

Qu'est-ce que l'annotation et pourquoi en parle-t-on de manière si inquiétante ?

Lou Burnard
lou.burnard@tge-adonis.fr



Un corpus langagier peut-il exister à l'état "pur"?

- Un corpus langagier contient des échantillons du langage authentique
- sélectionné selon des principes explicites...
- avec des enjeux précis...
- ... et *représentés* sous forme numérisée

une représentation "pure" sans interprétation ou réduction, serait-elle donc possible?

L'annotation : un mal nécessaire ? ou un aspect incontournable ?

'Annotation ... is anathema to corpus-driven linguists.' (Aarts, 2002)

'The interspersing of tags in a language text is a perilous activity, because the text thereby loses integrity.' (Sinclair, 2004)

'...the categories used to annotate a corpus are typically determined before any corpus analysis is carried out, which in turn tends to limit ... the kind of question that usually is asked.' (Hunston, 2002)

- l'annotation présente les intuitions dans une forme codifiée
- on risque la circularité



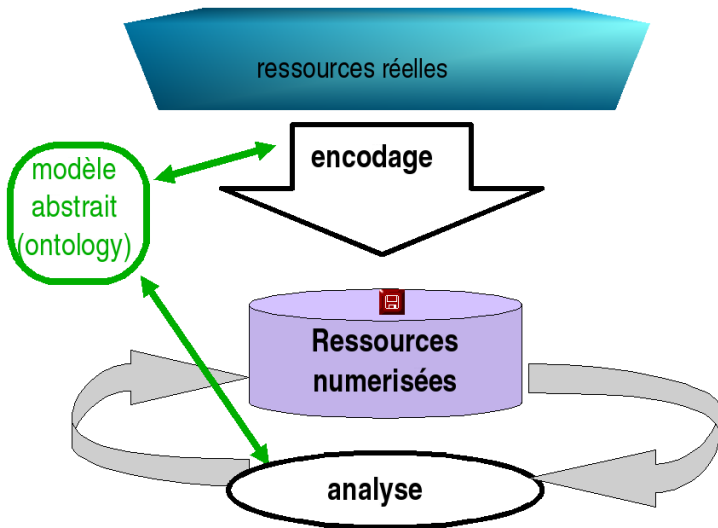
définition classique de l'annotation

the practice of adding interpretative linguistic information to a corpus (Leech 2005)

A noter: l'annotation est

- un acte interprétatif
- d'intérêt linguistique
- et son objectif serait : un corpus enrichi ... ou simplifié

Annotation: modèle classique



Principes d'annotation selon (Leech, 2005)

- the annotation should be separable from the text
- multiple annotations may co-exist within the text
- annotation should be
 - self-documenting
 - explicit
 - reproducible
 - formally verifiable



Quelques variétés d'annotation

- art-of-speech (POS)
- lemmatisation
- syntaxique
- sémantique
- de coréférences
- pragmatique
- stylistique
- prosodique
- évaluative (ex learner co)

L'annotation est liée étroitement avec une théorie langagière quelconque



Au niveau plus abstrait..

du point de vue technique, l'annotation implique

- une catégorisation des composants du corpus
- l'identification des composants supplémentaires
- l'identification des relations entre des composants spécifiés
 - des ensembles
 - des correspondences

Donc, tout système d' annotation ayant ces capacités devrait répondre aux besoins des annotateurs

Comment annoter?

à la main : danger d'incohérences; gros travaux ...

automatiquement : danger de circularité; nécessité de faire évoluer les outils

compromis britannique tous les deux ensemble

compromis google "crowd sourcing", "mechanical turk" etc.



Mechanical turk ?

https://www.mturk.com/mturk/welcome

Google Maps The Usual Press This bit.ly Sidebar BNC Simple Se... Legal sounds Zoho Invoice OxLIP+ - Find D... Other Bookmarks

amazon mechanical turk
Artificial Intelligence

Your Account HITs Qualifications

Already have an account?
Sign in as a [Worker](#) | [Requester](#)

Introduction | **Dashboard** | Status | Account Settings

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

288,430 HITs available. [View them now.](#)

Make Money

by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task

Find interesting and readily chosen tasks **TASKS** after government approval. Globally available.

Work

Find HITs Now

Earn money

Get Results

from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Register Now](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account

Load your tasks

Get Started

Get results

or [learn more about being a Worker](#)

[FAQ](#) | [Contact Us](#) | [Careers at Amazon](#) | [Developers](#) | [Press](#) | [Policies](#) | [Blog](#)

©2005-2011 Amazon.com, Inc. or its Affiliates

An **amazon.com** company

<http://blog.crowdfunder.com/2008/09/amt-fast-cheap-good-machine-learning/>

Exemples d'outils annotatifs

POS taggers

- Amalgam tagger : traduit entre plusieurs systèmes anglais d'annotation
<http://www.comp.leeds.ac.uk/amalgam/amalgam/ama>
- TnT (Tags and Trigrams):
<http://www.coli.uni-saarland.de/~thorsten/tnt/>
- TreeTagger : models for several languages
<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeT>

Parseurs

- TTT : boîte à outils utilisant l'XML
<http://www.ltg.ed.ac.uk/software/lt-ttt2/>
- Parseur avancé pour l'anglais
<http://nlp.stanford.edu/downloads/lex-parser.shtml>



Quelques exemples concrets

Un labo illustre d'Angleterre, à l'université de Lancaster, nous propose un "tasting plate" (plat de dégustation?) des variétés d'annotation qui y sont pratiquées:

<http://ucrel.lancs.ac.uk/annotation.html>

Des systèmes d'annotation qui ont évolué dans le temps, et qui sont dans leurs aspects notationnels entièrement *sui generis*...

part-of-speech (POS) tagging

hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN ,_, but_CC
not_XNOT when_WRB the_ATI guests_NNS have_HV to_TO sleep_VB
in_IN rows_NNS in_IN the_ATI cellar_NN !_!

- distinction mot/ annotation
- confusion mot/ ponctuation
- opacité des catégories

ou... en XML

```
<w pos="NN">hospitality</w>
<w pos="BEZ">is</w>
<w pos="AT">an</w>
<w pos="JJ">excellent</w>
<w pos="NN">virtue</w>
```

ou, également :

```
<w>
  <pos>NN</pos>
  <form>hospitality</form>
</w>
<w>
  <pos>BEZ</pos>
  <form>is</form>
</w>
```

ou meme :

```
<w pos="NN" form="hospitality"/>
<w pos="BEZ" form="is"/>
```

Grammatical parsing

```
[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N]
,_, [Fr [N who_PNQS N][V 'd_VHD grown_VVN [J too_RG
big_JJ [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1
Pier_NNL1 N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN
safely_RR [P at_II [N his_APP$ new_JJ home_NN1 [P in_II
[N Windsor_NP1 [ safari_NN1 park_NNL1 ]N]P]N]P]V] ._. S]
```

- syntaxe identique (labelled bracketting), mais notation tout à fait différente!
- redondance des labels

... ou en XML

```
<s type="S">
  <s type="N">
    <w pos="NP1">Nemo</w>
    <c>,</c>
    <s type="N">
      <w pos="AT">the</w>
      <w pos="NN1">killer</w>
      <w pos="NN1">whale</w>
    </s>
  </s>
  <s type="Fr">
    <s pos="N">
      <w pos="PNQS">who</w>
    </s>
    <s pos="V">
      <w pos="VHD">'d</w>
      <w pos="VVN">grown</w>
    </s>
    <s pos="J">
      <w pos="RG">too</w>
      <w pos="JJ">big</w>
    </s>
    <s type="P">
      <w pos="IF">for</w>
      <s type="N">
        <w pos="APP">his</w>
        <w pos="NN1">pool</w>
      </s>
    </s>
    <s type="P">
      <w pos="II">on</w>
      <s type="N">
        <w pos="NP1">Clacton</w>
        <w pos="NNL1">Pier</w>
      </s>
    </s>
  </s>
</s>
</s>
</s>
```


Semantic tagging

PPIS1	I	Z8
VV0	like	E2+
AT1	a	Z5
JJ	particular	A4.2+
NN1	shade	04.3
IO	of	Z5
NN1	lipstick	B4

- notation entièrement différente, et *ad hoc*

... mais simple à exprimer en XML

```
<w pos="PPIS1" cat="Z8">I</w>
<w pos="VV0" cat="E2" plus="Y">like</w>
<w pos="AT1" cat="Z5">a</w>
<w pos="JJ" cat="A4.2" plus="Y">particular</w>
<w pos="NN1" cat="04.3">shade</w>
<w pos="IO" cat="Z5">of</w>
<w pos="NN1" cat="B4">lipstick</w>
```

(Je reviens sur la question des catégories)

Annotation des anaphores

S.1 (0) The state Supreme Court has refused to release {1 [2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on bail . S.2 (1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction . S.3 (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision .

- syntaxe encore divergente
- besoin d'exprimer des correspondences ou liaisons

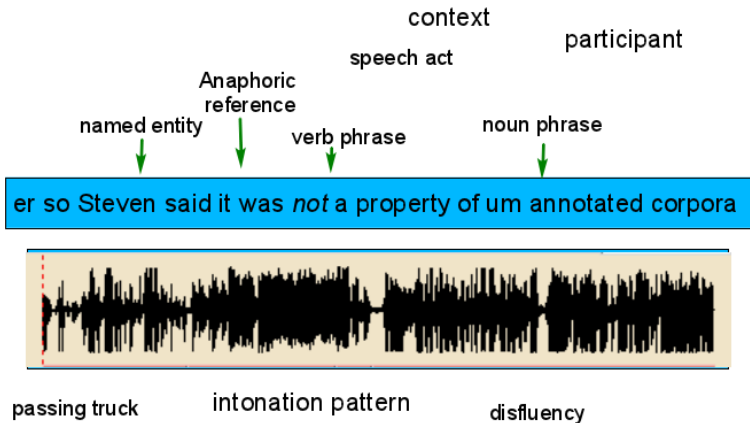
... ou en XML

```
<s n="1">
  <zero/>The state Supreme Court has refused to release
  <rs key="P1" type="copular">
    <rs key="P2">Rahway State Prison</rs> inmate </rs>
    <rs key="P1">James Scott</rs> on bail .
  </s>
<s n="2">
  <rs key="P1">The fighter</rs> is serving 30-40 years for a 1975 armed
  robbery conviction .
  </s>
<s n="3">
  <rs key="P1">Scott</rs> had asked for freedom while
  <rs key="P1" type="pronominal">he</rs> waits for an appeal decision .
  </s>
```

- ici, on se sert d'un balisage de coréférence (@key)
- un balisage explicite du linkage est également possible
- les relations syntaxiques différentes ne sont pas distinguées

Annotation de l'oral

Un cas extrême -- il n'y a pas de texte à annoter; on le crée en annotant.



Modélisation de l'oral

Outils de transcriptions les plus répandus :

- Anvil, CHAT, ELAN, EXMARaLDA, FOLKER, Praat, Transcriber.

Schmidt (2011) note que tous ces systèmes proposent un modèle commun, une simplification d'une *annotation graph* (Bird et Libermann):

- l'oral existe dans le temps : donc chaque morceau de transcription est associé avec un point de départ et une fin
- ces triplets sont regroupables en 'tiers' (couches?)
- une couche peut être associée avec un locuteur, et/ou un type

n'empêche un pluralité de moyens de l'exprimer..



EXMARaLDA, par exemple

EXMARaLDA Partitur-Editor 1.5.1 [S:\TP-Z2\Publikationen\jTEI_2010\Beispiel_EXMARaLDA.exb]

File Edit View Transcription Tier Event Timeline Format SFB 538/632 Help

très bien.

00:00 0.5 00:01.39

00:00 00:01 00:02 00:03 00:04 00:05

DS [sup] faster

DS [v] Okay. Très bien, très bien. Ah oui?

DS [en] Okay. Very good, very good.

DS [mv] right hand raised

FB [v] Alors ça dépend ((cough)) un petit peu.

FB [en] That depends, then, a little bit

FB [pho] [ʔtipə]

Done.

[13:38:37] Transcription S:\TP-Z2\Publikationen\jTEI_2010\Beispiel_EXMARaLDA.exb saved



EXMARaLDA: "Extensible Markup Language for Discourse Annotation" <http://www.exmaralda.org/>

Format EXMARaLDA

```
<common-timeline>
  <tli id="T0" time="0.0"/>
  <tli id="T1" time="1.309974117691172"/>
  <tli id="T2" time="1.899962460773455"/>
  <tli id="T3" time="2.3399537674788866"/> ....
</common-timeline>
<tier id="TIE0" speaker="SPK0" category="v" type="t" display-name="PRE
[v]">
  <event start="T2" end="T3">Good evening. </event>
  <event start="T5" end="T6">I have with me tonight Ann Elk Mistress Ann
Elk.
  </event>
</tier>
```


Voices of the Holocaust

```
<div xml:lang="de">  
  <u who="#boderD" start="127.732" end="x">[In German] Also, sagen Sie  
mir,  
  wie lautet Ihr Name, Frau Button?</u>  
  <u who="#buttonE" start="132.669" end="x">Deutsch sprechen?</u>  
  <u who="#boderD" start="135.403" end="x">Auf Deutsch.</u>  
  <u who="#buttonE" start="137.122" end="x">Ich heie Eda Button. Ich war  
deportiert von Athen im, h, April '44.</u>  
  <u who="#boderD" start="137.122" end="x">Und nach wo wurden sie  
deportiert?</u>  
  <u who="#buttonE" start="146.903" end="x">Ich war deportiert in h  
Bergen-Belsen.</u>  
  <u who="#boderD" start="149.496" end="x">Ja. Also, sagen Sie mal, h, wo  
ist  
  Ihr Mann?</u>  
  <u who="#buttonE" start="153.090" end="x">Mein Mann ist, h, weggelau-  
war,  
  h, in, h, Palstina, in Tel Aviv.</u>  
</div>
```

IFA Dialog Video corpus

```
<TIME_ORDER>
  <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="0"/>
  <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="10"/>
  <TIME_SLOT TIME_SLOT_ID="ts3" TIME_VALUE="462"/>
  <TIME_SLOT TIME_SLOT_ID="ts4" TIME_VALUE="840"/> ...
</TIME_ORDER>
<ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTA-
TION_ID="a1" TIME_SLOT_REF1="ts4" TIME_SLOT_REF2="ts7">
    <ANNOTATION_VALUE>beginnen we weer opnieuw?</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
</ANNOTATION>
```

Sentence Level

```
<u who=5 id=1 time=0.112> But I must see Mr <name> [smile again.]  
<u who=1 id=2 time=2.016> [<unclear> spoiled again?] ...  
</u>
```

Word level

```
<u who=5 id=1 time=0.112>  
<Audio word=BUT time=0.112 durn=0.176>But</Audio>  
<Audio word=I time=0.288 durn=0.064>I</Audio>  
<Audio word=MUST time=0.352 durn=0.304>must</Audio>  
<Audio word=SEE time=0.816 durn=0.352>see</Audio>  
<Audio word=MR time=1.168 durn=0.160>Mr</Audio> ...
```

Transcriber

```
<Turn speaker="spk2" startTime="0.557" endTime="5.851">  
  <Sync time="0.557"/> so what do you know of your family 's  
  <Sync time="2.255"/> history like <Sync time="3.410"/> do you know when and  
  why they came to Oxford  
</Turn>
```

Au niveau de transcription...

Même jeu ... On peut identifier (au moins) 5 systèmes de transcription pour les énoncés eux-mêmes:

HIAT	((coughs)) You must/ you (should) let • it be. ((laughs)) Pleease!
GAT	((coughs)) you must- you (should/could) let (-) it be; ((laughs)) plea:se-
CHAT	&=coughs you must... you should let # it be. &=laughs please!
DT1	(COUGH) you must-- you <X should X> let .. it be. @@ please?
cGAT	((coughs)) you must you (should/could) let (-) it be ((laughs)) please

(Table from Schmidt 2011)

ou, en TEI XML

```
<u>
  <seg type="interrupted">
    <incident>
      <desc>coughs</desc>
    </incident>
    <w>you</w>
    <w>must</w>
  </seg>
  <seg type="declarative">
    <w>you</w>
    <w>should</w>
    <w>let</w>
    <pause dur="short"/>
    <w>it</w>
    <w>be</w>
  </seg>
  <seg type="emphatic">
    <incident>
      <desc>laughs</desc>
    </incident>
    <w>please</w>
  </seg>
</u>
```

Méta-annotation

Revenons sur les catégories annotées:

```
<s>
  <w type="VVD">annotated</w>
  <w type="NN2">corpora</w>
  <w type="VV2">are</w>
  <w type="JJ1">cool</w>.
</s>
```

- un parser XML validera la structuration de ce fichier
- on désire également contrôler que chaque valeur d'attribut soit choisie dans un ensemble prédéfini
- on désire aussi fournir une explication des codes

```
<interp
  id="VVD"
  value="past tense adjectival form of lexical verb"/>
<interp id="NN2" value="plural form of common noun"/>
```

Définition des annotations

On pourrait donc représenter une taxinomie des POS-codes

```
<interpGrp xml:id="NN" value="common noun">  
  <interp xml:id="NN1" value="singular common noun"/>  
  <interp xml:id="NN2" value="plural common noun"/>  
</interpGrp>
```

Cette taxinomie peut être organisée de manière hiérarchique, distinguant par exemple les noms propres (NP) des noms communs (NN) :

```
<interpGrp value="nominal">  
  <interpGrp id="NN">  
    <interp id="NN1" value="singular common noun"/>  
    <interp id="NN2" value="plural common noun"/>  
  </interpGrp>  
  <interpGrp id="NP">  
    <interp id="NP1" value="singular proper noun"/>  
    <interp id="NP2" value="plural proper noun"/>  
  </interpGrp>  
</interpGrp>
```


Structures de traits

La TEI propose aussi un "meta-model" pour décrire n'importe quel système d'annotation linguistique... et de les rendre mutuellement compréhensibles, voire unifiables

Ce modèle opère à deux niveaux:

- **représentation** de l'analyse comme un lot de traits, typés, structurés, et linéarisés en XML
- définition d'un **système de traits**, représentant les contraintes sur les valeurs intégrées, et les règles à suivre pour les interpréter, surtout du point de vue d'une grammaire d'unification



Feature Structure Representation (ISO 24061)

```
<w ana="#NP1">ATILF</w>
<w ana="#NN2">corpora</w>
```

```
<fs xml:id="NP1">
  <f name="class">
    <symbol value="noun"/>
  </f>
  <f name="number">
    <symbol value="singular"/>
  </f>
  <f name="proper">
    <binary value="true"/>
  </f>
</fs>
<fs xml:id="NN2">
  <f name="class">
    <symbol value="noun"/>
  </f>
  <f name="number">
    <symbol value="plural"/>
  </f>
  <f name="proper">
    <binary valu="false"/>
  </f>
</fs>
```

Représentation simplifiée

Avec prédéfinition d'une librairie de traits ...

```
<fLib>
  <f name="class" xml:id="FCN">
    <symbol value="noun"/>
  </f>
  <f name="number" xml:id="FN1">
    <symbol value="singular"/>
  </f>
  <f name="number" xml:id="FN2">
    <symbol value="plural"/>
  </f>
  <f name="proper" xml:id="FPP">
    <binary value="true"/>
  </f>
  <f name="proper" xml:id="FPM">
    <binary value="false"/>
  </f>
</fLib>
```

... on arrive à simplifier cette représentation :

```
<fs xml:id="NN1" feats="FCN FPM FN1"/>
<fs xml:id="NN2" feats="#FCN #FPM #FN2"/>
<fs xml:id="NP1" feats="#FCN #FPP #FN1"/>
<fs xml:id="NN1" feats="#FCN #FPP #FN2"/>
```

Mais pourquoi réinventer la roue?

- ISO 12620:2009 Data Category Register : fournit une plateforme pour l'enregistrement des taxinomies linguistiques
- accessible on line ou comme service web
- démontre que ce n'est pas un problème insignifiant

```
<interp xml:id="NP1"  
  sameAs="http://www.isocat.org/datcat/DC-1892">  
  <desc>Nom propre au singulier</desc>  
</interp>  
<!-- .... -->  
<w ana="#NP1">ATILF</w>
```

C'est quoi un "noun"?

ISOcat

Welcome Guest Help

noun

My Workspace

- Public
 - Thematic Views
 - Metadata
 - Morphosyntax
 - Morphosyntax
 - Basics
 - Cases
 - FormRelated
 - MorphologicalFeaturesExclu
 - Operations
 - PartOfSpeech
 - RegisterDatingFrequency
 - Semantic Content Representa
 - Syntax
 - Language Resource Ontology
 - Lexicography
 - Language Codes
 - Terminology
 - Multilingual Information Manag
 - Lexical Resources
 - Lexical Semantics
 - Translation
 - Sign language
 - Audio
- Athena Core
- CLARIN-NL/VL
- Edisyn
- GOLD
- GilAndDan

PartOfSpeech

#	Name	Version	Administration stat	Registration status	Check	Type	Owned by	Scope
3854	manner noun	1:0	private	private		simple	Francopoulo, Gil	public
2277	mass noun	1:0	private	private	✓	simple	Francopoulo, Gil	public
1329	modal	1:0	private	private		simple	Francopoulo, Gil	public
1920	modal particle	1:0	private	private		simple	Francopoulo, Gil	public
1894	negative particle	1:0	private	private	✓	simple	Francopoulo, Gil	public
1925	negative pronoun	1:0	private	private		simple	Francopoulo, Gil	public
1333	noun	1:0	private	private	✓	simple	Wright, Sue Ellen	public
3852	number noun	1:0	private	private		simple	Francopoulo, Gil	public
1334	numeral	1:0	private	private	✓	simple	Francopoulo, Gil	public
1940	numeral approximation	1:0	private	private		simple	Francopoulo, Gil	public
1938	numeral both	1:0	private	private		simple	Francopoulo, Gil	public

noun - 1:0

Language French (fr)

2.3.1 Name Section

Name noun

Name Status standardized name

2.3.2 Definition Section

Definition Partie du discours attribuée aux mots qui désignent une personne, lieu, action, propriété ou chose etc. qui peut avoir des propriétés morphosyntaxiques comme le nombre ou le genre and des combinaisons syntaxiques comme la modification par un adjectif ou la détermination par un déterminant

Source adapté de 12620 avec les commentaires de Jan Odijk

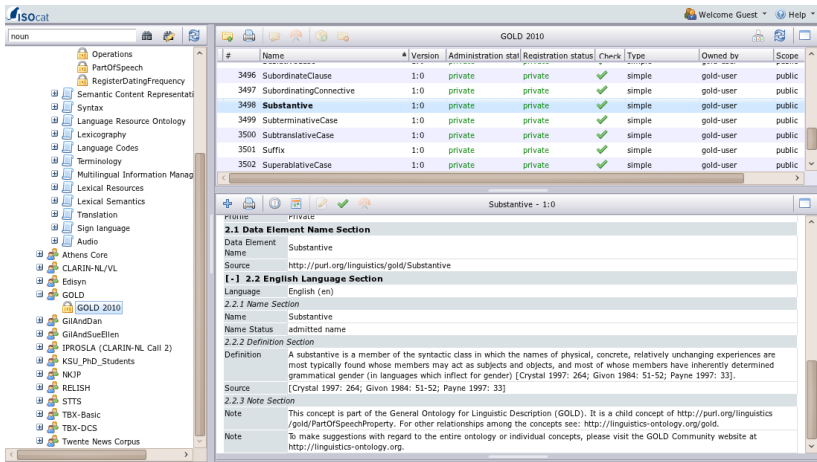
2.3.2.1 Note Section

Note Il est très difficile de caractériser les noms sémantiquement. Ajoutons que des propriétés dérivationnelles et compositionnelles spécifiques peuvent être utilisées pour distinguer les noms des autres parties du discours comme les prépositions et les déterminants

2.3.3 Example Section

Example table, présent, idée, Napoléon, Spiderman

C'est quoi un "substantive"?



The screenshot shows the ISOcat interface for the GOLD 2010 ontology. The left sidebar lists various categories, and the main window displays a table of ontology elements. The element 'Substantive' is highlighted, and its details are shown in the bottom pane.

#	Name	Version	Administration status	Registration status	Check	Type	Owned by	Scope
3496	SubordinateClause	1:0	private	private	✓	simple	gold-user	public
3497	SubordinatingConnective	1:0	private	private	✓	simple	gold-user	public
3498	Substantive	1:0	private	private	✓	simple	gold-user	public
3499	SubterminativeCase	1:0	private	private	✓	simple	gold-user	public
3500	SubtranslativeCase	1:0	private	private	✓	simple	gold-user	public
3501	Suffix	1:0	private	private	✓	simple	gold-user	public
3502	SuperlativeCase	1:0	private	private	✓	simple	gold-user	public

Substantive - 1:0

2.1 Data Element Name Section

Data Element Name: Substantive

Source: <http://purl.org/linguistics/gold/Substantive>

1-1 2.2 English Language Section

Language: English (en)

2.2.1 Name Section

Name: Substantive

Name Status: admitted name

2.2.2 Definition Section

Definition: A substantive is a member of the syntactic class in which the names of physical, concrete, relatively unchanging experiences are most typically found whose members may act as subjects and objects, and most of whose members have inherently determined grammatical gender (in languages which inflect for gender) [Crystal 1997: 264; Givón 1984: 51-52; Payne 1997: 33].

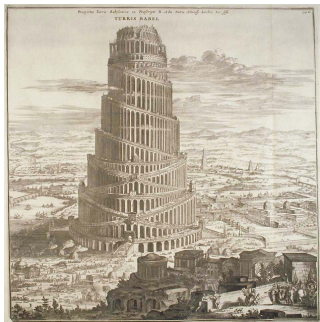
Source: [Crystal 1997: 264; Givón 1984: 51-52; Payne 1997: 33]

2.2.3 Note Section

Note: This concept is part of the General Ontology for Linguistic Description (GOLD). It is a child concept of <http://purl.org/linguistics/gold/PartOfSpeechProperty>. For other relationships among the concepts see: <http://linguistics-ontology.org/gold>.

Note: To make suggestions with regard to the entire ontology or individual concepts, please visit the GOLD Community website at <http://linguistics-ontology.org>.

Une tour (trop) bien connue



- Dans un seul labo vous avez vu une multiplication des systèmes de notation
- Entre labos divers, le cas est encore pire: le même phénomène linguistique se présentant dans des formats divergents — même si la notation est identique

Par conséquent...

- on ne peut pas mutualiser ou intégrer ses ressources
- chaque nouvelle ressource requiert un outil nouveau d'analyse
- les résultats scientifiques sont difficiles voire impossibles à répliquer ou tester

Donc on doit renoncer à prétendre faire de la science empirique :-)

Est-ce que XML nous fournit la réponse parfaite?

Oui!

- une technologie mûre, bien comprise, omniprésente
- beaucoup d'outils, des systèmes de développement, et de l'expertise
- bon compromis entre tractabilités humain/machine

Non!

- le modèle XML est trop restreint; trop axé sur le texte
- ne supporte pas les structures ayant plusieurs hiérarchies
- ne supporte guère les annotations imprécises ou incomplètes

Annotation graphs? RDF? microformats?



Problèmes de chevauchement

```
<lg>  
<l>Maître corbeau, sur un arbre perché</l>  
<l>Tenait en son bec un fromage.</l>  
<l>Maître renard par l'odeur alléché</l>  
<l>Lui tint à peu près ce langage:</l>  
<l>Hé! bonjour Monsieur du Corbeau</l>  
<l>Que vous êtes joli! que vous me semblez beau!</l> ...  
</lg>
```

avec superposition d'analyse linguistique...

```
<l>  
  <s part="I">Maître corbeau, sur un arbre perché</s>  
</l>  
<l>  
  <s part="F">Tenait en son bec un fromage.</s>  
</l>  
<l>  
  <s part="I">Maître renard par l'odeur alléché</s>  
</l>  
<l>  
  <s part="F">Lui tint à peu près ce langage:</s>  
</l>  
<l>  
  <s>Hé! bonjour Monsieur du Corbeau</s>  
</l>  
<l>  
  <s>Que vous êtes joli!</s>  
  <s> que vous me semblez beau!</s>  
</l> ...
```

Balisage débarqué

```
<lg>
  <l>
    <seg xml:id="S1">Maître corbeau, sur un arbre perché</seg>
  </l>
  <l>
    <seg xml:id="S2">Tenait en son bec un fromage.</seg>
  </l>
<!-- ... -->
</lg>
<join result="s" targets="#S1 #S2"/>
```

Au lieu de fournir des identifiants, on peut se servir de la syntaxe XPath pour les offset:

```
<lg>
  <l>Maître corbeau, sur un arbre perché</l>
  <l>Tenait en son bec un fromage.</l>
<!-- ... -->
</lg>
<join result="s" target="lg/l[1] lg/l[2]"/>
```

Overlap Happens

<i>tokens</i>	de	la	crème	glacé	
<i>phonemes</i>	dla		krEm	gla	se
<i>syntaxique</i>	P	NP			
<i>semantique</i>	some		icecream		

(Redrawn from Wörner et al, 2006)

Essentiellement, un problème de tokenisation/ annotation.

Une façon naturelle d'annoter?

Par exemple: voici du 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 a dollar forty.</u>  
<u xml:id="u5">Two dollars</u>  
<u xml:id="u6">Sixty, eighty, two dollars. Thank you.</u>
```

Ensuite, je veux catégoriser des unités:

```
<spanGrp type="transactions">  
  <span target="#u1">sale request</span>  
  <span target="#u2 #u3">sale compliance</span>  
  <span target="#u4">sale</span>  
  <span target="#u5 #u6">purchase</span>  
  <span from="#u6">purchase closure</span>  
</spanGrp>
```

L'annotation débarquée en 2 étapes

- une structuration basique, en unités (segments, tokens) identifiables
- un système pour mettre en relation les segments ainsi identifiés

les normes XPath et XInclude servent à compléter la sérialisation XML

Conclusions

- Tout corpus est intrinsèquement annoté
- l'annotation est un acte linguistique, interprétatif
- une structuration très simple peut supporter plusieurs niveaux d'annotation complexe

La quête d'une langue universelle d'annotation reste inachevée...



Bibliography

- Hunston, S. 2002 *Corpora in Applied Linguistics*
- Langendoen, D T and Gary F. Simons. 1995 *Rationale for the TEI Recommendations for Feature-Structure Markup* in N. Ide. and J. Veronis, eds. *The Text Encoding Initiative: Background and Contexts*
- Leech, G 2005 'Adding Linguistic Annotation', in M. Wynne, *Developing Linguistic Corpora: a Guide to Good Practice*
- Schmidt, T. 2011 'A TEI-based approach to standardising spoken language transcription' *Journal of the Text Encoding Initiative* 1 (June 2011).
- Sinclair, J. 2004 *Trust the Text: Language, Corpus and Discourse*
- Wörner, K., Witt, A., Rehm, G., Dipper, S. (2006). 'Modelling Linguistic Data Structures'. Presented at Extreme Markup Languages 2006, Montréal, Québec.

