Corpus Linguistics
(L615)
Linguistic Annotation

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015

Corpus mark-up

Corpus mark-up is information about the corpus included in the corpus itself

- Some of it governs formatting, printing, other processing
- Mark-up is important to:
  - help recover the original context
  - allow for a wider range of questions to be explored (e.g., sociolinguistic variables)
  - encode extralinguistic details (e.g., laughs, formulas)

Motivating Annotation

Why would we want annotation?

- For training NLP tools
- For finding examples
  - what is the plural form of fish?
  - which nouns can occur as bare nouns, without a determiner?
  - are there subjectless sentences in German?
    - Yes, e.g., Mir ist kalt. (‘To me is cold.’)
  - is it possible in English to have something between a noun and its modifying relative clause?

Annotation makes it possible to find phenomena that would otherwise disappear in masses of data

Potential benefits of annotation

Corpus annotation adds much value to a corpus

- Extracting information is easier:
  - Can easily isolate left in its adjectival uses
  - Can find information on a language you don’t speak
- Corpus is more reusable
  - Insights are more accessible to others
  - Corpus is more multifunctional
  - Annotation provides a clear record of analysis, open to future scrutiny
    - Decisions are more reproducible

Potential criticisms of annotation

- Annotation is cluttered
  - But: it can be easily ignored
- Annotation imposes one particular linguistic analysis
  - But: users can reject the analysis, and annotation at least makes the analysis process clearer
- Annotation makes a corpus less accessible & expandable
  - But: corpora can be extended without annotation, if need be
- Annotation cannot be done completely consistently
  - But: humans are fallible, and there are ways to ensure better consistency

Corpus annotation

Corpus annotation = “interpretative, linguistic information” added to a corpus

- Corpus mark-up is relatively objective factual information
- Corpus annotation is more subjective, interpretative information

Corpus mark-up (Review)

Corpus mark-up is information about the corpus included in the corpus itself

- Some of it governs formatting, printing, other processing

Mark-up is important to:

- help recover the original context
- allow for a wider range of questions to be explored (e.g., sociolinguistic variables)
- encode extralinguistic details (e.g., laughs, formulas)
Leech’s Seven Maxims of Annotation

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
2. It should be possible to extract the annotations by themselves from the text. (flip side of maxim 1)
3. The annotation scheme should be based on guidelines which are available to the end user.
4. It should be made clear how and by whom the annotation was carried out.

Leech’s Seven Maxims (2)

5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
7. No annotation scheme has the a priori right to be considered as a standard. Standards emerge through practical consensus.

How corpus annotation is achieved

Some options:
- Fully manually
  - Pro: has the potential to be of highest quality
  - Con: time-consuming + humans are prone to errors
- Fully automatically (assuming appropriate technology)
  - Pro: quick and consistent
  - Con: will often be consistently wrong
- Semi-automatically, e.g., automatic analysis + manual post-editing
  - Pro: Can combine the best of both worlds
  - Con: Have to avoid the pitfalls of each

We’ll browse some annotation tools soon

Levels of linguistic annotation

- Morphological (e.g. inflection, derivation, compounding)
- Morpho-syntactic: part-of-speech (POS) tagging
- Syntactic (e.g. named entities, phrasal chunking, full syntactic analysis)
- Semantic (e.g. word-sense disambiguation, anaphora & coreference resolution, information structure)
- Discourse (e.g. dialog turns, speech acts)

Some annotation uses

- POS annotation: compare occurrences of word classes
- Lemmatization: study distribution of lemmes for lexicography
- Parsing: study clauses types in a language or teach grammatical analysis
- Semantic annotation: analyze content

Other types of annotation can cover specific purposes: stylistic annotation, error tagging, problem-oriented annotation, ...

We’ll examine POS and syntactic annotation in greater detail later in the semester
Preprocessing the Text: Tokenization

Tokenization refers to the annotation step of dividing the input text into units called tokens.

Each token consists of one of the following:
- a morpho-syntactic word
- a punctuation mark or a special character (e.g., &, @, %)
- a number

Lemmatization

- Process of relating individual word forms to their citation form (lemma) by means of morphological analysis
  - e.g., stopped ⇒ stop
- Distinguishes between the total number of word tokens and distinct lemmata that occur in a corpus
  - e.g., helps to find all occurrences of buy
- Indispensable for highly inflectional languages with a large number of distinct word forms for a given lemma

Example Tools for Lemmatization

- XEROX Morphological Analyzer: comprehensive morphological analyzers for many languages including English, French, German, Hungarian, Italian, Czech, Polish, Russian, Turkish.

Search online for lemmatizers to find more, e.g.,
- http://www.nltk.org

Tokenization

Example

before tokenization:
Milton wrote "Paradise Lost." Then his wife dies and he wrote "Paradise Regained."

after tokenization:
Milton wrote "Paradise Lost. " Then his wife dies and he wrote "Paradise Regained . "
(taken from a "student blooper" list, http://www.snopes.com/humor/lists/student.asp)

German Example

wie wie +Adv+Wh+#lex+COWIE
wie wie +Conj+Coord+#lex+COWIE
sie sie +Pron+Pers+3P+Sg+Nom+#lex+PERSPRO
sie sie +Conj+Subord+#lex+COWIE
offenbar offenbaren +Verb+Imp+2P+Sg+Nom+PERSPRO
offenbar offenbar +Adj+Pos+Pred+#lex+ADJD
gedacht gedanken +Verb+PPast+#lex+VVPP
gedacht dachen +Verb+PPast+#lex+VVPP
gedacht denken +Verb+PPast+#lex+VVPP
hat haben +Verb+Indc+3P+Sg+Pres+#lex+VAFIN

Lingsoft:

"half" <Quant> DET PRE SG/PL @QNR>
"half" <NonMod> <Quant> PRON SG/PL
"half" N NOM SG
"half" ADV

Morphological Analysis

Xerox:
half half+Adj
half half+Adv
half half+Noun+Sg
Standoff annotation

Instead of embedding annotation in the base document, one can use standoff, or standoff, annotation:

- The annotation is linked to particular points in the original document
- Allows base documents to be treated as read-only & can be distributed separately
- Allows for overlapping & alternative annotations
- Allows for new annotation levels to be easily added

Disadvantages:

- Potentially difficult to allow for annotation to point to other annotation
- Less pre-built tools for standoff annotation

Attributes vs. Elements (1)

There is often a question of where to store information: in element text, in attributes, as children elements, ... Here are three different ways to say the "same" thing

```xml
<note date="12/11/2002">
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don’t forget me this weekend!</body>
</note>
```

http://www.w3schools.com/xml/

Attributes vs. Elements (3)

If you want to do something with the day, month, or year, the last example is the most effective (i.e., most structured)

```xml
<note date="12/11/2002">
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don’t forget me this weekend!</body>
</note>
```

Attributes vs. Elements (2)

```xml
<note>
  <date>12/11/2002</date>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don’t forget me this weekend!</body>
</note>
```
By adding XML annotation, we are able to preserve the

Some possibilities for linguistic data:

<terminal tag=""NN">dog</terminal>

<terminal word=""dog""> tag=""NN"></terminal>

<!-- or: <terminal word=""dog""> tag=""NN"></terminal>

<terminal>
<word>dog</word>
<tag>NN</tag>
</terminal>

Part of Gigaword DTD

Adding POS information: BNC

New York Times data from the English Gigaword corpus:

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE DOC SYSTEM "Gigaword.dtd" [ ]>
<DOC id="NYT19940701.0001" type="story" >
</DOC>

<DOCUMENT word=""She"" tag=""PRP"">frequented</DOCUMENT> the restaurant quite often,'' DeBello said.

(WITNESS SAYS O.J. SIMPSON BOUGHT KNIFE WEEKS BEFORE SLAYINGS)

LOS ANGELES (BC-SIMPSON-KILLINGS-1stLd-3Takes-Writethru-LADN)

<DATELINE>LOS ANGELES</DATELINE>
<HEADLINE>WITNESS SAYS O.J. SIMPSON BOUGHT KNIFE WEEKS BEFORE SLAYINGS</HEADLINE>

(STORY CAN END HERE. OPTIONAL 2ND TAKE FOLLOWS.)

</P>
</TEXT>
</DOC>

Attributes

Putting data in attributes needs to be done with care:

- Attributes cannot contain multiple values (child elements can)
- Attributes are not easily expandable (for future changes)
- Attributes cannot describe structures (child elements can)
- Attributes are more difficult to manipulate by programs

But: converting all attributes to elements can result in an inflated structure

XML for linguistic purposes

- XML allows us to add structured information to a text
  - By adding XML annotation, we are able to preserve the original document—i.e., we only add to it
- The complexity of the XML used depends on exactly what the task is.
- Using XML and a corresponding XSD or DTD allows corpus or software designers to specify how the annotation information should be formatted

Linguistic example

Putting data in attributes needs to be done with care:

- Attributes cannot contain multiple values (child elements can)
- Attributes are not easily expandable (for future changes)
- Attributes cannot describe structures (child elements can)
- Attributes are more difficult to manipulate by programs

But: converting all attributes to elements can result in an inflated structure

XML for "raw" text

New York Times data from the English Gigaword corpus:
XML for syntactic trees: TigerXML (WSJ)

<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<corpus xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="/home/compling/TIGERSearch/1.1rc2/schema/TigerXML.xsd" id="temp"
>
  <head external="file:temp_generated_header.xml"/>
  <body>
    <s id="s1">
      <graph root="s1_500">
        <terminals>
          <t id="s1_1" word="Pierre" pos="NNP" />
          <t id="s1_2" word="Vinken" pos="NNP" />
          <t id="s1_3" word="," pos="," />
          <t id="s1_4" word="61" pos="CD" />
          <t id="s1_5" word="years" pos="NNS" />
          <t id="s1_6" word="old" pos="JJ" />
          <t id="s1_7" word="," pos="," />
          <t id="s1_8" word="will" pos="MD" />
          <t id="s1_9" word="join" pos="VB" />
          <t id="s1_10" word="the" pos="DT" />
          <t id="s1_11" word="board" pos="NN" />
          <t id="s1_12" word="as" pos="IN" />
          <t id="s1_13" word="a" pos="DT" />
          <t id="s1_14" word="nonexecutive" pos="JJ" />
          <t id="s1_15" word="director" pos="NN" />
          <t id="s1_16" word="Nov." pos="NNP" />
          <t id="s1_17" word="29" pos="CD" />
          <t id="s1_18" word="." pos="." />
        </terminals>
        <nonterminals>
          <nt id="s1_502" cat="NP" >
            <edge idref="s1_1" label="--" />
            <edge idref="s1_2" label="--" />
          </nt>
          <nt id="s1_504" cat="NP" >
            <edge idref="s1_4" label="--" />
            <edge idref="s1_5" label="--" />
          </nt>
          ...
          <nt id="s1_500" cat="S" >
            <edge idref="s1_501" label="SBJ" />
            <edge idref="s1_505" label="--" />
            <edge idref="s1_18" label="--" />
          </nt>
        </nonterminals>
      </graph>
    </s>
  </body>
</corpus>