Corpus Linguistics: corpus annotation

Karën Fort
karen.fort@inist.fr

November 30, 2010

Introduction

Sources

Most of this course is largely inspired by:

- Corpus Annotation [Garside et al., 1997],
- Annotation Science, from theory to practice and use [Ide, 2007].
- A Formal Framework for Linguistic Annotation [Bird and Liberman, 2000].
- Sylvain Pogodalla's course on the same subject [`http://www.loria.fr/~pogodall/enseignements/TAL-Nancy/notes-2008-2009.pdf`],

Annotation

Definition

"[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process" [Leech, 1997]

"Enhancing (raw) data with relevant linguistic annotations (relevant with what respect? Depends on the usage)" [Pogodalla]

"'Linguistic annotation' covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings - or it may be textual." [Bird and Liberman, 2000]

Scope [Bird and Liberman, 2000]

- morphological analysis
- POS tagging
- syntactic bracketing

Scope [Bird and Liberman, 2000]

- morphological analysis
- POS tagging
- syntactic bracketing
- co-reference marking
- 'named entities' tagging
- sense tagging

Scope [Bird and Liberman, 2000]

- morphological analysis
- POS tagging
- syntactic bracketing
- co-reference marking
- 'named entities' tagging
- sense tagging
- orthographic transcription
- phonetic segmentation and labeling
- disfluencies
- prosodic phrasing, intonation, gesture
- discourse structure

Scope [Bird and Liberman, 2000]

- morphological analysis
- POS tagging
- syntactic bracketing
- co-reference marking
- 'named entities' tagging
- sense tagging
- orthographic transcription
- phonetic segmentation and labeling
- disfluencies
- prosodic phrasing, intonation, gesture
- discourse structure
- phrase-level or word-level translation

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

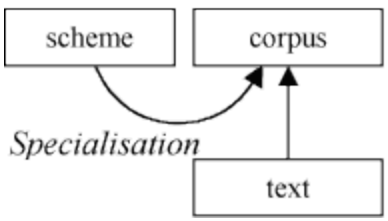
Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
7. No annotation schema should consider itself a standard (it possibly becomes one)

Different Methodological Stances

"you only get out what you put in" [Wallis, 2007]

Top-down approach



Knowledge is in the scheme $\Rightarrow$ the corpus is secondary

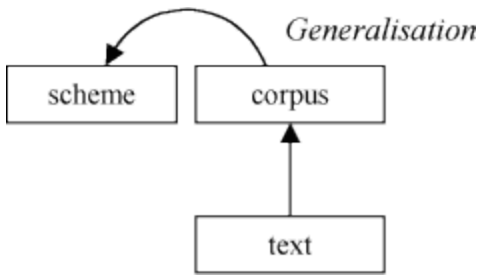
It's all in the annotation!

Top-down approach

- theory-led corpus linguistics
- problems arising in annotation are:
  - fixed by altering the algorithm or
  - excused as 'input noise' (performance)

→ NLP?

Bottom-up approach



Knowledge is in the text $\Rightarrow$ the corpus is primary [Sinclair]

Bottom-up approach

- data-driven corpus linguistics
- "*those who select facts from theory are ignoring linguistic evidence*"
- describe real linguistic utterances and the choices speakers make (not consider them as mere 'performance')
- annotation is secondary, if it has a status (!)
- there is no point in annotating or correcting the analysis (!)

$\rightarrow$ study of collocations, concordancing, lexical frames

Bottom-up approach

- **but** success of POS tagging!

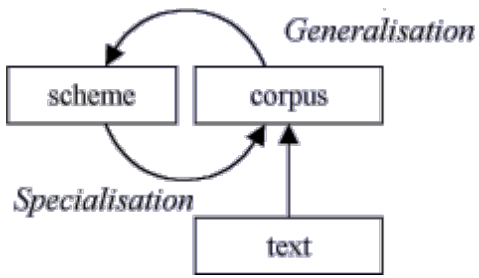
$\Rightarrow$ text-only position has been watered down!

- Today: "minimum necessary" annotation
- How much annotation is necessary/useful (see Active Learning)?

Pause for reflection

- first approach: allowed for POS tagging and parsing tools but too many frameworks, focus on rare issues
- second approach: necessary corrective

Third way?



Knowledge is in the scheme **and** in the corpus

Cyclic corpus annotation

- New observations generalise hypotheses
- Theory is needed to interpret and classify information

- Evolutionary circle: each loop enhances our knowledge by refining and testing our theories against real data

$\Rightarrow$ Both a more accurate corpus representation is constructed over time and a more sophisticated tagger (for example) is produced.

Issues

1. Juridical problems (text ownership) $\rightarrow$ not treated here.
2. Quality vs Cost
3. Technical problems (formats, recommendations, standardisation)

$\Rightarrow$ Reusability

Quality Cost?

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: ? words an hour, ? hours a day
- correcting skeleton "**treebanking**": ? words an hour, ? hours a day

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: 3,000 words an hour, ? hours a day
- correcting skeleton "**treebanking**": ? words an hour, ? hours a day

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: 3,000 words an hour, 3 hours a day
- correcting skeleton "**treebanking**": ? words an hour, ? hours a day

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: 3,000 words an hour, 3 hours a day
- correcting skeleton "**treebanking**": 750 words an hour, ? hours a day

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: 3,000 words an hour, 3 hours a day
- correcting skeleton "**treebanking**": 750 words an hour, 3 hours a day

Penn Treebank (PTB)

PTB 1 experiments on performance:

- correcting **POS tagging**: 3,000 words an hour, 3 hours a day
- correcting skeleton "**treebanking**": 750 words an hour, 3 hours a day
- + learning curve from 1 month (POS tagging) to 2 months (bracketing)!

Prague Dependency Treebank (PDT)

- 1996-2004 [Böhmová et al., 2001],
- built on the CNC (Czech National Corpus),
- 3-level structure:
  1. morphological (semi-automatic): 1.8 mil. tokens
  2. analytical (dependency syntax, with adapted tool)
  3. tectogrammatical (linguistic meaning using the Functional Generative Description): 1 mil. tokens

Prague Dependency Treebank (PDT)

Version 1.0:

- includes manual annotation of the morphological and analytical levels
- Time: ?
- Number of people involved: ?
- Cost estimate: ?

Prague Dependency Treebank (PDT)

Version 1.0:

- includes manual annotation of the morphological and analytical levels
- Time: 5 years
- Number of people involved: ?
- Cost estimate: ?

Prague Dependency Treebank (PDT)

Version 1.0:

- includes manual annotation of the morphological and analytical levels
- Time: 5 years
- Number of people involved: 22 people involved, with 17 simultaneously at the peak time
- Cost estimate: ?

Prague Dependency Treebank (PDT)

Version 1.0:

- includes manual annotation of the morphological and analytical levels
- Time: 5 years
- Number of people involved: 22 people involved, with 17 simultaneously at the peak time
- Cost estimate: $600,000

GENIA

GENIA: 400,000 words annotated in biology.

GENIA

GENIA: 400,000 words annotated in biology.

$\Rightarrow$ 5 part-time annotators, 1 senior coordinator, 1 junior coordinator for 1.5 year [Kim et al., 2008]

GENIA

GENIA: 400,000 words annotated in biology.

$\Rightarrow$ 5 part-time annotators, 1 senior coordinator, 1 junior coordinator for 1.5 year [Kim et al., 2008]

$\Rightarrow$ quality should be high!

Conclusion

Depends on the annotation and on the application!

Training, is, as of today, the best way to improve speed and quality of all annotations [Marcus et al., 1993, Chamberlain et al., 2008, Dandapat et al., 2009]

We'll see other solutions during the class on solutions for annotations (next class).

Which annotation formats do you already know?

Linear formats

('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD') [Brown corpus]

The DT the
TreeTagger NP TreeTagger
is VBZ be
easy JJ easy
to TO to
use VB use
. SENT .

PAT: <boy [*] no>[//] girl [/] girl truck # girl +... [CHILDES]

$\Rightarrow$ simple, but little expressivity (interpretation needed)

TEI (Text Encoding Initiative): history

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part
- Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

TEI (Text Encoding Initiative): objectives

+ give a standardised format for data exchange
+ give guidelines for encoding
+ be independent from applications
+ ? enable the encoding of any kind of information for any kind of text

TEI (Text Encoding Initiative): characteristics

+? provides multiple options for annotating a given phenomenon: **or** <p>

\+ SGML, then XML

\+ distinction between required practices, recommended practices and optional practices

+? provides ways for users to extend basic schemas

(X)CES: Corpus Encoding Standard

+ extends the TEI to provide a single representation format for **linguistic** annotations:
  + no more **or** <p>
  - ... but generic categories like <msd> (morpho-syntactic description), with linguistic annotation category in the attribute or tag content!
  ⇒ Specifications for linguistic category description is left to projects like EAGLES/ISLE (of which CES was a part)

++ standoff annotation

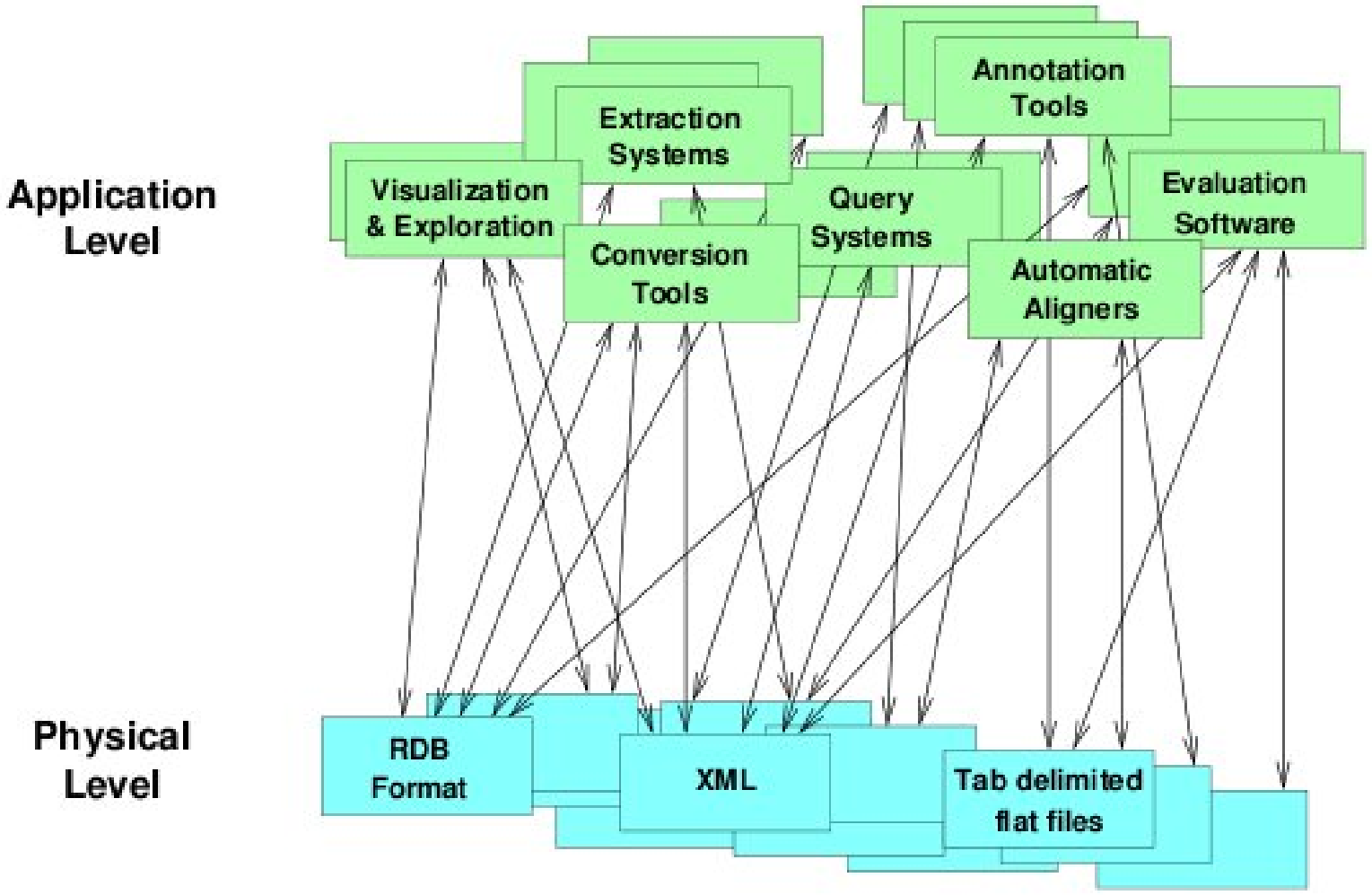
[Bird and Liberman, 2000]

- file formats, tags and attributes are secondary
- logical structure of annotations is primary (commonality appears here)

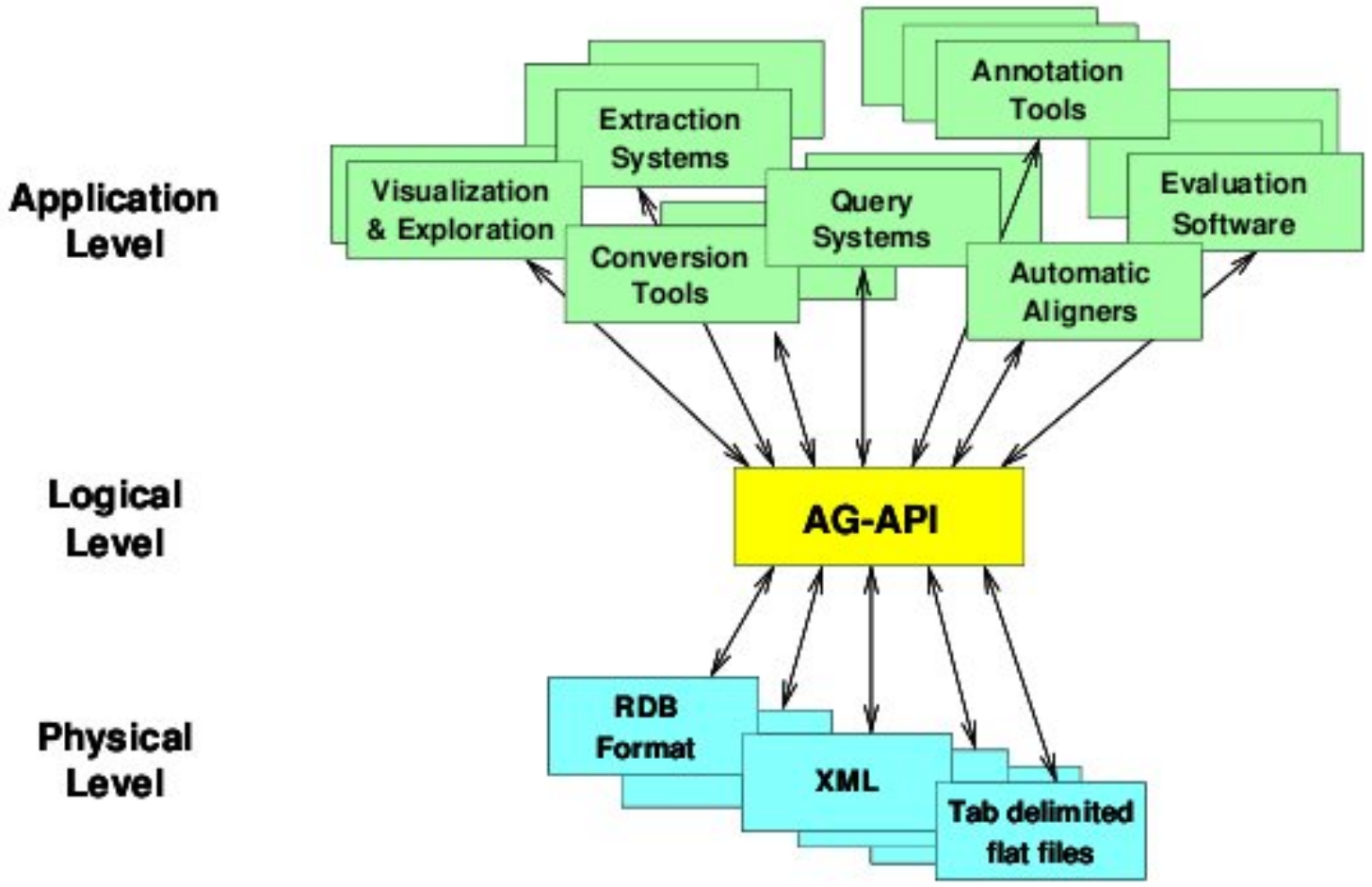
→ parallel made with DB systems:

  - interoperability
  - create and manipulate annotations according to your task/need/preferences
  - data independence principle

From Two-level Architectures...



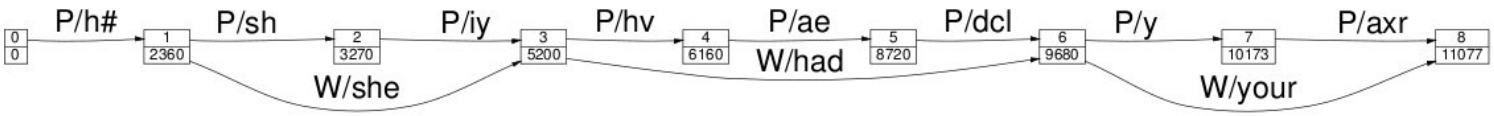
... To Three-level Architectures



Annotation Graphs for TIMIT

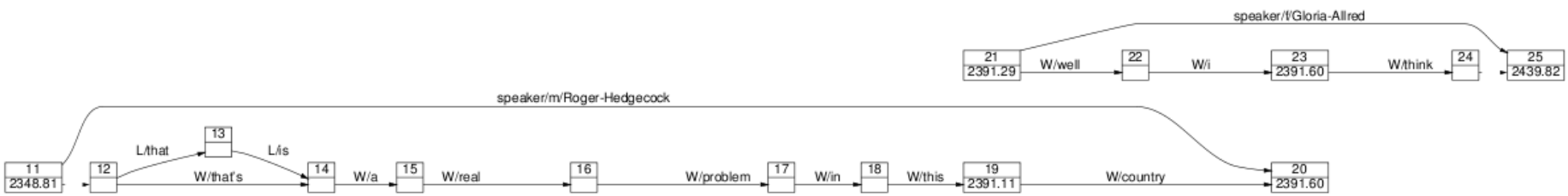
```
train/dr1/fjsp0/sa1.wrd:
2360 5200 she
5200 9680 had
9680 11077 your
11077 16626 dark
16626 22179 suit
22179 24400 in
24400 30161 greasy
30161 36150 wash
36720 41839 water
41839 44680 all
44680 49066 year
```

```
train/dr1/fjsp0/sa1.phn:
0 2360 h#
2360 3720 sh
3720 5200 iy
5200 6160 hv
6160 8720 ae
8720 9680 dcl
9680 10173 y
10173 11077 axr
11077 12019 dcl
12019 12257 d
...
```



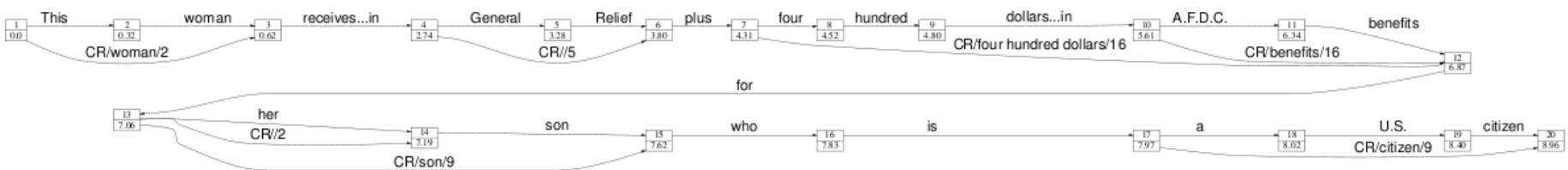
Annotation Graphs for UTF

```
<turn speaker="Roger_Hedgecock" spkrtype="male" dialect="native"
    startTime="2348.811875" endTime="2391.606000" mode="spontaneous" fidelity="high">
  ...
  <time sec="2378.629937">
  now all of those things are in doubt after forty years of democratic rule in
  <b_enamex type="ORGANIZATION">congress<e_enamex>
  <time sec="2382.539437">
  {breath because <contraction e_form="[you=>you]['ve=>have]">you've got quotas
  {breath and set<hyphen>asides and rigidities in this system that keep you
  <time sec="2387.353875">
  on welfare and away from real ownership
  {breath and <contraction e_form="[that=>that]['s=>is]">that's a real problem in this
  <b_overlap startTime="2391.115375" endTime="2391.606000">country<e_overlap>
</turn>
<turn speaker="Gloria_Allred" spkrtype="female" dialect="native"
    startTime="2391.299625" endTime="2439.820312" mode="spontaneous" fidelity="high">
  <b_overlap startTime="2391.299625" endTime="2391.606000">well i<e_overlap>
  think the real problem is that %uh these kinds of republican attacks
  <time sec="2395.462500">
  i see as code words for discrimination
  ...
</turn>
```



Annotation Graphs for Coreference

```
<COREF ID="2" MIN="woman">This woman</COREF> receives three hundred dollars a month under
<COREF ID="5">General Relief</COREF>, plus <COREF ID="16" MIN="four hundred dollars"> four
hundred dollars a month in <COREF ID="17" MIN="benefits" REF="16">A.F.D.C.  benefits</COREF></COREF>
for <COREF ID="9" MIN="son"><COREF ID="3" REF="2">her</COREF>son</COREF>, who is
<COREF ID="10" MIN="citizen" REF="9">a U.S. citizen</COREF>.
<COREF ID="4" REF="2">She</COREF>'s among <COREF ID="18" MIN="aliens">an estimated five hundred
illegal aliens on <COREF ID="6" REF="5">General Relief</COREF> out of
<COREF ID="11" MIN="population"><COREF ID="13" MIN="state">the state</COREF>'s total illegal
immigrant population of <COREF ID="12" REF="11"> one hundred thousand </COREF></COREF></COREF>
<COREF ID="7" REF="5">General Relief</COREF> is for needy families and unemployable adults who
don't qualify for other public assistance.  Welfare Department spokeswoman Michael Reganburg says
<COREF ID="15" MIN="state" REF="13">the state</COREF> will save about one million dollars a year
if <COREF ID="20" MIN="aliens" REF="18">illegal aliens</COREF> are denied
<COREF ID="8" REF="5">General Relief</COREF>.
```



Annotation Graphs [Bird and Liberman, 2000]

- Directed Acyclic Graphs (DAGs) $\Rightarrow$ expressive power
- with fielded records on the arcs
- with optional time references on the nodes

Linguistic Annotation Framework, LAF [Ide and Romary, 2006]

- ISO TC37 SC4 standard project (or standard?)
- aims at:
  1. accommodating all types of linguistic annotations
  2. providing means to represent complex linguistic information

LAF principles

- separation of data (read-only) and annotations (stand-off)
- separation of user annotation format and exchange format (mappable)
- separation of structure and content in the exchange format (list = alternatives or inclusive or prioritized list?)

$\Rightarrow$ annotation = directed graph, instantiated in XML (TEI)

GrAF: Application of LAF

While AGs allow to represent layers of annotation, each associated with primary data...
... GrAF allow for annotations linked to other annotations (multiple annotations form a single graph)

Formats vs Schemes

TEI
is XML

Formats vs Schemes

TEI
is XML
is Tree-structured?

Formats vs Schemes

TEI
is XML
is Tree-structured?

LAF
is DAG

Formats vs Schemes

TEI
is XML
is Tree-structured?

LAF
is DAG
is Graph-structured?

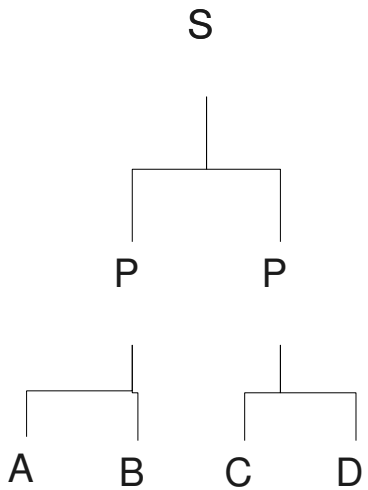
Formats vs Schemes

TEI
is XML
is Tree-structured?

LAF
is DAG
is Graph-structured?
in TEI??

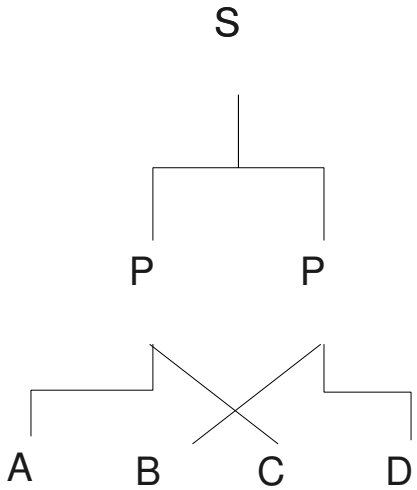
is XML about Syntax or Semantics?

Trees vs Graphs



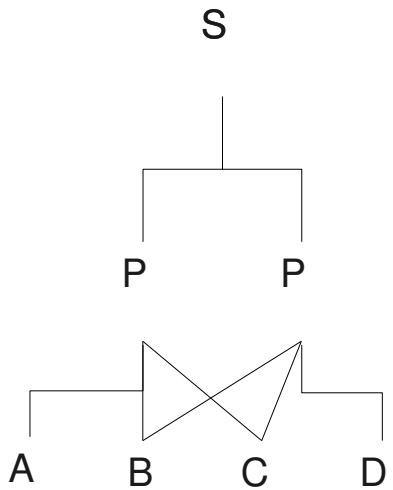
Tree

XML



Tree

Decorrelated XML



Graph

Decorrelated XML

Structure vs Interpretation

- XML allows to represent both trees and graphs
- interpretation is in the structure or outside the structure: expressivity
- XML expressivity is limited $\Rightarrow$ use stand-off annotations (decorrelated)

Conclusion

- evolution towards more complex (semantic) annotations
- evolution towards the use of non-expert annotators for simple annotations
- from trees to graphs: evolution towards more expressivity
- still room for more methodology!

- Annotation costs and solutions
- Methodology
- Structure vs Interpretation

- Read carefully: [Dandapat et al., 2009] (http://www.aclweb.org/anthology/W/W09/W09-3002.pdf)
- Apply the grid we saw in the second course to this article.


Bird, S. and Liberman, M. (2000).
A Formal Framework for Linguistic Annotation (revised version).
CoRR, cs.CL/0010033:pp 23–60.


Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).
The prague dependency treebank: Three-level annotation scenario.
In Abeillé, A., editor, Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.


Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).
Phrase Detectives: a Web-based Collaborative Annotation Game.
In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz.


Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
Complex Linguistic Annotation - No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks.

In Proceedings of the third ACL Linguistic Annotation Workshop.


Garside, R., Leech, G., and McEnery, T., editors (1997).
Corpus Annotation: Linguistic Information from Computer Text Corpora.
Longman, London.


Ide, N. (2007).
Annotation science: From theory to practice and use. (invited talk) data structures for linguistics resources and applications.
In Proceedings of the Bienniel GLDV Conference, Tübingen, Germany.


Ide, N. and Romary, L. (2006).
Representing linguistic corpora and their annotations.
In Proceedings of the Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy.


Kim, J.-D., Ohta, T., and Tsujii, J. (2008).

Corpus annotation for mining biomedical events from literature.
BMC Bioinformatics, 9(1):10.


Leech, G. (1993).
Corpus Annotation Schemes.
Literary and Linguistic Computing, 8(4):275–281.


Leech, G. (1997).
Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, London.


Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of english : The penn treebank.
Computational Linguistics, 19(2):313–330.


Wallis, S. (2007).
Annotating Variation and Change, chapter Annotation, Retrieval and Experimentation.

Varieng, University of Helsinki, Helsinki, Finland.