L'oral, et l'analyse linguistique

January 2010

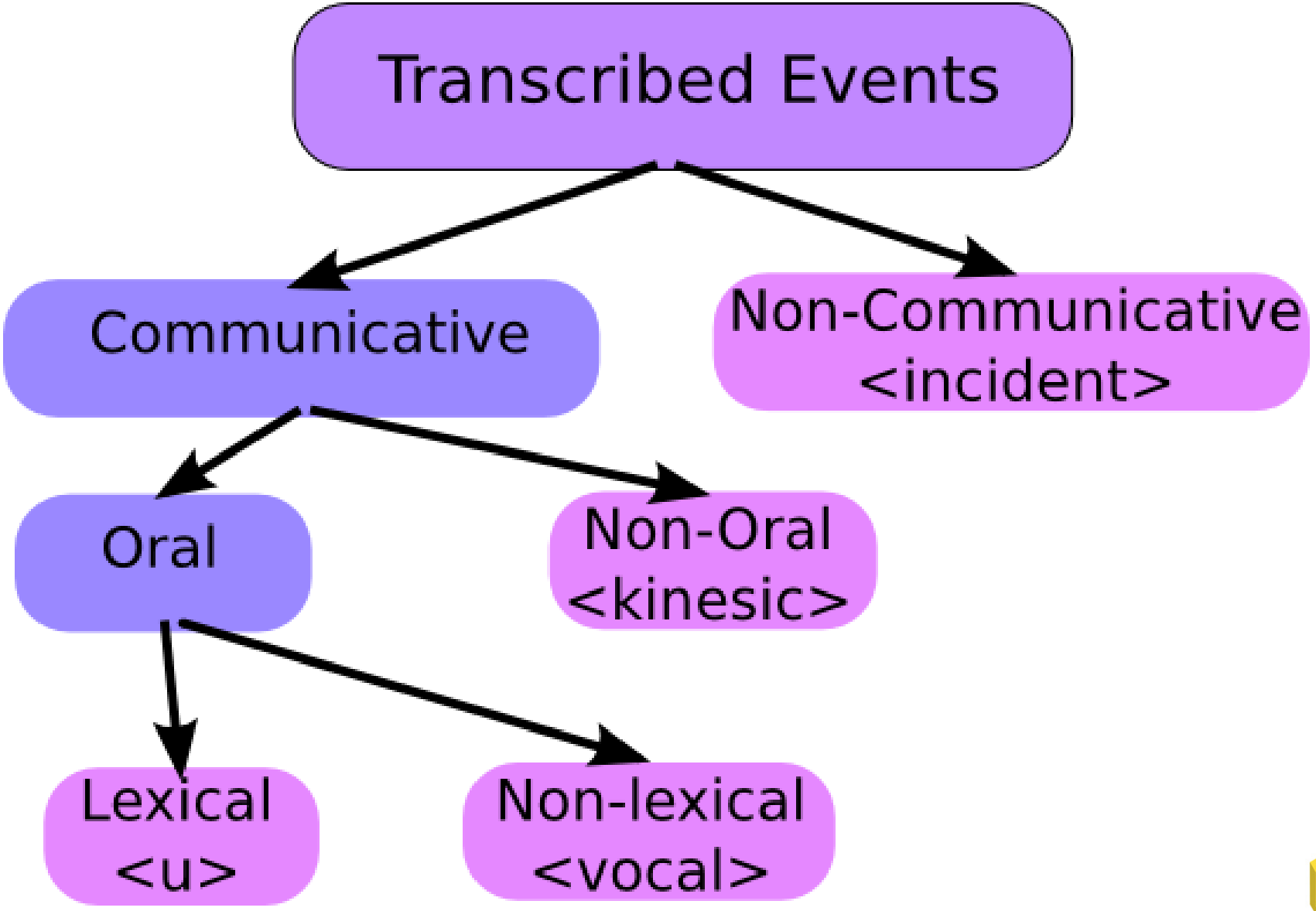
L'oral

Dans un texte oral on peut reconnaître :

- des énoncés (utterances)
- des pauses
- des phénomènes vocalisés mais pas lexicalisés, par exemple la toux, des quasi-mots comme "hein", "uh" etc.
- des phénomènes *kinésiques* (non vocalisés, non lexicaux, mais servant à communiquer) notamment les gestes
- des événements entièrement non linguistiques, mais ayant un effet sur le discours, par exemple un camion qui passe etc.
- des événements écrits, que l'on peut donc transcrire, par exemple des titres ou des diapositives affichés pendant une communication
- des changements dans la qualité de la voix, par exemple de volume

Un sous-ensemble des phénomènes de discours qui sont d'utilité lexicale

En résumé...


Transcribed Events
Communicative
Non-Communicative
<incident>
Oral
Non-Oral
<kinesic>
Lexical
<u>
Non-lexical
<vocal>

Propositions du module spoken

Des éléments supplementaire dans l'en-tête <broadcast>, <equipment>, <recording>, <recordingStmt>, <scriptStmt>

Des éléments pour la transcription de l'oral <incident>, <kinesic>, <pause>,<shift>, <u>, <vocal>, <writing>

Des attributs supplémentaires *@dur@start@end@sync* etc.

Par exemple...

```
<u who="#Jan">mmm delicieux</u>
<incident>
  <desc>téléphone sonne</desc>
</incident>
<u who="#Kim">j'y vais</u>
<u who="#Tom">ya longtemps <vocal>
    <desc>tousse</desc>
  </vocal> jne
fume plus</u>
<u who="#Bob">
  <vocal>
    <desc>sniffs</desc>
  </vocal>il se croit dur
</u>
<vocal who="#Ann">
  <desc>grognement</desc>
</vocal>
<u who="#Tom">oueh
<kinesic>
    <desc>fait un geste avec le doigt</desc>
  </kinesic>
</u>
<u who="#Bob">donc j'aurais dû <vocal who="#Ann">
    <desc>faisant tss-tss</desc>
  </vocal> faire quoi</u>
```

Le concept d'"énoncé"

- une séquence de discours d'un seul locuteur
- peut être regroupé dans des sections
- peut être fragmenté dans des segments <seg> ou <s>
- L'attribut @*who* sert à indiquer le locuteur

Exemple utilisant d'autres éléments TEI

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

Changements de voix

- Problème de chevauchement classique !
  - on peut se servir de la balise <shift> comme <milestone> pour indiquer les frontières...
  - ... ou des <seg> typés
- applicable également au "code shifting"

```
<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>
<listPerson type="speakers">
  <person xml:id="LB"/>
  <person xml:id="EB"/>
</listPerson>
```

Liste non exhaustive de caractéristiques en prose

(basée sur Boase, Survey of English Usage, 1990)

tempo	rapide, lent, de plus en plus rapide, de plus en plus lent, etc.
volume	fort, faible, de plus en plus fort, de plus en plus faible
hauteur	aigu, grave,...
tension	lié, tendu, staccato, legato...
rythme	régulier, irrégulier...
qualité de la voix	murmures, voix enrouée, voix de fausset, gloussements, sanglots, bâillements, soupirs...

Les chercheurs ont besoin de définir leur propre terminologie

<shift/> : exemple

```
<u who="#a">écoutez <shift new="reading"/>Matignon se déclare
confiant que les problèmes financiers actuels seront
entièrement maîtrisés fin juin<shift/> mon cul</u>
```

Ou bien :

```
<u who="#a">écoutez
<incident>
   <desc>lit à haute voix du journal</desc>
 </incident>mon cul</u>
```

<writing> exemple

```
<u who="#a">regardez ceci</u>
<writing who="#a" type="newspaper" gradual="false">
Matignon se déclare
<soCalled>confiant de maîtriser</soCalled> les problèmes financiers actuels
</writing>
<u who="#a">mon cul!</u>
```

Questions relatives à la temporalité

- pour les pauses : élément <pause>
- pour la durée : attribut @*dur*
- synchronisation : attribut @*synch*
- chevauchement : attribut @*trans*

<pause> : exemple

```
<u>Okay <pause dur="PT2M"/>U-m<pause dur="PT75S"/>la scène ouvre
<pause dur="PT50S"/> avec <pause dur="PT20S"/> um <pause dur="PT145S"/>
on voit un arbre okay?</u>
```

Chevauchement

```
Mutt: vous avez entendu l - -
Jeff: les résultats?
Mutt: quel désastre !
```

```
<u who="#mutt">vous avez entendu l</u>
<u trans="latching" who="#jeff">les résultats</u>
<u who="#mutt">quel désastre</u>
<u who="#jeff" trans="overlap">quel miracle </u>
```

Synchronisation

```
<u who="#mutt">vous avez entendu <anchor synch="#t1"/>l</u>
<u who="#jeff" synch="#t1">les résultats</u>
<u who="#mutt" synch="#t2">quel désastre</u>
<u who="#jeff" synch="#t2">quel miracle</u>
<!-- Elsewhere in Document -->
<timeline origin="#t1">
  <when xml:id="t1" since="00001728281"/>
  <when xml:id="t2" since="00001728302"/>
</timeline>
```

<timeline> : exemple

```
<timeline unit="s" origin="#TS-P1">
  <when xml:id="TS-P1" absolute="12:20:01"/>
  <when xml:id="TS-P2" interval="4:05" since="#TS-P1"/>
  <when xml:id="TS-P6"/>
  <when xml:id="TS-P3" interval="1:05" since="#TS-P6"/>
</timeline>
```

```
<u xml:id="TS-U1" start="#TS-P2" end="#TS-P3">This is my
<anchor synch="#TS-P6" xml:id="TS-P6A"/> turn</u>
```

Le début de l'énoncé TS-U1 s'aligne avec le point temporel TS-P2, il est donc 4.5 unités après TS-P1, i.e. à 12:24:06. Sa fin est synchronisée avec le point temporel TS-P3. La transition entre les mots 'my' et ' turn' arrive à un moment synchronisé avec le point temporerel TS-P6.

Description des participants

```
<particDesc>
 <listPerson>
  <person xml:id="P-1234" sex="2" age="mid">
   <p>informateur, sexe féminin, bonne éducation, née à Shropshire UK, 12
Jan 1950,
        commerçante parle français couramment. Statut
        socio-économique (SSE) :
        commerçante.</p>
  </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>
```

Description des enregistrements

```
<recordingStmt>
 <recording type="audio" dur="P10M">
  <equipment>
   <p>podcast</p>
  </equipment>
  <broadcast>
   <bibl>
    <title>Questions sur la souffrance et la santé au travail :
pénibilité, stress,
            dépression, harcèlement, maladies et accidents...</title>
    <author>France Inter</author>
    <respStmt>
     <resp>Présentateur</resp>
     <name>Alain Bédouet</name>
    </respStmt>
    <respStmt>
     <resp>Personne interrogée</resp>
     <name> Marie Pezé, Docteur en psychologie, psychanalyste, expert
judiciaire ; dirige
               la consultation « souffrance et travail » à l'Hôpital de
Nanterre (92), auteure de
      <title>ils ne mourraient pas tous mais tous étaient frappés</title>,
Editions
               Pearson.</name>
    </respStmt>
    <series>
```

Ou peut-être plus simplement...

```
<recordingStmt>
  <recording type="audio" dur="P15M" xml:id="rec-3001">
    <date>14 Feb 2001</date>
  </recording>
  <recording type="audio" dur="P15M" xml:id="rec-3002">
    <date>17 Feb 2001</date>
  </recording>
  <recording type="audio" dur="P15M" xml:id="rec-3003">
    <date>22 Feb 2001</date>
  </recording>
</recordingStmt>
```

...et pour le contexte

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

L'analyse, l'annotation

- Tout balisage resort d'une analyse; tout balisage s'exprime en annotation.
- Quand même, il y a un sentiment qu'il vaut mieux distinguer des assertions telles que

```
ceci est un paragraphe
```

  des assertions telles que

```
ceci est un nom verbal
```

- La standardisation des annotations linguistiques, en particulier à l'usage de la communauté du TAL, fut un des enjeux originaux de la TEI
- (bien que l'historique subséquent démontre que cet enjeu aurait pu être un peu ambitieux)
- La TEI propose une gamme de mécanismes génériques

Principes d'annotation linguistique (G.Leech, 1998)

- l'annotation devrait être séparable du texte
- des annotations multiples peuvent co-exister dans le texte
- l'annotation devrait être
  - auto-documentée
  - explicite
  - reproductible
  - formellement vérifiable

Diversité des annotations

Segmentation identification des segments, locations, et "empans"

Alignement et correspondance identification d'associations entre segments (e.g. équivalence, anaphore...)

Catégorisation classification des structures identifiées, e.g. classement morpho-syntaxique, fonction syntaxique, catégorie analytique

L'importance de la segmentation

- La segmentation permet l'identification et l'accès aux composants du texte à n'importe quel niveau
  - pour les référencer e.g. il y a un this à ...
  - pour définir un périmètre e.g. trouver un this au sein d'un that
  - pour l'analyse e.g. 90% des these de type that contiennent un the other

(des chevauchements peuvent se produire)

Éléments de segmentation

- génériques:
  - <s> segmentation de bout en bout (sur tout le texte)
  - <seg> segmentation arbitraire, éventuellement imbriquée
- de motivation linguistique :
  - <cl> (clause) proposition syntaxique
  - <phr> (phrase) syntagme
  - <w> (word) mot
  - <m> (morpheme) morphème
  - <c> (character) caractère

Tous ces éléments portent des attributs *@type*, *@subtype* et *@function*

L'annotation au niveau des mots ou des phrases est facile...

```
<p>
  <s>The export of sardines in oil from
     Sweden is prohibited. </s>
</p>
```

```
<s  n="11">
  <w>The</w>
  <w>export</w>
  <w>of</w>
  <w>sardines</w>
  <w>in</w>
  <w>oil</w>
  <w>from</w>
  <w>Sweden</w>
  <w>is</w>
  <w>prohibited</w>
  <c>.</c>
</s>
```

...les structures syntaxiques sont un peu moins évidentes

```
((The export of (sardines in
oil) (from Sweden)) is prohibited.)
```

```
((The export of (sardines (in
oil from Sweden)) is prohibited.)
```

```
<s  n="11">
  <seg>
    <w>The</w>
    <w>export</w>
    <w>of</w>
    <seg>
      <w>sardines</w>
      <w>in</w>
      <w>oil</w>
    </seg>
    <seg>
      <w>from</w>
      <w>Sweden</w>
    </seg>
  </seg>
  <w>is</w>
  <w>prohibited</w>
  <c>.</c>
</s>
```

La complexité intrinsèque est rendue explicite

bien que XML ait été conçu pour représenter des structures arborescentes linéarisées,

- il y a des problèmes de discontinuité et de chevauchement
- catégoriser les relations peut être problématique

Voyez la phrase ci-après...

```
<s>
  <cl>Some resentment
     is felt <phr>at the order</phr>
    <phr>by the
        Germans</phr>
  </cl>, <cl>who <phr>with their customary
        ingenuity</phr> have <phr>for some time</phr> been
     importing <phr>india-rubber sardines in
        petrol</phr>
    <phr>without detection</phr>
  </cl>
</s>
```

Discontinuité : utilisation des pointeurs

```
... (Germans, who (with their customary ingenuity) have
(for some time) been importing)...
```

```
<seg>
  <w xml:id="s1" next="#s2">who</w>
  <phr>with their customary ingenuity</phr>
  <w xml:id="s2" prev="#s1" next="#s3">have</w>
  <phr>for some time</phr>
  <w xml:id="s3" prev="#s2">been</w>
  <w>importing</w>
</seg>
```

On peut aussi utiliser l'attribut @*part* pour indiquer que les segments sont incomplets.

Discontinuité : utilisation des techniques de type "standoff markup"

```
<w  xml:id="W1">who</w>
<phr>with their customary ingenuity</phr>
<w  xml:id="W2">have</w>
<phr>for some time</phr>
<w  xml:id="W3">been</w>
<w>importing</w>
<!-....  -->
<join  targets="#W1 #W2 #W3"  result="seg"/>
```

Traduction

L'attribut *@corresp* est proposé pour la traduction :

```
<s corresp="#ALRTP1" xml:lang="EN" xml:id="RTP1">For a long time I used to
go to bed early</s>
<!-- ... -->
<s xml:id="ALRTP1" corresp="#RTP1" xml:lang="FR">Longtemps je me couchais
de bonne heure</s>
```

et/ou...

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

Référence anaphorique

L'attribut *@corresp* peut servir également à la résolution des anaphores :

```
<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 <seg corresp="#nbc">the network</seg> says
<seg corresp="shirl">the
show</seg> still is being considered.
```

ou, d'une manière externe ("standoff markup"):

```
<title xml:id="SHIRL">Shirley</title>, qui a débuté le vendredi soir il
n'y a seulement un mois, ne figure pas dans le listing
actuel du programme de <name xml:id="NBC">NBC</name>, bien que
<seg xml:id="NWK">le réseau</seg> prétende que
<seg xml:id="SHOW">ce spectacle</seg> est toujours prévu.


<linkGrp type="anaphor">
  <link targets="#SHIRL #SHOW"/>
  <link targets="#NWK #NBC"/>
</linkGrp>
```

L'attribut @*ana*

- fournit une manière générique d'associer un élément avec son analyse
- pointe sur une analyse, définie en se servant de n'importe laquelle des méthodes suivantes :
  - description pure en prose
  - un élément <interp>
  - une définition formelle en *structure de trait* (feature structure)

L'attribut @*type* propose une manière alternative de catégoriser les éléments.

Analyse au niveau des mots

```
<s  n="11">
  <w  ana="#DT">The</w>
  <w  ana="#NN">export</w>
  <w  ana="#IN">of</w>
  <w  ana="#NNS">sardines</w>
  <w  ana="#IN">in</w>
  <w  ana="#NN">oil</w>
  <w  ana="#IN">from</w>
  <w  ana="#NP">Sweden</w>
  <w  ana="#VBZ">is</w>
  <w  ana="#VVN">prohibited</w>
  <c  ana="#SENT">.</c>
</s>
```

Ceci implique qu'il existe quelque part une définition de DT, NN, IN etc. Par exemple:

```
<interp  xml:id="DT">
  <desc>déterminant</desc>
</interp>
<interp  xml:id="NN">
  <desc>nom au singulier</desc>
</interp>
<interp  xml:id="NNS">
  <desc>nom au pluriel</desc>
</interp>
<!-- ... etc -->
```

Ou bien...

```
<s  n="11">
  <w  type="DT">The</w>
  <w  type="NN">export</w>
  <w  type="IN">of</w>
  <w  type="NNS">sardines</w>
<!-- ... -->
</s>
```

Dans ce cas il faut définir un ODD où les valeurs possibles pour l'attribut *@type* seront spécifiées, de cette manière

```
<elementSpec  ident="w"  mode="change">
  <attList>
    <attDef  ident="type"  mode="replace">
      <valList>
        <valItem  ident="DT">
          <desc>déterminant</desc>
        </valItem>
        <valItem  ident="NN">
          <desc>nom au singulier</desc>
        </valItem>
        <valItem  ident="NNS">
          <desc>nom au pluriel</desc>
        </valItem>
<!-- ... -->
      </valList>
    </attDef>
  </attList>
</elementSpec>
```

La représentation en structures de trait

- La *structure de trait* est un concept bien connu en linguistique théorique.
- Toute analyse peut être representée par des paires de traits nom + valeurs, structurées d'une manière spécifique
- La représentation TEI a été standardisée au niveau de l'ISO; elle fournit une solution pragmatique et neutre aux problèmes de communication machine-machine dans le domaine du TAL industriel

Des exemples

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

Encore un exemple...

```
<fs type="word structure">
  <f name="spelling">
    <string>pólesi</string>
  </f>
  <f name="lemma">
    <string>pólis</string>
  </f>
  <f name="gloss" xml:lang="en">city</f>
  <f name="category">
    <symbol value="noun"/>
  </f>
  <f name="proper">
    <binary value="false"/>
  </f>
  <f name="gender">
    <symbol value="feminine"/>
  </f>
  <f name="inflection">
    <fs type="inflection structure">
      <f name="case">
        <symbol value="dative"/>
      </f>
      <f name="number">
        <symbol value="plural"/>
      </f>
    </fs>
  </f>
</fs>
```

Niveaux d'annotation (1)

Ces mécanismes très génériques peuvent s'appliquer à n'importe quel niveau de structuration

Par exemple, on peut distinguer les syntagmes :

```
<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>
```

Niveaux d'annotation (2)

On peut décomposer les mots :

```
<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>
```

ou

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

Niveaux d'annotation (3)

ou des unités de discours ...

```
<u xml:id="u1">Can I have ten oranges and a kilo of bananas please?</u>
<u xml:id="u2">Yes, anything else?</u>
<u xml:id="u3">No thanks.</u>
<u xml:id="u4">That'll be dollar forty.</u>
<u xml:id="u5">Two dollars</u>
<u xml:id="u6">Sixty, eighty, two dollars. Thank you.</u>
<spanGrp type="transactions">
  <span from="#u1">sale request</span>
  <span from="#u2" to="#u3">sale compliance</span>
  <span from="#u4">sale</span>
  <span from="#u5">purchase</span>
  <span from="#u6">purchase closure</span>
</spanGrp>
```