Linguistic annotation

Mark up, XML, DTD, TEI

Serge Sharoff

Centre for Translation Studies
University of Leeds

s.sharoff@leeds.ac.uk

Outline

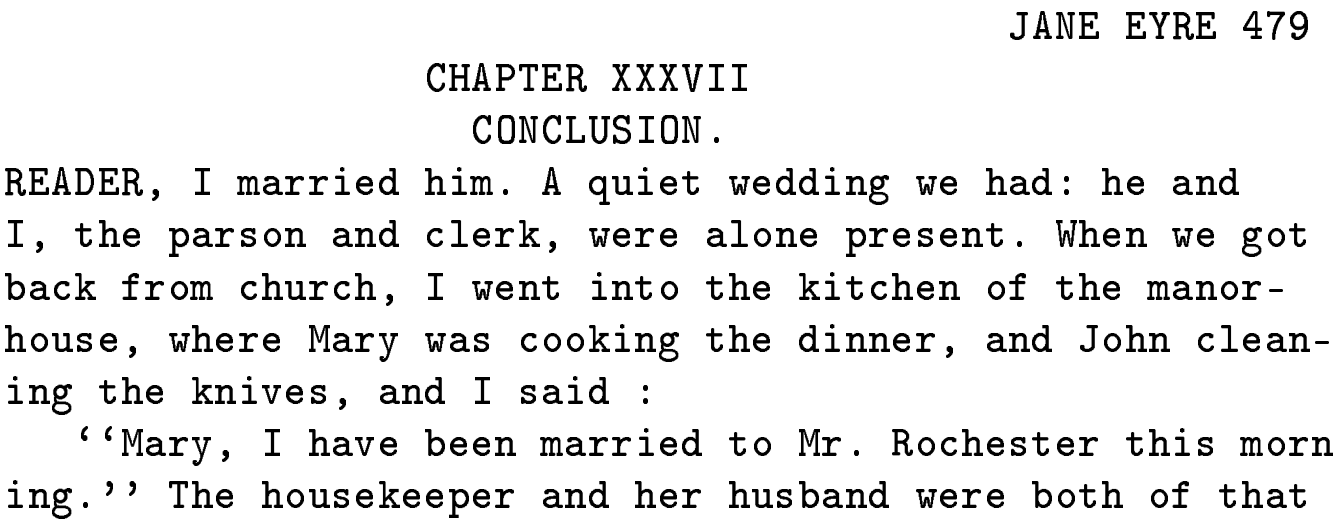
CHAPTER XXXVIII.

CONCLUSION.

READER, I married him. A quiet wedding we had; he and I, the parson and clerk, were alone present. When we got back from church, I went into the kitchen of the manor-house, where Mary was cooking the dinner, and John cleaning the knives, and I said:

"Mary, I have been married to Mr. Rochester this morning." The housekeeper and her husband were both of that decent, phlegmatic order of people, to whom one may at any

Why annotation?



JANE EYRE 479

CHAPTER XXXVII
CONCLUSION.

READER, I married him. A quiet wedding we had: he and I, the parson and clerk, were alone present. When we got back from church, I went into the kitchen of the manor-house, where Mary was cooking the dinner, and John cleaning the knives, and I said :

‘‘Mary, I have been married to Mr. Rochester this morning.’’ The housekeeper and her husband were both of that

Any problems with OCR?

What is annotation?

- Data vs. interpretation
- Markup contributes towards explicit interpretation of a text
- Early examples of markup: alphabet, punctuation, typographic setup
- Different annotations for different interpretations:
  Chomsky set out his theory in *Syntactic Structures* (1957)

→ `Chomsky set out his theory in <i>Syntactic Structures</i> (1957)`

`Chomsky set out his theory in <cite type="title" ref="Chomsky1957"/>`

XML: example of code

```
<pb n="479"/>
<div1 type="chapter"   n="38">
<h>Conclusion .</h>
<p><s>Reader , I married him .</s>
<s>A quiet wedding we had : he and I , the parson and clerk ,
were alone present . </s>
<s>When we got back from church , I went into the kitchen of
the manor-house , where Mary was cooking the dinner , and John
cleaning the knives, and I said : </s></p>
<p><q><s> Mary, I have been married to Mr. Rochester this morning . </s> </q>
...
</p>
```

Markup languages

- XML: eXtensible Markup Language (1999)
  Representing the logical structure of data
- XSL(T)= style sheet and transformation language
- HTML: HyperText Markup Language
  a relative of XML, xHTML
- TMX: Translation Memory eXchange format
- TBX: TermBase eXchange format
- SRX: Segmentation Rules eXchange format
- XML for dictionary entries

XML basics

- tags, words in angular brackets: `<p>` `<div1>`
- elements, the content within a tag
- attributes and values: `<div1 type="chapter" n="38">`
- empty elements: `<pb n="474"/>`
- entities: `&mdash;` (—), `&amp;` (&), `&auml;` (ä)
- DTD (Document Type Definition) — definition of all the elements, attributes and entities

Well-formed XML document

- every tag which is opened has to be closed
* `<p> text <p> new paragraph`
- The only exception is for empty elements:
  `<pb n="474"/>` $\mapsto$ `<pb n="474"></pb>`
- tags are properly embedded:
* `<p> <q> quote </p></q>`
- Tags are case-sensitive:
  `<publicationStmt>` `<gramGrp>`

Outline

Formats of annotation

- Tab-separated format

```
N12:0510    VVD    studied    study    [Coord.N12:0504]
```

- COCOA: `<w VVD>studied`
- XML:

```
<s n="12">
        ...
<w id="N12:0510">studied
<ana lemma="study" pos="VVD" />
</w>
</s>
```

Types of annotation

- Metatext annotation (author, audience, domain, genre)
- Text annotation (quotes, comments, page breaks)
- Typographic annotation (fonts, headings, text alignment)
- Linguistic annotation:

  Part-of-speech (POS): `<w pos="VVZ">studies</w>`

  Lemmatisation: `He <w lemma="leave">left</w>`

  Functional relationships, e.g., coreference:
  `<w ref="X1" lemma="he">He</w>`

  Word senses: power$_1$ →énergie, power$_2$ →pouvoir

The owner	of	the company	has	the power	to fire
Le propriétaire	de	la société	a	le pouvoir	de licencier

Other kinds of annotation

- Named Entities
  Mark boundaries of names of type PERson, ORGanization, GPE, LOCation,...
  <enamex type="ORGANIZATION">FBI</enamex> agents arrested <enamex type="PERSON">Kaczynski</enamex> on <enamex type="DATE">April 3, 1996</enamex>, at his remote cabin outside <enamex type="LOCATION">Lincoln</enamex>
- Gazetteers, normalisers, disambiguation
  Lech Kaczynski vs *Jaroslaw Kaczynski* vs *Ted Kaczynski*
  Lincoln, US vs *Lincoln, UK*
- Terminology annotation: Identify terms, detect their canonical form and link to their database records
  <te id="1123">Fast Breeder Reactors</te> (<te id="1123" type="abbr">FBR</te>) can produce more <te id="1134">fissile fuel</te> than <te id="1123" type="ana">they</te> consume.

Standards of annotation

- TEI (Text Encoding Initiative), TEI-Lite (a subset) — for text and metatext annotation
- EAGLES (European Advisory Group on Language Engineering Standards) — for linguistic annotation
- *The man still saw her*

  The (AT) man ($NN|VV$) still ($NN|VV|RB$) saw ($NN|VVD$) her ($PP|PP\$$) . ($SENT|PUN$)
- Rules vs statistical training
- Statistical hidden Markov model: $p(t_i|w_i) \times p(t_i|t_{i-1})$

$$p(PP|her) = 0.3; \quad p(PP|SENT) = 0.0022; \quad =0.000654$$
$$p(PP\$|her) = 0.7; \quad p(PP\$|SENT) = 0.000019 \quad =0.000013$$

POS tags

- The standard inventory for English (the Penn set):
  VV – base verb, VVD – past tense, VVN – pp (taken)
  NN – common noun, singular; NNS – common noun, plural;
  CS – conjunction, subordinative, *so that*: so_CS21
  that_CS22

  Penn set: 55 tags; Lancaster set: 146 tags
- Specific sets for languages and taggers:
  German STTS: VVFIN – finite verb, VVINF – infinitive,
  VVIZU – infinitive with "zu",
  NN – common noun (50 tags)

  Japanese mecab tagset: 16 tags

  Russian tagset: 1066 tags

  Ncmsan : Noun, Type = common, Gender = masculine,
  Number = singular, Case = accusative, Animate = no

Accuracy of tagging

- Accuracy – the percentage of words (i.e. word tokens) which are correctly tagged

  95% accuracy: one in 20 is wrong

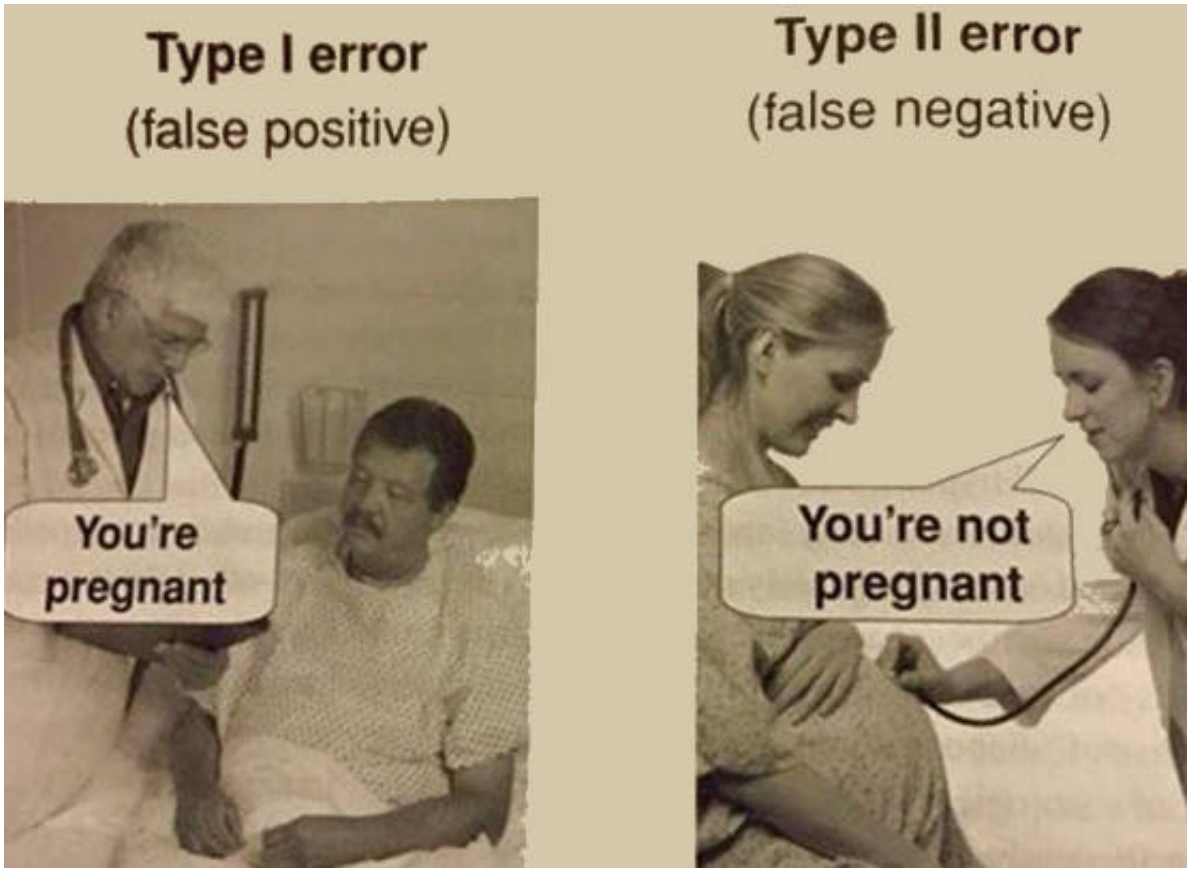
  96% accuracy: one in 25 is wrong

  an improvement of 25% from 95% to 96% ???
- Domain and genre influence:
  newspapers for training (Wall Street Journal in Penn Treebank)

Types of errors

- *I can light a fire. You can open a can.*

→ PP VV NN DT NN. PP VV VV DT VV.



- Precision = TP/(TP+ FP) vs
  Recall = TP/(TP + FN)

Text level annotation

- Linguistic features which can be counted

  Lexical features:
  publicVerbs = *acknowledge*, *admit*, *agree*, *assert*, *claim*...
  amplifiers = *absolutely*, *altogether*, *completely*, *enormously*, ...

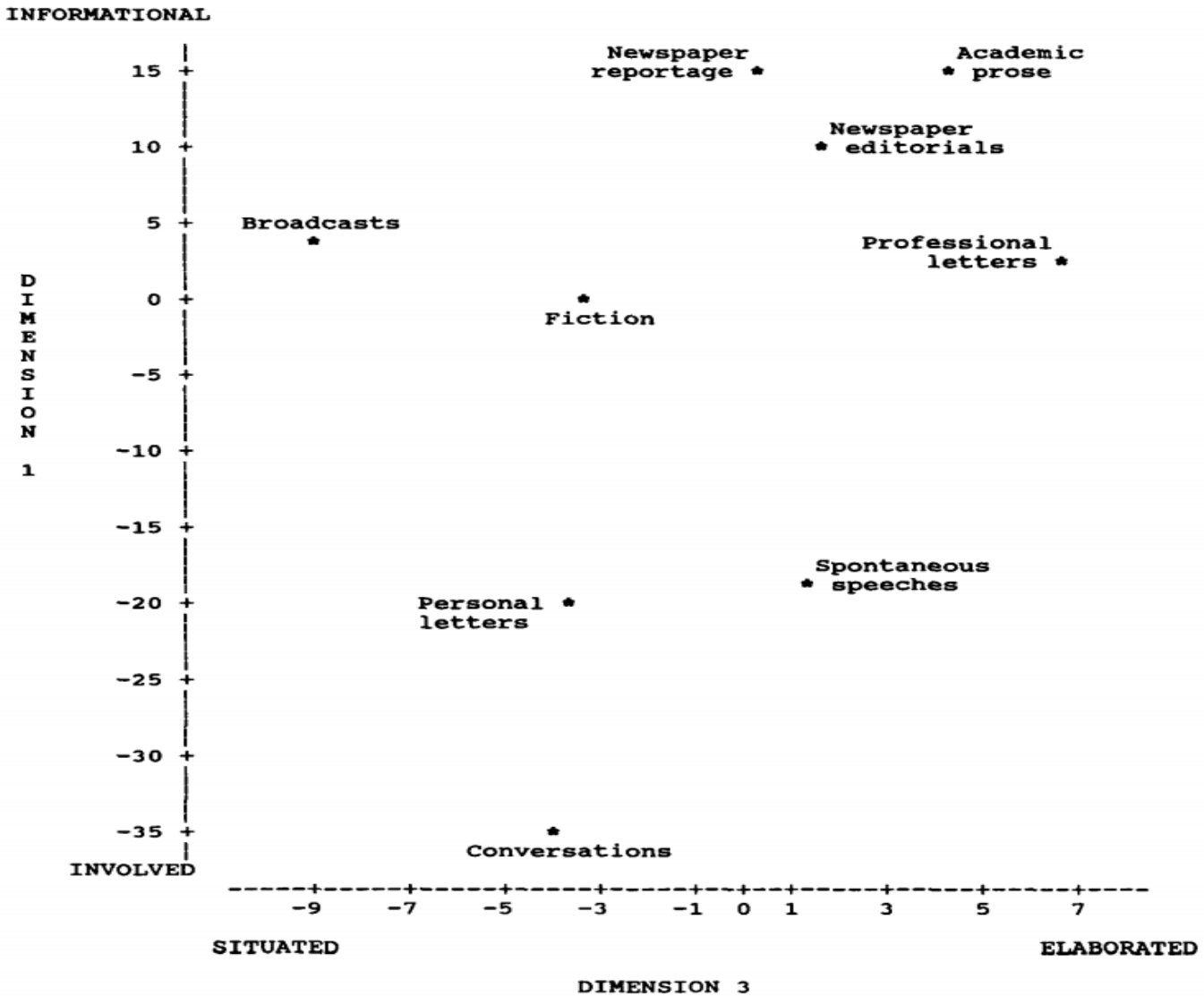
  Part-of-speech features:
  Nominalisations (nouns ending in −*tion*, −*ness*, −*ment*)
  main POS tags
  Past tense verbs (VVD)

  Syntactic features:
  that deletions
  pied piping (*Which house did she buy ...?*)

  Text-level features, such as:
  Average word length
  Average sentence length

- Factor analysis: which features group with each other

Multi-dimensional analysis (Biber, 1988)



Multi-dimensional analysis (Biber, 1988)

Functions	Linguistic features	Characteristic genres
Dimension 1		
Monologue Careful production Faceless	nouns, adjectives prepositional phrases long words	informational exposition e.g., official documents academic prose
Interactive Personal focus Involved Online production	1^{st} and 2^{nd} PPs questions, reductions stance verbs, hedges emphatics	conversations (public and private)
Dimension 3		
Elaborated	wh-relative clauses pied-piping phrasal coordination	official documents professional letters (exposition)
Situation-dependent	time and place adverbials	broadcasts (fiction)

Basic points

- Raw texts vs. annotated texts
- Annotation formats and standards: XML, TEI, EAGLES
- Metatextual annotation
- POS tagging
- Higher-level annotation

For the next seminar

Prepare for a short presentation of the most exciting findings in your projects to discuss them with another student in the group