Spoken Texts, Linguistics and Dictionaries

TEI@Oxford

July 2009
A spoken text may contain any of the following components:

- utterances
- pauses
- vocalized but non-lexical phenomena such as coughs
- kinesic (non-verbal, non-lexical) phenomena such as gestures
- entirely non-linguistic incidents occurring during and possibly influencing the course of speech
- writing, regarded as a special class of incident in that it can be transcribed, for example captions or overheads displayed during a lecture
- shifts or changes in vocal quality
What sort of events?

Transcribed Events

Communicative

Oral

Lexical <u>

Non-Oral <kinesic>

Non-Communicative

<incident>

Non-lexical <vocal>
Grouping documents into a corpus allows you to factor out the metadata they have in common:

```xml
<teiCorpus>
  <teiHeader>
    <!-- shared metadata -->
  </teiHeader>
  <TEI>
    <teiHeader>
      <!-- specific metadata -->
    </teiHeader>
    <text>
      <!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```
The notion of "utterance"

- problematic, but pragmatic
- a sequence of speech from a single speaker
- may be grouped into higher-level `<div>`s
- or fragmented into smaller segments `<seg>` or `<s>`
- the `@who` attribute points to speaker information
Transcriptions of Speech

Elements defined: <broadcast>, <equipment>, <incident>, <kinesic>, <pause>, <recording>, <recordingStmt>, <scriptStmt>, <shift>, <u>, <vocal>, <writing>,

Classes defined: att.duration, model.divPart.spoken, model.global.spoken, model.recordingPart
Simple examples

Mixture of utterance and ‘paralinguistic’ information:

<u who="#Jan">This is just delicious</u>
<incident>
   <desc>telephone rings</desc>
</incident>
<u who="#Kim">I'll get it</u>
<u who="#Tom">I used to <vocal>
   <desc>coughs</desc>
</vocal> smoke a lot</u>
<u who="#Bob">
   <vocal>
      <desc>sniffs</desc>
   </vocal> He thinks he's tough
</u>
<vocal who="#Ann">
   <desc>snorts</desc>
</vocal>
<u who="#Tom">Yeah</u>
<kinesic>
   <desc>gives uplifted middle finger sign</desc>
</kinesic>
</u>
Back channelling

<u who="#a">So what could I have done <vocal who="#b"> 
<desc>tut-tutting</desc>
</vocal> about it anyway?</u>
Example using other TEI elements

<u who="#mar">you never <pause/> take this cat for show and tell</u>

<u who="#ros">yeah well I dont want to</u>

<incident>
  <desc>toy cat has bell in tail which continues to make a tinkling sound</desc>
</incident>

<u who="#ros">because it is so old</u>

<u who="#mar">how <choice>
  <orig>bout</orig>
  <reg>about</reg>
</choice>

<emph>your</emph> cat <pause/>yours is <emph>new</emph>

<kinesic>
  <desc>shows Father the cat</desc>
</kinesic>

<u trans="pause" who="#fat">thats <pause/> darling</u>

<u who="#mar">no <emph>mine</emph> isnt old mine is just um a little dirty</u>
Shifts in voice quality

- Classic multiple hierarchy problem
  - can use `<shift>` or `<milestone>` to mark boundaries...
  - ... or can use typed `<seg>` elements

- useful also for code shifting

```xml
<u who="#LB">  
  <shift feature="loud" new="f"/>Elizabeth
</u>
<u who="#EB">Yes</u>
<u who="#LB">  
  <shift feature="loud"/>Come and try this <pause/>
  <shift feature="loud" new="ff"/>come on
  <shift feature="code" new="fr-mru"/> 'tin va!
</u>
```
Sample prosodic feature list

(based on Boase, Survey of English Usage, 1990)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tempo</td>
<td>(fast, slow, getting faster, slower, etc.)</td>
</tr>
<tr>
<td>loud</td>
<td>loud, soft, getting louder, slower</td>
</tr>
<tr>
<td>pitch range</td>
<td>high, low, wide, narrow, ascending...</td>
</tr>
<tr>
<td>tension</td>
<td>slurred, tense, staccato, legato...</td>
</tr>
<tr>
<td>rhythm</td>
<td>regular, irregular, spiky rising or falling...</td>
</tr>
<tr>
<td>voice quality</td>
<td>whisper, husky, falsetto, giggle, sobbing, yawning, sighing...</td>
</tr>
</tbody>
</table>

Researchers need to define their own terms
<u who="#a">Listen to this <i>Listen to this</i> <shift new="reading"/>The government is confident, he said, that the current economic problems will be completely overcome by June<shift/> what nonsense!</u>
or as an <incident>

<u who="#a">Listen to this</u>

<incident>
    <desc>reads aloud from newspaper</desc>
</incident> what nonsense!</u>
<vocal> vs <u>

Compare:

<vocal who="#ann">
  <desc>snorts</desc>
</vocal>

and

<u who="#ann">
  <vocal>
    <desc>snorts</desc>
  </vocal>
</u>
<u who="#a">look at this</u>
<writing who="#a" type="newspaper" gradual="false">
Government claims economic problems <soCalled>over by June</soCalled>
</writing>
<u who="#a">what nonsense!</u>
Timing issues

- pausing: use `<pause>` element
- duration: use `@dur` attribute
- synchronization: use `@synch` attribute
- overlap: use `@trans` attribute
Okay <pause dur="PT2M"/> U-m <pause dur="PT75S"/> the scene opens up <pause dur="PT50S"/> with <pause dur="PT20S"/> um <pause dur="PT145S"/> you see a tree okay?
Mutt: Have you heard the --
Jeff: the election result?
Mutt: It's a disaster!

<u who="#mutt">have you heard the</u>
<u trans="latching" who="#jeff">the election result</u>
<u who="#mutt">its a disaster</u>
<u who="#jeff" trans="overlap">its a miracle</u>
More overlap

<u who="#tom">I used to smoke <anchor xml:id="TS-p10"/> a lot more than this <anchor xml:id="TS-p20"/> but I never inhaled the smoke</u>

<u start="#TS-p10" end="#TS-p20" who="#bob">You used to smoke</u>
Synchronization

<u who="#mutt">have you heard <anchor synch="#t1"/>the</u>
<u who="#jeff" synch="#t1">the election result</u>
<u who="#mutt" synch="#t2">its a disaster</u>
<u who="#jeff" synch="#t2">its a miracle</u>

<!-- Elsewhere in Document -->
<timeline origin="#t1">
  <when xml:id="t1"/>
  <when xml:id="t2"/>
</timeline>
Participant Description

<particDesc>
  <listPerson>
    <person xml:id="P-1234" sex="2" age="mid">
    </person>
    <person xml:id="P-4332" sex="1">
      <persName>
        <surname>Hancock</surname>
        <forename>Antony</forename>
        <forename>Aloysius</forename>
        <forename>St John</forename>
      </persName>
      <residence notAfter="1959">
        <address>
          <street>Railway Cuttings</street>
          <settlement>East Cheam</settlement>
        </address>
      </residence>
      <occupation>comedian</occupation>
    </person>
  </listPerson>
</particDesc>
<sourceDesc>
  <scriptStmt xml:id="CNN12">
    <bibl>
      <author>CNN Network News</author>
      <title>News headlines</title>
      <date when="1991-06-12">12 Jun 91</date>
    </bibl>
  </scriptStmt>
</sourceDesc>
Similarly for recordings...

<recordingStmt>
  <recording type="audio" dur="P30M">
    <respStmt>
      <resp>Location recording by</resp>
      <orgName>Sound Services Ltd.</orgName>
    </respStmt>
    <equipment>
      <p>Multiple close microphones mixed down to stereo Digital Audio Tape, standard play, 44.1 KHz sampling frequency</p>
    </equipment>
    <date>12 Jan 1987</date>
  </recording>
</recordingStmt>
<recording type="audio" dur="P10M">
  <equipment>
    <p>Recorded from FM Radio to digital tape</p>
  </equipment>
  <broadcast>
    <bibl>
      <title>Interview on foreign policy</title>
      <author>BBC Radio 5</author>
      <respStmt>
        <resp>interviewer</resp>
        <name>Robin Day</name>
      </respStmt>
      <respStmt>
        <resp>interviewee</resp>
        <name>Margaret Thatcher</name>
      </respStmt>
    </bibl>
  </broadcast>
</recording>
... and for settings

<setting xml:id="KDFSE002" n="063505" who="#PS0M6">
  <name type="place">Lancashire: Morecambe</name>
  <locale>at home</locale>
  <activity>watching television</activity>
</setting>
Linguistics

• associating simple linguistic analyses and interpretations with text elements
• semantic or syntactic interpretations which an encoder wishes to attach to all or part of a text
• mainly covering linguistic information
• as often in the TEI, you can do the same thing in many ways:
  • using generic <seg> elements with @type attributes
  • using the straightforward canned analyses described here
  • using the more powerful and general TEI Feature Structures
Linguistic units

To mark up text for linguistic purposes:

<s>  (s-unit) contains a sentence-like division of a text.
<br>  (clause) represents a grammatical clause.
<phr> (phrase) represents a grammatical phrase.
<w>  (word) represents a grammatical (not necessarily orthographic) word.
<m>  (morpheme) represents a grammatical morpheme.
<c>  (character) represents a character.
<pc> (punctuation character) represents a single punctuation mark.

From the att.segLike class, these elements all have @type and @function attributes.
Example of linguistic markup

Compare

<u>Like a suck of one of my sweets?</u>
<u>No I don't take sweets from strangers, oh God</u>

with....
Like a suck of one of my sweets?

No I don't take sweets from strangers, oh God.
Mixing analysis with structure

Analytic units often cross structural boundaries. The `<cl>` (clause) elements here cross the verse lines (`<l>`). We can use the `@part` attribute to show how a `<cl>` can be assembled:

```html
<div type="stanza">
  <l>
    <cl part="I">Tweedledum and Tweedledee</cl>
  </l>
  <l>
    <cl part="F">Agreed to have a battle;</cl>
  </l>
  <l>
    <cl part="I">For Tweedledum said Tweedledee</cl>
  </l>
  <l>
    <cl part="F">Had spoiled his nice new rattle.</cl>
  </l>
</div>
```
Phrase segmentation

<s>
  <cl type="finite-declarative" function="independent">
    <phr type="NP" function="subject">It</phr>
    <phr type="VP" function="predicate">
      <phr type="V" function="verb-main">was</phr>
      also
      <phr type="NP" function="predicate-nom.">
        a crucial year for me
      </phr>
    </phr>
  </cl>
</s>
Words with lemmas and morphemes with types

<s xml:lang="la">
  <w lemma="timeo">timeo</w>
  <w lemma="danaii">Danaos</w>
  <w lemma="et">et</w>
  <w lemma="donum">dona</w>
  <w lemma="fero">ferentes</w>
</s>

or

<w type="adjective">
  <m type="prefix" baseForm="con">com</m>
  <m type="root">fort</m>
  <m type="suffix">able</m>
</w>
Nested <w>

<s>
  <w>I</w>
  <w>did</w>
  <m>n't</m>
  <w>do</w>
  <w>it</w>
  <pc>!</pc>
</s>
The victim's friends told police that Kruger drove into the quarry and never surfaced.
Interpretation

<interpGrp type="POS">
  <interp xml:id="AT0">Definite article</interp>
  <interp xml:id="AV0">Adverb</interp>
  <interp xml:id="CJC">Conjunction</interp>
  <interp xml:id="CJT">Relative that</ interp>
  <interp xml:id="NN1">Noun singular</ interp>
  <interp xml:id="NN2">Noun plural</ interp>
  <interp xml:id="NP0">Proper noun</ interp>
  <interp xml:id="POS">Genitive marker</ interp>
  <interp xml:id="PRP">Preposition</ interp>
  <interp xml:id="VVD">Verb past tense</ interp>
</interpGrp>
The TEI defines a module for encoding human-oriented monolingual and multilingual dictionaries, glossaries, and similar documents. These are not just for standalone use, but could be for a wordlist or glossary accompanying a digital edition.
Dictionary Structures

• `<entry>` contains a reasonably well-structured dictionary entry
• `<entryFree>` (unstructured entry) contains a dictionary entry which does not necessarily conform to the constraints imposed by the entry element
• `<superEntry>` groups successive entries for a set of homographs
And other structures like...

- `<hom>` (homograph) groups information relating to one homograph within an entry
- `<sense>` groups together all information relating to one word sense in a dictionary entry, for example definitions, examples, and translation equivalents
Inside these structures

- `<form>` groups all the information on the written and spoken forms
- `<gramGrp>` groups morpho-syntactic information about a lexical item
- `<def>` contains a definition
- `<cit>` contains a cited quotation
- `<usg>` contains usage information
- `<xr>` contains a cross-reference
- `<etym>` encloses the etymological information
- `<re>` contains a related entry
- `<note>` contains a note or annotation.
<entry> example

<entry>
  <form>
    <orth>competitor</orth>
    <hyph>com|peti|tor</hyph>
    <pron>k@m"petit@(r)</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <def>person who competes.</def>
</entry>
Multiple senses

<entry>
  <form>
    <orth>disproof</orth>
    <pron>dIs"pru:f</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <sense n="1">
    <def>facts that disprove something.</def>
  </sense>
  <sense n="2">
    <def>the act of disproving.</def>
  </sense>
</entry>
Inside `<form>`

- `<orth>` gives the orthographic form
- `<pron>` contains the pronunciation(s)
- `<hyph>` contains a hyphenated form
- `<syll>` contains the syllabification
- `<stress>` contains the stress pattern
- `<lbl>` contains a label for a form, example, translation, or other piece of information
What? There is more inside `<form>`?

- `<gram>` for grammatical information
- `<gen>` identifies the morphological gender
- `<number>` indicates grammatical number
- `<case>` contains grammatical case
- `<per>` contains the grammatical person (1st, 2nd, 3rd, etc.)
- `<tns>` indicates the grammatical tense
- `<mood>` contains information about the grammatical mood of verbs
- `<iType>` indicates the inflectional class
<form>
  <orth>brag</orth>
</form>
<gramGrp>
  <pos>vb</pos>
</gramGrp>
<form type="infl">
  <orth>brags</orth>
  <orth>bragging</orth>
  <orth>bragged</orth>
</form>
<entry>
  <form>
    <orth>rémoulade</orth>
    <pron>Remulad</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
    <gen>f</gen>
  </gramGrp>
  <cit type="translation" xml:lang="en">
    <quote>remoulade</quote>
    <quote>rémoulade</quote>
    <def>dressing containing mustard and herbs</def>
  </cit>
</entry>