

## Data representation

October 17<sup>th</sup>, 2012

Pesquisa de Informação em corpora, Lecture 4,  
Michel Génèreux

*I am grateful to Iris Hendrickx for letting me use and adapt her course material.*

2

## Raw material

- Corpus: sample of language use
- What type of material to start with?
  - raw text, html, pdf, ps, RTF, word documents
- Nowadays: digital corpora, but input can come from written material OCR documents
- Spoken: speech recording, transcribed speech, multimodal data → video

3

## Overview

### How do you represent your data?

- Data cleaning
- Encoding problems
- Mark-up

2

## Automatic conversion

Programs can convert:

Written material

- PDF → text
- Word → text
- HTML → text
- OCR → text

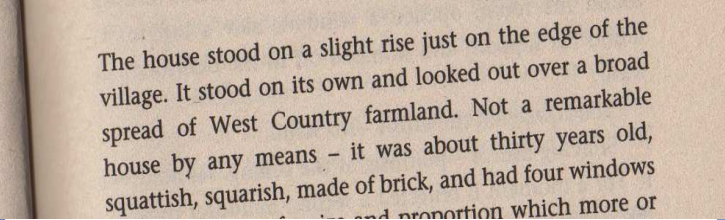
Spoken material

- Automatic speech recognition → text (in practice: let humans transcribe speech into clean text)

However: not straightforward, not without errors

4

## Optical Character Recognition



The house stood on a slight rise just on the edge of the village. It stood on its own and looked out over a broad spread of West Country farmland. Not a remarkable house by any means - it was about thirty years old, squattish, squarish, made of brick, and had four windows

The house stood on a slight rise just on the edge of the village. It stood on its own and looked out over a broad spread of West Country farmland. Not a remarkable house by any means - it was about thirty years old, squattish, squarish, made of brick, and had four windows

<http://www.onlineocr.net/>

5

## Textual raw input not clean

Input text is messed up with:

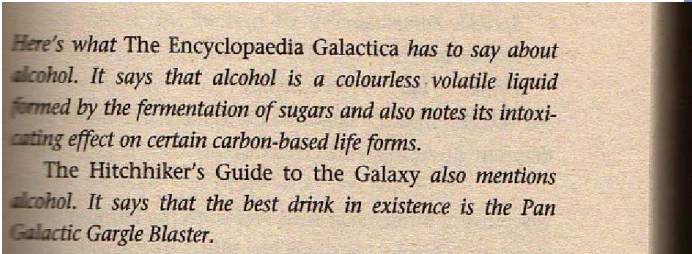
- advertisements
- photos, tables, graphics
- layout or design information
- disclaimers, copyright statements

HTML:

- related links, lists of links, java or php scripts, navigation bars

7

## OCR



*Here's what The Encyclopaedia Galactica has to say about alcohol. It says that alcohol is a colourless volatile liquid formed by the fermentation of sugars and also notes its intoxicating effect on certain carbon-based life forms.*

*The Hitchhiker's Guide to the Galaxy also mentions alcohol. It says that the best drink in existence is the Pan Galactic Gargle Blaster.*

licre's what The Encyclopaedia Galactica has to say about alcohol. It says that alcohol is a colourless - volatile liquid frned by the fennentation of sugars and also notes its intoxi-mting effect on certain carbon-based life forms. The Hitchhiker's Guide to the Galaxy also mentions dicohol. it says that the best drink in existence is the Pan Gabaic Gargle Blaster.

6

## Methods HTML corpus cleaning

Body Text Extraction (BTE) algorithm  
[Finn et al., 2001]

Finn's heuristic: the informative text is in parts where there are less HTML tags:

$Maximize N(words) - N(HTML\ tags)$

8

## HTML input

Meias finais » Itália 3 - Holanda 1

### Jogadores-chave de Itália

#### NESTA.

Maldini é um caso especial, por isso o destaque, mas Nesta também foi fantástico ontem. Desnecessário confirmar, numa olhadela rápida para os televisores que repetem as jogadas na bancada de Imprensa: sempre que houve um corte espectacular, foi ele quem o fez. Se quisermos ser justos para com Fernando Couto, podemos dizer que a Lazio tem os dois melhores centrais deste Europeu.



9

## HTML source code

```
<html><head>
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<title>O Jogo Online</title>
<SCRIPT LANGUAGE="JavaScript">
  var visibleVar="null";
.....
<td bgcolor="#000080"><font face="Arial" size="3" color="#AAAAAA">
<strong>Meias finais</strong></font> » <a href="artigo27787.htm"><font
face="Arial" size="3" color="#FFFFFF"><strong>Itália 3 - Holanda 1
</strong></font></a></strong></font></td> </tr>
</table>
```

```
<h1>Jogadores-chave de Itália</h1>
<p><br><b>NESTA.</b>
<br>Maldini é um caso especial, por isso o destaque, mas Nesta
também foi fantástico ontem. Desnecessário confirmar,
numa olhadela rápida para os televisores que repetem as jogadas na
bancada de Imprensa:
```

11

## HTML source code

```
<html><head>
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<title>O Jogo Online</title>
<SCRIPT LANGUAGE="JavaScript">
  var visibleVar="null";
.....
<td bgcolor="#000080"><font face="Arial" size="3" color="#AAAAAA">
<strong>Meias finais</strong></font> » <a href="artigo27787.htm"><font
face="Arial" size="3" color="#FFFFFF"><strong>Itália 3 - Holanda 1
</strong></font></a></strong></font></td> </tr>
</table>
```

```
<h1>Jogadores-chave de Itália</h1>
<p><br><b>NESTA.</b>
<br>Maldini é um caso especial, por isso o destaque, mas Nesta
também foi fantástico ontem. Desnecessário confirmar,
numa olhadela rápida para os televisores que repetem as jogadas na
bancada de Imprensa:
```

10

## Cleaned version

```
<h> Jogadores-chave de Itália
```

```
<p> Maldini é um caso especial, por isso o destaque, mas Nesta também foi
fantástico ontem. Desnecessário confirmar, numa olhadela rápida para os
televisores que repetem as jogadas na bancada de Imprensa: sempre que houve
um corte espectacular, foi ele quem o fez. Se quisermos ser justos para com
Fernando Couto, podemos dizer que a Lazio tem os dois melhores centrais deste
Europeu.
```

12

## CLEANEVAL competition

Shared task: cleaning web pages to prepare them for use as a linguistic corpus

CLEANEVAL 2007: Chinese and English

### Results:

- Development of many cleaner systems
- Compare systems against each other on the same data with the same evaluation method

13

## In conclusion

Data conversion & cleaning to obtain 'pure' text material are necessary, time consuming, error prone, and complicated steps in the process of corpus creation.

15

## Example of cleaners

- FIASCO - (D. Bauer et al., WAC3 2007)
- Victor (P. Pecina, LREC 2008)
- Ncleaner (S. Evert, LREC 2008)

	F-score	precision	recall
Baseline	88.72	83.11	95.15
NCLEANER (HTML)	92.73	94.70	90.83
NCLEANER (text)	90.18	90.30	90.05
Non-lexical (HTML)	92.31	91.65	92.97
Non-lexical (text)	89.86	89.88	89.85

Table 3: Micro-averaged evaluation results of the standard NCLEANER parameter files on the official CLEANVAL test set (percentages calculated at word level).

14

## Overview

- Data cleaning
- **Encoding problems**
- Mark-up

16

## In the beginning there was ASCII

- Computers understand binary: 1 and 0.
- **Character encoding**: translates a binary string to a character
- ASCII is a 7-bit encoding based on English alphabet

ASCII: American Standard Code for Information Interchange

17

## Extended ASCII: 256 chars

128	Ç	144	È	161	í	177	Ï	193	±	209	ƒ	225	ß	241	±
129	à	145	é	162	ò	178	Ï	194	±	210	ƒ	226	Γ	242	±
130	á	146	ê	163	ó	179	Ï	195	±	211	ƒ	227	π	243	±
131	â	147	ë	164	ô	180	Ï	196	±	212	ƒ	228	Σ	244	±
132	ã	148	ì	165	õ	181	Ï	197	±	213	ƒ	229	σ	245	±
133	ä	149	í	166	ö	182	Ï	198	±	214	ƒ	230	μ	246	±
134	å	150	î	167	÷	183	Ï	199	±	215	ƒ	231	τ	247	±
135	æ	151	ï	168	ø	184	Ï	200	±	216	ƒ	232	ϕ	248	±
136	ë	152	—	169	—	185	Ï	201	±	217	ƒ	233	Ω	249	±
137	è	153	Ö	170	—	186	Ï	202	±	218	ƒ	234	⊙	250	±
138	é	154	Û	171	¼	187	Ï	203	±	219	ƒ	235	δ	251	±
139	í	155	£	172	½	188	Ï	204	±	220	ƒ	236	∞	252	±
140	î	156	¥	173	¾	189	Ï	205	±	221	ƒ	237	ϕ	253	±
141	ï	157	—	174	—	190	Ï	206	±	222	ƒ	238	ε	254	±
142	À	158	—	175	—	191	Ï	207	±	223	ƒ	239	—	255	±
143	Á	159	—	176	—	192	Ï	208	±	224	ƒ	240	—	—	±

Source: www.LookupTables.com

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	MUL (null)	32	20	040	#32;	Space	64	40	100	#64;	@	96	60	140	#96;	`
1	1	001	SOH (start of heading)	33	21	041	#33;	!	65	41	101	#65;	A	97	61	141	#97;	a
2	2	002	STX (start of text)	34	22	042	#34;	"	66	42	102	#66;	B	98	62	142	#98;	b
3	3	003	ETX (end of text)	35	23	043	#35;	#	67	43	103	#67;	C	99	63	143	#99;	c
4	4	004	EOT (end of transmission)	36	24	044	#36;	\$	68	44	104	#68;	D	100	64	144	#100;	d
5	5	005	ENQ (enquiry)	37	25	045	#37;	%	69	45	105	#69;	E	101	65	145	#101;	e
6	6	006	ACK (acknowledge)	38	26	046	#38;	&	70	46	106	#70;	F	102	66	146	#102;	f
7	7	007	BEL (bell)	39	27	047	#39;	'	71	47	107	#71;	G	103	67	147	#103;	g
8	8	010	BS (backspace)	40	28	050	#40;	(	72	48	110	#72;	H	104	68	150	#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	#41;	)	73	49	111	#73;	I	105	69	151	#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	#42;	*	74	4A	112	#74;	J	106	6A	152	#106;	j
11	B	013	VT (vertical tab)	43	2B	053	#43;	+	75	4B	113	#75;	K	107	6B	153	#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	#44;	,	76	4C	114	#76;	L	108	6C	154	#108;	l
13	D	015	CR (carriage return)	45	2D	055	#45;	-	77	4D	115	#77;	M	109	6D	155	#109;	m
14	E	016	SO (shift out)	46	2E	056	#46;	.	78	4E	116	#78;	N	110	6E	156	#110;	n
15	F	017	SI (shift in)	47	2F	057	#47;	/	79	4F	117	#79;	O	111	6F	157	#111;	o
16	10	020	DLE (data link escape)	48	30	060	#48;	0	80	50	120	#80;	P	112	70	160	#112;	p
17	11	021	DC1 (device control 1)	49	31	061	#49;	1	81	51	121	#81;	Q	113	71	161	#113;	q
18	12	022	DC2 (device control 2)	50	32	062	#50;	2	82	52	122	#82;	R	114	72	162	#114;	r
19	13	023	DC3 (device control 3)	51	33	063	#51;	3	83	53	123	#83;	S	115	73	163	#115;	s
20	14	024	DC4 (device control 4)	52	34	064	#52;	4	84	54	124	#84;	T	116	74	164	#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	#53;	5	85	55	125	#85;	U	117	75	165	#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	#54;	6	86	56	126	#86;	V	118	76	166	#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	#55;	7	87	57	127	#87;	W	119	77	167	#119;	w
24	18	030	CAN (cancel)	56	38	070	#56;	8	88	58	130	#88;	X	120	78	170	#120;	x
25	19	031	EM (end of medium)	57	39	071	#57;	9	89	59	131	#89;	Y	121	79	171	#121;	y
26	1A	032	SUB (substitute)	58	3A	072	#58;	:	90	5A	132	#90;	Z	122	7A	172	#122;	z
27	1B	033	ESC (escape)	59	3B	073	#59;	;	91	5B	133	#91;	[	123	7B	173	#123;	[
28	1C	034	FS (file separator)	60	3C	074	#60;	<	92	5C	134	#92;	\	124	7C	174	#124;	\
29	1D	035	GS (group separator)	61	3D	075	#61;	=	93	5D	135	#93;	]	125	7D	175	#125;	]
30	1E	036	RS (record separator)	62	3E	076	#62;	>	94	5E	136	#94;	^	126	7E	176	#126;	^
31	1F	037	US (unit separator)	63	3F	077	#63;	?	95	5F	137	#95;	_	127	7F	177	#127;	DEL

Source: www.LookupTables.com

18

## How about the rest of the world?

corpus linguistics

कोष भाषा विज्ञान (hindi)

اللغويات الإحصار (arabic)

語料庫語言學 (chinese)

בלשנות קורפוס (hebrew)

การ ศึกษา ภาษาศาสตร์(thai)

Цорпус Лингуистиц (serbian)

20

## Every language its own encoding

- Efficient for one language, but incompatible with others
- Also different operating systems, software, regional settings, fonts led to incompatibility problems
- From the 1980's → work on unifying encodings

21

## UTF-8

- Implementations of Unicode:
  - UTF-8: most common encoding (compatibility with ASCII)
  - UTF-16
  - UTF-32
- For maximum compatibility (forward and backward) encode corpora in UTF-8
- Can handle pretty much anything
- Web minded corpora

23

## Unicode

- Universal standard for all writing systems
- +100.000 characters
- Independent of platform, software, vendor
- Represent characters in a descriptive way:
  - e.g. ä is "a + umlaut"
- Actual rendering of the character is done by the implementation (e.g. UTF-8)

22

## Overview

- Data cleaning
- Encoding problems
- **Mark-up**

24

## Text Markup

### Information about the text and its context

- Origin, location, genre, author, etc.

### Indications of how a text should look like

- Newlines, spaces, division in sections → also forms of textual markup
- Letter type and size, bold or italics
- In printing: page size, word hyphenation at the end of a line.

25

## How does it look?

You just saw an example of Markup for online documents: HTML

- Text appearance:

```
<b>this is bold</b>
```

```
<i>this is italics</i>
```

- Meta data, e.g. author information

```
<author>Fernando Pessoa </author>
```

```
<author id="Fernando Pessoa"> </author>
```

27

## Markup language

A set of markup conventions used together for encoding texts.

It defines:

- what is allowed
- what is required
- markup format
- meaning of the markup

26

## XML

- eXtensible Markup Language
- Meta language: formal description of a language

XML characteristics:

- descriptive markup
- has document type concept
- independent of hardware or software

28

## Descriptive vs. procedural

- **Procedural:** specific instructions of how to process a part of text

Proc markup: [LaTeX](#) and [HTML](#)

- **Descriptive:** assigns boundaries and a name to a piece of text

Actual processing is defined outside text

Desc markup: [CSS](#)

29

## Basic SGML/XML Concepts

Basic SGML/XML Concepts

- structured, semantic markup
- elements
- attributes
- entities

31

## SGML

### Standard Generalized Markup Language

- SGML provides a way to define markup languages and sets the standard for their form.
- XML is a simplification of SGML
- Nowadays: XML
- Every XML document is a valid SGML document.  
But not the other way around!
- SGML is more flexible:
  - end tags are optional
  - attributes with or without quotes

30

## Structured, semantic markup

- The markup is clearly separated from the text
- `<.>`
- Markup is written between these tags
- Markup has a hierarchical structure
- Markup expresses a meaning, an intention

32



## XML Elements

- Each XML unit is called 'element'
- Denoted with start and end tag
- Different elements have different names
  - start tag: <>
  - end tag: </>
  
  - <title>The SICK ROSE</title>

33

## Document types

- The type of a document is formally defined by its constituent parts and their structure.
  - A Document Type Definition (DTD)

35

## XML Attributes

- Each element can have attributes  
<element attribute="value">  
<section> </section>  
However, multiple sections in an article-> give them unique number:  
  
<section id="1" length="40"> In this first section we introduce ... </section>  
<section id="2" length="35"> In the next step we show how ... </section>

34

## Types of documents

Definition of a **report**:

- a title
- possibly an author
- followed by an abstract and a sequence of one or more paragraphs.

Anything lacking a title, according to this formal definition, would not formally be a report

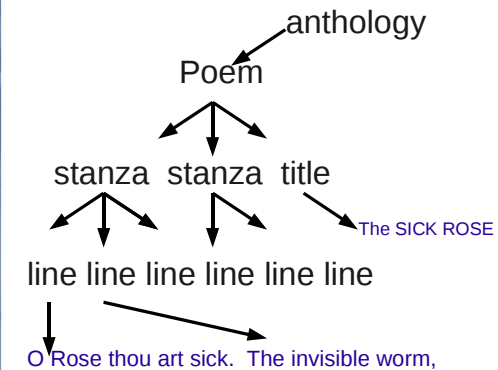
36

## Poem example

```
<anthology>
  <poem>
    <title>The SICK ROSE</title>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
    </stanza>
  </poem>
</anthology>
```

37

## Tree representation



39

## Description in words

- An anthology contains a number of poems and nothing else.
- A poem always has a single title element which precedes the first stanza and contains no other elements.
  - Apart from the title, a poem consists only of stanzas.
  - Stanzas consist only of lines and every line is contained by a stanza.
  - Nothing can follow a stanza except another stanza or the end of a poem.
  - Nothing can follow a line except another line or the start of a new stanza

38

## DTD Document Type Description

A DTD is expressed in SGML as a set of declarative statements, using a simple syntax defined in the standard.

```
<!ELEMENT anthology      (poem+) >
  <!ELEMENT poem          (title?, stanza+) >
  <!ELEMENT title          (#PCDATA) >
  <!ELEMENT stanza        (line+) >
  <!ELEMENT line           (#PCDATA) >
```

*#PCDATA is just a way of saying 'textual content'*

40

## Internal DTD Declaration

```
<?xml version="1.0"?>
<!DOCTYPE note [
<ELEMENT note (to,from,heading,body)>
<ELEMENT to (#PCDATA)>
<ELEMENT from (#PCDATA)>
<ELEMENT heading (#PCDATA)>
<ELEMENT body (#PCDATA)>
]>
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend</body>
</note>
```

<http://www.w3schools.com/dtd>

41

## DTD building blocks

- Elements: `<body>some text</body>`
- Attributes: ``
- Entities:
  - `&lt;` `&gt;` `&gt;` `&amp;` `&`
  - `&quot;` `"` `&apos;` `'`
- PCDATA: parsed text
- CDATA: unparsed text

<http://www.w3schools.com/dtd>

43

## External DTD Declaration

```
<?xml version="1.0"?>
<!DOCTYPE note SYSTEM "note.dtd">
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

---

```
<ELEMENT note (to,from,heading,body)>
<ELEMENT to (#PCDATA)>
<ELEMENT from (#PCDATA)>
<ELEMENT heading (#PCDATA)>
<ELEMENT body (#PCDATA)>
```

<http://www.w3schools.com/dtd>

42

## DTD – Elements (1)

- Declaring Elements
  - `<!ELEMENT elmt-name (elmt-content)>`
- Empty Elements
  - `<!ELEMENT element-name EMPTY>`
- Elements with Parsed Character Data
  - `<!ELEMENT element-name (#PCDATA)>`
- Elements with Children (*same sequence*)
  - `<!ELEMENT elmt-name (child1,child2,...)>`

<http://www.w3schools.com/dtd>

44

## DTD – Elements (2)

- Declaring Only One Occurrence of an Element:
  - <!ELEMENT element-name (child-name)>
- Declaring Minimum One Occurrence of an Element
  - <!ELEMENT elemt-name (child-name+)>
- Declaring Zero or More Occurrences of an Element
  - <!ELEMENT element-name (child-name\*)>

<http://www.w3schools.com/dtd>

45

## DTD – Attributes (1)

- Declaring Attributes
  - <!ATTLIST elmt-name att-name att-type default-value>
- The attribute-type:
  - CDATA: The value is character data
  - (en1|en2|..) Must be one from an enumerated list
- The default-value:
  - value The default value of the attribute
  - #REQUIRED The attribute is required
  - #IMPLIED The attribute is not required
  - #FIXED value The attribute value is fixed

<http://www.w3schools.com/dtd>

47

## DTD – Elements (3)

- Declaring Zero or One Occurrences of an Element
  - <!ELEMENT elemt-name (child-name?)>
- Declaring either/or Content
  - <!ELEMENT note (to,from,header, (message|body))>
- Declaring Mixed Content
  - <!ELEMENT note (#PCDATA|to|from| header|message)\*>

<http://www.w3schools.com/dtd>

46

## DTD – Attributes (2)

- <!ATTLIST square width CDATA "0">
- <!ATTLIST person number CDATA #REQUIRED>
- <!ATTLIST contact fax CDATA #IMPLIED>
- <!ATTLIST sender company CDATA #FIXED "Microsoft">
- <!ATTLIST payment type (check|cash) "cash">

<http://www.w3schools.com/dtd>

48

## Document type checking

If documents are of known types, a special purpose program (called a *parser*) can be used to process a document claiming to be of a particular type and check that all the elements required for that document type are indeed present and correctly ordered.

Different documents of the same type can be processed in a uniform way. Programs can be written which take advantage of the knowledge encapsulated in the document structure information, and which can thus behave in a more intelligent fashion.

49

## Stand-off vs. In-line

So far:

- In-line markup: placed inside the text

Alternative:

- Stand-off markup: outside the text, stored in another file

51

## XML Checking

On-line checkers available to validate XML:

- W3

[http://www.w3schools.com/dom/dom\\_validate.asp](http://www.w3schools.com/dom/dom_validate.asp)

- RUWF (are you well formed)

<http://www.xml.com/lpt/a/tools/ruwf/check.html>

- XML with DTD

<http://www.xmlvalidation.com>

- Linux: xmllint

50

## Advantages stand-off

- Allows levels of annotation with crossing branches (not possible in XML)
- Different levels of annotation without interfering with each other
  - different versions of same annotation
- New levels of annotation can be added later on without changing original
- Multiple people can annotate same data at the same time

52

## Example in-line

Era uma vez um príncipe ...

```
<sentence id="1">Era uma vez  
<markable id="aa" head="principe">um  
príncipe</markable>  
...  
</sentence>
```

53

## multiple annotation layers

As many different annotation layers as needed, each stored in separate file:

- file with **sentence** boundaries:  

```
<markable id="markable_1" span="word_1..word_10"  
mmax_level="sentences" />
```
- file with **noun phrases**:  

```
<markable id="markable_aa" span="word_4..word_5"  
head="principe" />
```

55

## Example stand-off

```
<words>  
<word id="word_1">Era</word>  
<word id="word_2">uma</word>  
<word id="word_3">vez</word>  
<word id="word_4">um</word>  
<word id="word_5"> príncipe</word>  
...  
</words>
```

54

## TEI=Text Encoding Initiative

**Goal:** developing a standard for digital text documents using SGML, and now of XML.

**One uniform independent standard format:**

- Search engines, editors, browsers, delivery-tools
- Parsers: automatic check whether the document has been encoded correctly

56

## TEI=Text Encoding Initiative

### Sponsors:

**ACH:** Association for Computers and the Humanities

**ACL:** Association for Computational Linguistics

**ALLC:** Association for Literary and Linguistic Computing

TEI Guidelines: 1300 pages

<http://www.tei-c.org/Guidelines/P5/>

57

## Core tags

### Some example core tags:

```
<TEI>
<teiHeader>

</teiHeader>
<text>
<body>
  <p> paragraph</p>
  <abbr>abbreviation<abbr>

</body>
</text>
</TEI>
```

59

## TEI Tagset

- **Core tag set:** standard components of the TEI main DTD in all its forms; these are always included .
- **Base tag sets:** basic building blocks for specific text types; (poetry, spoken, prose)
- **Additional tag sets:** extra tags useful for particular purposes. (transcription, names or dates, tables etc. )

58

## TEI Header

- **File description** *<fileDesc>* full bibliographic description for the source of electronic doc
- **a text profile** *<profileDesc>*
  - languages used
  - situation in which the text was produced,
  - participants (for speech),
  - topic or classification of the document,
  - demographic or social background of the authors, ...

60

## TEI Header

- Encoding description <encodingDesc>
- Revision description <revisionDesc> history of changes made during development of text. (version control)
- TEI header can be simple, or large and complex, depending on doc type

```
<teiHeader> ← simple minimal header
<fileDesc>
<!-- ... -->
</fileDesc>
</teiHeader>
```

61

## TEI Spoken DTD

- A spoken text may contain any of the following components:
- Utterances <u>
- Pauses <pause>
- Vocalized but non-lexical phenomena such as coughs <vocal>
- Kinesic (non-verbal, non-lexical) such as gestures <kinesic>
- Entirely non-linguistic events <event>
- etc.

63

## Speech annotation

- TEI has separate DTD for spoken text
- More in Lecture 5 on Spoken corpora
- here an example of TEI of speech  
*Sketch Monty python, My Theory*  
full description at:

[http://www1.uni-hamburg.de/exmaralda/files/demokorpus/MyTheory/export/MyTheory\\_TEI.xml](http://www1.uni-hamburg.de/exmaralda/files/demokorpus/MyTheory/export/MyTheory_TEI.xml)

62

## TEI speech example(a)

```
<TEI.2>
<teiHeader>
<fileDesc>
<titleStmt/>
  <sourceDesc>
    Miss Ann Elk is in a TV show to present her theory about the
    brontosaurus.
  </sourceDesc>
</fileDesc>
<profileDesc>
<particDesc>
  <person id="PRE"/>
  <person id="ELK"/>
</particDesc>
</profileDesc>
</teiHeader>
```

64



## TEI speech example(b)

```
<text>
<timeline>
  <when id="T0" absolute="0.0"/>
  <when id="T1" absolute="1.309974117691172"/>
  <when id="T2" absolute="1.899962460773455"/>
  <when id="T3" absolute="2.3399537674788866"/>
  ....
</timeline>
<event start="T0" end="T1" desc="((laughter, 1,3s)) " type="nn"/>
<u who="SPK0" start="T1" end="T2">
  <seg type="utterance" mode="declarative">
    <w>Good</w>
    <w>evening</w>
  </seg>
</u>
```

65

## SMIL

SMIL = Synchronized Multimedia Integration Language

- Based on XML
- --> multimedia databases

Description of the time-based structure of a multimedia object

Precise rendering on the screen

Links to multimedia objects

SMIL specification can be found here:

<http://www.w3.org/TR/REC-smil>

66