</linkGrp>
```

Word class categorization

```
<s n="00011">
   <w pos="NN1">Difficulty</w>
   <w pos="VBZ">is</w>
   <w pos="VBG">being</w>
   <w pos="VVN">expressed</w>
... </s>
```

```
<s id="Oen.1.1.1.1">
<w lemma="it" ana="#Pp3ns">It</w>
<w lemma="be" ana="#Vmis3s">was</w>
<w lemma="a" ana="#Di">a</w>
<w lemma="bright" ana="#Af">bright</w>
<w lemma="cold" ana="#Afp">cold</w>
<w lemma="day" ana="#Ncns">day</w>
```

Making analysis explicit

```
<w ana="#NN1">Difficulty</w>
```

- The *ana* attribute is a pointer
- What does *NN1* identify?
  - a prose description
  - an **<interp>** element
  - a feature structure

for example…

```
<w ana='#VVD'>annotated</w>
<w ana='#NN2'>corpora</w>
```

```
<interp xml:id='VVD'>
  <desc>verb past tense</desc>
</interp>
<interp xml:id='NN2'>
  <desc>plural common noun</desc>
</interp>
```

Formal encoding of analyses

- Linguistic Annotation Frameworks and standards
  - the philosophers stone
- Generic feature structure system
  - any analysis can be represented by bundles of named *feature-value* pairs
  - embedded within text or indirectly linked
- Ancillary feature system declaration
- Theoretically neutral (?) pragmatic solution to real world problem of intermachine communication

Feature structures

- a *feature structure* consists of a set bundle of *feature*s
- a feature has a *name* and a *value*
- values may be binary switches, symbols, strings, feature structures, or operations on them
- bundling may constrained in various (not necessarily hierarchic) ways

... or, in XML:

- The <fs> element represents a feature structure, which contains...
- One or more <f> elements, each of which has
  - a name
  - a value
- Feature values may be
  - atomic: <binary> <string> <numeric> <symbol>
  - complex: <fs> <coll>
  - expressions: <vNot> <vAlt> <vColl> ... or <var>

Using a feature structure...

```
<w ana='#NN2'>corpora</w>
```

```
<fs xml:id='NN2'>
   <f name='class'><symbol value='noun'/></f>
   <f name='number'><symbol value='plural'/></f>
   <f name='proper'><binary value="false"/></f>
</fs>
```

An example of a lexical definition

```
<fs type='word structure'>
  <f name='lemma'>
       <string>goose</str>
  </f>
  <f name='category'>
       <symbol value='noun'/>
  </f>
  <f name='barLevel'>
       <numeric value='0'/>
  </f>
  <f name='number'>
       <symbol value='plural'/>
  </f>
</fs>
```

lemma: goose
category: noun
number: plural
bar level: 0

A more complex example

```
<fs>
    <f name="lexicalForm">
      <symbol value="auxquels"/></f>
   <f name="analyses">
     <coll org="list">
        <fs>
           <f name="cat"><symbol value="prep"/></f>
        </fs>
        <fs>
         <f name="cat"><symbol value="pronoun"/></f>
         <f name="kind"><symbol value="rel"/></f>
         <f name="num"><symbol value="pl"/></f>
         <f name="gender"><symbol value="masc"/></f>
        </fs>
     </coll>
   </f>
</fs>
```

The TEI header

The TEI header gives the meta-data on the TEI document and consists of four elements:

- <fileDesc>
  the file description (the only obligatory element)
- <encodingDesc>
  the encoding description
- <profileDesc>
  the text profile
- <revisionDesc>
  the revision history

<fileDesc>

- *file description*, containing a full bibliographical description of the computer file itself
- e.g. the title statement, edition statement, size, …
- includes also information about the source or sources (<sourceDesc>) from which the electronic text was derived

File description (2)

- all subelements

```
<teiHeader>
  <fileDesc>
    <titleStmt/>
    <editionStmt/>
    <extent/>
    <publicationStmt/>
    <seriesStmt/>
    <notesStmt/>
    <sourceDesc/>
  </fileDesc>
</teiHeader>
```

- minimal header

```
<teiHeader>
  <fileDesc>
    <titleStmt/>
    <publicationStmt/>
    <sourceDesc/>
  </fileDesc>
</teiHeader>
```

fileDesc/titleStmt

- Short Example

```
<titleStmt>
 <title>Two stories by Edgar
   Allen Poe: electronic
   version</title>
 <author>Poe, Edgar Allen
   (1809-1849)</author>
 <respStmt>
   <resp>compiled by</resp>
   <name>James
   Benson</name>
 </respStmt>
</titleStmt>
```

- Longer example:

```
<titleStmt>
    <title>Yogadarśanam (arthāt
     yogasūtrap¯&tdot;ha&hdot;):
     a machine readable transcription.</title>
    <title>The Yogasūutras of Patañjali:
     a machine readable transcription.</title>
    <funder>Wellcome Institute for the History
of Medicine</funder>
    <principal>Dominik Wujastyk</principal>
    <respStmt>
      <name>Wieslaw Mical</name>
      <resp>data entry and proof
correction</resp>
    </respStmt>
    <respStmt>
      <name>Jan Hajic</name>
      <resp>conversion to TEI-conformant
markup</resp>
    </respStmt>
</titleStmt>
```

fileDesc/publicationStmt

- publication statement groups information concerning the publication or distribution of an electronic or other text
- It may contain either a simple prose description
- or groups of the elements described below:
  - **<publisher>** who is responsible for the publication
  - **<distributor>** who is responsible for the distribution
  - **<authority>** (release authority) who is responsible for making an electronic file available, other than a publisher or distributor
- each of the above elements may be followed by one or more of the following elements: **<pubPlace>, <address>, <idno>, <availability>** , **<date>**

fileDesc/sourceDesc

- The Source Description records details of the source or sources from which a computer file is derived.
- An electronic file may also have no source, if what is being catalogued is an original text created in electronic form.
- The <sourceDesc> element may contain a simple prose description, or, more usefully, a bibliographic citation:
  - <bibl>, <biblItem>, <biblFull>, <biblStruct>, <listBibl>

<sourceDesc> examples

```
<sourceDesc>
      <bibl>The first folio of Shakespeare, prepared by
       Charlton Hinman (The Norton Facsimile, 1968)</bibl>
</sourceDesc>
```

```
<sourceDesc>
 <p>No source: created in machine-readable form.</p>
</sourceDesc>
```

```
<sourceDesc>
 <biblStruct xml:lang="FR">
  <monogr>
   <author>Eugène Sue</author>
   <title>Martin, l'enfant trouvé</title>
   <title type="sub">Mémoires d'un valet de
    chambre</title>
   <imprint>
    <pubPlace>Bruxelles et Leipzig</pubPlace>
    <publisher>C. Muquardt</publisher>
    <date value="1846">1846</date>
   </imprint>
  </monogr>
 </biblStruct>
</sourceDesc>
```

<encodingDesc>

- *encoding description*
- describes the relationship between an electronic text and its source or sources
- allows for detailed description of whether (or how) the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, etc.

<encodingDesc> (2)

The content of the encoding description may be a prose description, or it may contain elements from the following list, in the order given:

- <projectDesc> (project description)
- <samplingDecl> (sampling declaration)
- <editorialDecl> (editorial practice declaration)
- <tagsDecl> (tagging declaration)
- <refsDecl> (references declaration)
- <classDecl> (classification declarations)
- <fsdDecl> (FSD (feature-system declaration) declaration)
- <metDecl> (metrical declaration)
- <variantEncoding> declares method used to encode text-critical variants.

For more details, see the TEI P5 guidelines,
5.3 The Encoding Description

<profileDesc>

- *text profile*
- contains classificatory and contextual information about the text
- e.g. its subject matter, the individuals described by or participating in producing it, etc.
- of particular use in structured composite texts such as corpora, where it is often desirable to enforce a controlled descriptive vocabulary or to perform retrievals from a body of text in terms of text type or origin

For more details, see the TEI P5 guidelines,
5.4 The Profile Description

<revisionDesc>

- *revision history*
- allows the encoder to provide a history of changes made during the development of the electronic text
- important for version control and for resolving questions about the history of a file.

For more details, see the TEI P5 guidelines,
5.5 The Revision Description

Dublin core

- The Dublin Core Metadata Initiative (dublincore.org) was founded during a joint workshop of the National Center for Supercomputing Applications (NCSA) and the Online Computer Library Center (OCLC) held in Dublin, Ohio, March 1995.
- The aim was to create a core set of meta-data descriptions for Web-based resources that would be useful for categorizing the Web for easier search and retrieval.
- **Dublin Core Metadata Element Set (DCES)** defines 15 elements:
  - *Title*: A name given to the resource.
  - *Creator*: An entity primarily responsible for making the content of the resource.
  - *Subject*: The topic of the content of the resource.
  - *Description*: An account of the content of the resource.
  - *Publisher*: An entity responsible for making the resource available.
  - *Contributor*: An entity responsible for making contributions to the content of the resource.
  - *Date*: A date associated with an event in the life cycle of the resource.
  - *Type*: The nature or genre of the content of the resource.
  - *Format*: The physical or digital manifestation of the resource.
  - *Identifier*: An unambiguous reference to the resource within a given context.
  - *Source*: A Reference to a resource from which the present resource is derived.
  - *Language*: A language of the intellectual content of the resource.
  - *Relation*: A reference to a related resource.
  - *Coverage*: The extent or scope of the content of the resource.
  - *Rights*: Information about rights held in and over the resource.

Tools for corpus building

1. Google, BootCat
2. Perl, Python,…
3. XML, XSLT
4. tokenisers, taggers, aligners, parsers…
5. annotation editors

And exploitation

- PC concordancers:
  - WordSmith, ParaConc
- Web concordancers / viewers
  - With corpus: BNC View etc.
  - Engines: CQP, Manatee
  - Specialised: TIGER search
- Data extraction:
  - XSLT / XQuery
  - SAX + Perl, Java
  - Database
- Analysis:
  - Excel
  - SQL
- HLT uses..