Corpora and Linguistic Annotation

L645

Dept. of Linguistics, Indiana University
Fall 2015
The goal today is to get us all on the same page about where linguistic data comes from

▶ ... since we’ll be assuming, and at times using, such data & annotation all semester

Outline

▶ Corpus design
▶ Preprocessing
▶ Annotation & annotation schemes
  ▶ (& linguistic distinctions therein)
▶ Mark-up encoding

If you enjoy today’s material, consider taking L615, *Corpus Linguistics*
Corpus design

Representativeness

**Representativeness:** the extent to which a sample includes the full range of variability in a population

- Distinguishes corpora from archives
- Allows findings to be generalized to a particular variety of language

Corpus: sample of language use (i.e., from a particular population)

- Balance: types of genres
- Sampling: how the text is selected
Corpus design

Balance

What should be covered in a balanced corpus?

- **Balanced**: covers a range of text categories
  - Definition depends upon the intended uses
  - No real objective measure of balance
  - Usually based on proportional sampling
Corpus design

Sampling

To achieve representativeness & balance, one samples language (since it cannot be exhaustively described)

▶ The sample should be representative of the larger population

To properly sample, need to define:

▶ **Sampling unit**: book, periodical, newspaper, articles, chapters, . . .

▶ **Sampling frame**: list of all possible units, from which that actual ones are selected
  ▶ Brown sampling frame: list of books & periodicals in Brown University Library & Providence Athenaeum
  ▶ (written English text published in 1961)
Corpus design

Corpus size

Generally speaking, the bigger the corpus the better, but ...

- For annotation, text cannot be as big
  - e.g., not possible to have 100 billion words with high-quality linguistic annotation without many errors
Preprocessing

Some issues arise as the corpus is being built:

▶ Segmentation: what counts as a sentence for this corpus?
▶ Tokenization: what counts as a word for this corpus?
▶ Lemmatization: do we include word stems for each word in the corpus?
Intuitively, anything ending in a ., ?, or ! is a sentence.
- But :; and — often also end sentences
- And periods can also be abbreviation markers

Quotation marks also vary in their consistency of use

Spoken language data has the further complication of separating utterances from sentences

Usually, a heuristic-based algorithm attempts to find sentence boundaries (possibly in tandem with a tokenizer)
Heuristic sentence boundary detection
Manning & Schütze, figure 4.1

- Place putative sentence boundaries after all occurrences of . ? ! (and maybe ; : —)
- Move boundary after following quotation marks, if any
- Disqualify a period boundary if:
  - it is preceded by a known abbreviation and not followed by an uppercase word (e.g., etc.)
  - it is preceded by a known abbreviation which is commonly followed by a capitalized word (e.g., Prof.)
- Disqualify a boundary with a ? or ! if:
  - it is followed by a lowercase letter (or a known name)
- Regard other putative sentence boundaries as sentence boundaries.
Preprocessing

Tokenization

- Intuitively, a word is anything between whitespace
- But punctuation can cause problems because they mean multiple things
  - Periods: sentence enders or abbreviations (or both, e.g., etc.)?
  - Apostrophes: contractions treated as one word (phonetically, I’m) or two (two syntactic classes, I ’m)?
  - Hyphens: can cause two otherwise similar words to be distinct (co-operate vs. cooperate)
- Whitespace not indicating a break: data base, the New York-New Haven railroad
- Speech: fillers (uh, um), meta-information ([cough])

Word segmentation in other languages is even more complicated
Corpora often also contain the **lemma** for each word, the base form or stem of a word

- Need to be able to do this accurately: e.g., what is the lemma of *leaves*? *leaf* or *leave*?

**Stemming** is a similar process popular in Information Retrieval (IR)

- Instead of finding the dictionary form, simply strip off affixes
Annotation and annotation schemes

Building from the text, many corpora add various linguistic annotations

- Choose a corpus which has the annotation of interest
  - Or find a tool which can automatically add the annotation with high precision
- Remember that annotation is almost always an opinion (even spelling “errors”), so:
  - You may have to consult the guidelines which come with a corpus
Grammatical tagging

One of the most common forms of annotation is that of part-of-speech classes, or grammatical classes

▪ Part-of-speech (POS) tags are a combination of morphological and syntactic information
  ▪ Occasionally require semantic distinctions to be made
▪ Tagsets differ in size and granularity, and there is not always a way to map between tagsets
Brown corpus example

[ The/DT Fulton/NNP County/NNP Grand/NNP Jury/NNP ] said/VBD
[ Friday/NNP ]

[ an/DT investigation/NN ]
of/IN
[ Atlanta/NNP 's/POS recent/JJ primary/JJ election/NN ]
produced/VBD ‘‘/‘‘
[ no/DT evidence/NN ]
’’/’’ that/IN
[ any/DT irregularities/NNS ]
took/VBD
[ place/NN ]
./.

====================================================================
## Comparison of different tagsets

Manning & Schütze, figure 4.2

<table>
<thead>
<tr>
<th>Sentence</th>
<th>CLAWS c5</th>
<th>Brown</th>
<th>PTB</th>
<th>ICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>PNP</td>
<td>PPS</td>
<td>PRP</td>
<td>PRON(pers,sing)</td>
</tr>
<tr>
<td>was</td>
<td>VBD</td>
<td>BEDZ</td>
<td>VBD</td>
<td>AUX(pass,part)</td>
</tr>
<tr>
<td>told</td>
<td>VVN</td>
<td>VBN</td>
<td>VBN</td>
<td>V(ditr,edp)</td>
</tr>
<tr>
<td>that</td>
<td>CJT</td>
<td>CS</td>
<td>IN</td>
<td>CONJUNC(subord)</td>
</tr>
<tr>
<td>the</td>
<td>AT0</td>
<td>AT</td>
<td>DT</td>
<td>ART(def)</td>
</tr>
<tr>
<td>journey</td>
<td>NN1</td>
<td>NN</td>
<td>NN</td>
<td>N(com,sing)</td>
</tr>
<tr>
<td>might</td>
<td>VM0</td>
<td>MD</td>
<td>MD</td>
<td>AUX(modal,past)</td>
</tr>
<tr>
<td>kill</td>
<td>VVI</td>
<td>VB</td>
<td>VB</td>
<td>V(montr,infin)</td>
</tr>
<tr>
<td>her</td>
<td>PNP</td>
<td>PPO</td>
<td>PRP</td>
<td>PRON(poss,sing)</td>
</tr>
<tr>
<td>.</td>
<td>PUN</td>
<td>.</td>
<td>.</td>
<td>PUNC(per)</td>
</tr>
</tbody>
</table>
A common layer above POS is that of syntactic (phrase structure) annotation

- Corpus with syntactic annotation is often called a treebank

- Again, you may have to consult the guidelines for the annotation scheme

- Treebanks generally shoot for “theory-neutrality”

- Dependency & function labeling are also quite common
Phrase structure tree

S
  / \  
NP  VP
  /  /  
D  N  V  PP
     /  
    N  NP
       /  
      P  N
        /  
       to  heaven

All  dogs  go  to  heaven
Dependency tree

All dogs go to heaven
Corpora and Linguistic Annotation

Corpus design

Preprocessing

Annotation

Mark-up

Syntactic annotation example

SUSANNE

A01:0010.03 - YB <minbrk> - [Oh.Oh]
A01:0010.06 - AT The the [O[S[Nns:s.
A01:0010.09 - NPLs Fulton Fulton [Nns.
A01:0010.12 - NNL1cb County county .Nns]
A01:0010.15 - JJ Grand grand .
A01:0010.18 - NN1c Jury jury .Nns:s]
A01:0010.21 - VVDv said say [Vd.Vd]
A01:0010.24 - NPD1 Friday Friday [Nns:t.Nns:t]
A01:0010.27 - AT1 an an [Fn:o[Ns:s.
A01:0010.30 - NN1n investigation investigation .
A01:0020.03 - IO of of [Po.
A01:0020.06 - NPlt Atlanta Atlanta [Ns[G[Nns.Nns]
A01:0020.09 - GG +apos>s - .G]
A01:0020.12 - JJ recent recent .
A01:0020.15 - JJ primary primary .
A01:0020.18 - NN1n election election .Ns]Po]Ns:s]
A01:0020.21 - VVDv produced produce [Vd.Vd]
A01:0020.24 - YIL <ldquo> - .
A01:0020.27 - ATn +no no [Ns:o.
A01:0020.30 - NN1u evidence evidence .
A01:0020.33 - YIR +<rdquo> - .
A01:0020.39 - CST that that [Fn.
A01:0030.03 - D Dy any any [Np:s.
A01:0030.06 - NN2 irregularities irregularity .Np:s]
A01:0030.09 - VVDv took take [Vd.Vd]
A01:0030.15 - YF +. - .0]
Syntactic annotation example
Brown part of PTB

```
(S (NP-SBJ-1 (NP Scotty 's)
    father)
   (VP sat
      (S-ADV (NP-SBJ *-1)
        (VP sprawled
           (PP-LOC in
              (NP his chair))))
   ,
   (S-ADV (NP-SBJ *-1)
      (ADJP-PRD (ADJP angular)
       (ADJP alert
        (PP as
         (NP a cricket))))
   ,
   (S-ADV (NP-SBJ *-1)
      (VP looking
       (ADVP about)
       (PP at
         (NP (NP the huge stainless-steel appointments)
          (PP of
           (NP the room)))
         (PP with
          (NP (NP an expression)
           (PP of
            (NP proprietorship)))))))
```

20 / 31
Syntactic annotation example
PTB converted to dependencies

<table>
<thead>
<tr>
<th></th>
<th>Syntactic annotation example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pierre</td>
</tr>
<tr>
<td>2</td>
<td>Vinken</td>
</tr>
<tr>
<td>3</td>
<td>,</td>
</tr>
<tr>
<td>4</td>
<td>61</td>
</tr>
<tr>
<td>5</td>
<td>years</td>
</tr>
<tr>
<td>6</td>
<td>old</td>
</tr>
<tr>
<td>7</td>
<td>,</td>
</tr>
<tr>
<td>8</td>
<td>will</td>
</tr>
<tr>
<td>9</td>
<td>join</td>
</tr>
<tr>
<td>10</td>
<td>the</td>
</tr>
<tr>
<td>11</td>
<td>board</td>
</tr>
<tr>
<td>12</td>
<td>as</td>
</tr>
<tr>
<td>13</td>
<td>a</td>
</tr>
<tr>
<td>14</td>
<td>nonexecutive</td>
</tr>
<tr>
<td></td>
<td>director</td>
</tr>
<tr>
<td></td>
<td>Nov.</td>
</tr>
<tr>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>18</td>
<td>.</td>
</tr>
</tbody>
</table>
Semantic annotation

- There are different kinds of semantic/pragmatic annotation one could in principle have
  - Word sense annotation, anaphoric relations, etc.
- One popular form of semantic annotation is to label sentences with verbal semantic relations and the arguments of that relation
PropBank example (standoff)

wsj/00/wsj_0001.mrg 0 8 gold join.01 vf--a 0:2-ARG0 7:0-ARGM-MOD 8:0-rel 9:1-ARG1
wsj/00/wsj_0001.mrg 1 10 gold publish.01 p---a 10:0-rel 11:0-ARG0
wsj/00/wsj_0002.mrg 0 16 gold name.01 pp--p 16:0-rel 0:2*17:0-ARG1 18:2-ARG2
wsj/00/wsj_0003.mrg 0 5 gold use.01 p---p 4:1-ARGM-TMP 5:0-rel 0:2*6:0-ARG1
   7:2-ARG2-PNC
wsj/00/wsj_0003.mrg 0 9 gold make.01 i---a 7:0-ARG0 9:0-rel 10:1-ARG1
wsj/00/wsj_0003.mrg 0 14 gold cause.01 pnp3a 0:3-ARG0 14:0-rel 15:2-ARG1
wsj/00/wsj_0003.mrg 0 26 gold expose.01 p---p 26:0-rel 28:1-ARG2-to 30:3-ARGM-TMP
   22:1,24:0,25:1*27:0-ARG1
wsj/00/wsj_0003.mrg 0 37 gold report.01 vp--a 36:1-ARG0 37:0-rel 0:4*39:0-ARG1
wsj/00/wsj_0003.mrg 1 11 gold enter.01 vn-3a 10:1-ARG0 11:0-rel 12:1-ARG1
wsj/00/wsj_0003.mrg 1 21 gold cause.01 p---a 16:2-ARG0 21:0-rel 22:2-ARG1
File it refers to

(S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken) )
    (, ,)
    (ADJP
      (NP (CD 61) (NNS years) )
      (JJ old) )
    (, ,) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
  (.. ..))
(S
  (NP-SBJ (NNP Mr.) (NNP Vinken) )
  (VP (VBZ is)
    (NP-PRD
      (NP (NN chairman) )
      (PP (IN of)
        (NP
          (NP (NNP Elsevier) (NNP N.V.) )
          (, ,)
          (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
  (.. ..))
Corpus mark-up schemes

To add linguistic properties to a corpus requires mark-up

- XML (eXtensible Markup Language)
- In contrast to HTML, XML does not have built-in “meaning” for labels: you must define your own tags

Benefits of XML:

- Pretty standard internationally and across fields
- Good tools available & convenient for a lot of programs
- Widely used in real text processing
- Handles multilinguiality very well
XML example

```xml
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<corpus xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <head external="file:wsj_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" />
          ...
        </terminals>
      </graph>
    </s>
  ...
```
XML tags

Several things to note in the previous example:

- Needs to be one root tag, e.g., `<corpus>` in this case
- Every (opening) tag needs a closing tag:
- Along with that, tags must be properly nested:
  - `<b><i>word</i></b>` is legitimate;
  - `<b><i>word</i></b>` is not

So, structure ends up like:

```xml
<root>
  <child>
    <subchild>.....</subchild>
  </child>
</root>
```
## Attributes

Each tag (or element) can have a variety of attributes (provided such attributes have been declared; see below)

- Attributes are noted within an element tag; there can be multiple attributes & can be put in different orders
  
  ```xml
  <t id="s1_1" word="Pierre" pos="NNP" />  
  <t word="Pierre" pos="NNP" id="s1_1" />  
  ```

- Attribute values are put in quotes after the attribute
- `<t XYZ />` is the same as `<t XYZ></t>`
Child & Attribute possibilities

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

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

<terminal>
  <word>dog</word>
  <tag>NN</tag>
</terminal>
New York Times data from the English Gigaword corpus:

```xml
<DOC id="NYT19940701.0001" type="story">
  <HEADLINE>
  WITNESS SAYS O.J. SIMPSON BOUGHT KNIFE WEEKS BEFORE SLAYINGS
  </HEADLINE>
  <DATELINE>
  LOS ANGELES (BC-SIMPSON-KILLINGS-1stLd-3Takes-Writethru-LADN)
  </DATELINE>
  <TEXT>
    <P>
    With the nation's attention riveted again on a Los Angeles courtroom, a knife dealer testified that O.J. Simpson bought a 15-inch knife five weeks before the slashing deaths of his ex-wife and her friend.
    </P>
    ...
    <P>
    "She frequented the restaurant quite often," DeBello said.
    </P>
    <P>
    (STORY CAN END HERE. OPTIONAL 2ND TAKE FOLLOWS.)
    </P>
  </TEXT>
</DOC>
```
NLP Perspective

Follow-up material we’ll zip through:
http://cl.indiana.edu/~md7/14/715/slides/03-nlp/03-nlp.pdf

And some material to specifically help us with the Stanford CoreNLP tools: http://cl.indiana.edu/~md7/14/715/slides/03-nlp/03b-practical.pdf