Overview

➢ Revision of Introduction
  ➢ What is Corpus Linguistics

➢ Mark-up

➢ Annotation

➢ Regular Expressions

➢ Lab One
Revision
What is a Corpus?

*corpus* (pl: *corpora*):

> In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database.

- machine-readable (i.e., computer-based)
- authentic (i.e., naturally occurring)
- sampled (bits of text taken from multiple sources)
- representative of a particular language or language variety.

> Sinclair’s (1996) definition:

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.
Why Are Electronic Corpora Useful?

➢ as a collection of examples for linguists
  ➢ intuition is unreliable

➢ as a data resource for lexicographers
  ➢ use natural data to exemplify usage

➢ as instruction material for language teachers and learners

➢ as training material for natural language processing applications
  ➢ training of speech recognizers, parsers, MT
The British National Corpus (BNC)

- 100 million words of written and spoken British English

- Designed to represent a wide cross-section of British English from late 20th century: balanced and representative

- POS tagging (2 million word sampler hand checked)

<table>
<thead>
<tr>
<th>Written (90%)</th>
<th>Domain</th>
<th>Date</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Imaginative (22%)</td>
<td>1960-74 (2%)</td>
<td>Book (59%)</td>
</tr>
<tr>
<td></td>
<td>Arts (8%)</td>
<td>1975-93 (89%)</td>
<td>Periodical (31%)</td>
</tr>
<tr>
<td></td>
<td>Social science (15%)</td>
<td>Unclassified (8%)</td>
<td>Misc. published (4%)</td>
</tr>
<tr>
<td></td>
<td>Natural science (4%) . . .</td>
<td></td>
<td>Misc. un-pub (4%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spoken (10%)</th>
<th>Region</th>
<th>Interaction type</th>
<th>Context-governed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>South (46%)</td>
<td>Monologue (19%)</td>
<td>Informative (21%)</td>
</tr>
<tr>
<td></td>
<td>Midlands (23%)</td>
<td>Dialogue (75%)</td>
<td>Business (21%)</td>
</tr>
<tr>
<td></td>
<td>North (25%) . . .</td>
<td>Unclassified (6%)</td>
<td>Institutional (22%)</td>
</tr>
</tbody>
</table>

Markup and Annotation
General vs. specialized corpora

➤ General corpora (such as “national” corpora) are a huge undertaking. These are built on an institutional scale over the course of many years.

➤ Specialized corpora (ex: corpus of English essays written by Japanese university students, medical dialogue corpus) can be built relatively quickly for the purpose at hand, and therefore are more common.

➤ Characteristics of corpora:
   1. Machine-readable, authentic
   2. Sampled to be balanced and representative

➤ Trend: for specialized corpora, criteria in (2) are often weakened in favor of quick assembly and large size.
   Rare phenomena only show up in large collections.
Mark Up
Mark-Up vs. Corpus Annotation

➤ **Mark up** provides objectively verifiable information
  ➢ Authorship
  ➢ Publication dates
  ➢ Paragraph boundaries
  ➢ Source text (URL, Book, …)

➤ **Annotation** provides interpretive linguistic information
  ➢ Sentence/Utterance boundaries
  ➢ Tokenization
  ➢ Part-of-speech tags, Lemmas
  ➢ Sentence structure
  ➢ Domain, Genre

Many people use the terms interchangeably.
The Need for Corpus Mark-Up

Mark up and Annotation guidelines are needed in order to facilitate the accessibility and reusability of corpus resources.

- Minimal information:
  - authorship of the source document
  - authorship of the annotated document
  - language of the document
  - character set and character encoding used in the corpus
  - conditions of licensing
Dublin Core Ontology

➢ Goals

➢ Provides a semantic vocabulary for describing the “core” information properties of resources (electronic and “real” physical objects)
➢ Provide enough information to enable intelligent resource discovery systems

➢ History

➢ A collaborative effort started in 1995
➢ Initiated by people from computer science, librarianship, on-line information services, abstracting and indexing, imaging and geospatial data, museum and archive control.

http://dublincore.org/
Dublin Core - 15 Elements

➢ Content (7)
   ➢ Title, Subject, Description, Type, Source, Relation and Coverage

➢ Intellectual property (4)
   ➢ Creator, Publisher, Contributor, Rights

➢ Instantiation (4)
   ➢ Date, Language, Format, Identifier
Dublin Core – discussion

➤ Widely used to catalog web data

➤ OLAC: Open Language Archives Community
➤ LDC: Linguistic Data Consortium
➤ ELRA: European Language Resources Archive
➤ …
An attempt to give Language Resources (such as corpora) a unique Identifier

Like ISBNs for books

“The main purpose of the metadata schema used in ISLRN, is the identification of LRs. Inspired by the broadly known OLAC schema, a minimal set of metadata was chosen to ensure that any resource can be correctly distinguished and identified. We emphasize on the simplicity of the fields, that are easy and quick to fill in beyond misunderstanding.”

Only a distributor, creator or rights holder for a resource is entitled to submit it for ISLRN assignment.

Task: Pick a corpus, write its MetaData
# ISLRN Metadata schema

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>The name given to the resource.</td>
<td>1993-2007 United Nations Parallel Text</td>
</tr>
<tr>
<td>Full Official Name</td>
<td>The name by which the resource is referenced in bibliography.</td>
<td>1993-2007 United Nations Parallel Text</td>
</tr>
<tr>
<td>Resource Type</td>
<td>The nature or genre of the content of the resource from a linguistic standpoint.</td>
<td>Primary Text</td>
</tr>
<tr>
<td>Source/URL</td>
<td>The URL where the full metadata record of the resource is available.</td>
<td><a href="https://kitwiki.csc.fi/twiki/bin/view/">https://kitwiki.csc.fi/twiki/bin/view/</a></td>
</tr>
<tr>
<td>Format/MIME Type</td>
<td>The file format (MIME type) of the resource. Examples: text/xml, video/mpeg, etc.</td>
<td>text/xml</td>
</tr>
<tr>
<td>Size/Duration</td>
<td>The size or duration of the resource.</td>
<td>21416 KB</td>
</tr>
<tr>
<td>Access Medium</td>
<td>The material or physical carrier of the resource.</td>
<td>Distribution: 3 DVDs</td>
</tr>
<tr>
<td>Description</td>
<td>A summary of the content of the resource.</td>
<td>1.0</td>
</tr>
<tr>
<td>Version</td>
<td>The current version of the resource.</td>
<td>Text</td>
</tr>
<tr>
<td>Media Type</td>
<td>A list of types used to categorize the nature or genre of the resource content.</td>
<td></td>
</tr>
<tr>
<td>Language(s)</td>
<td>All the languages the resource is written or spoken in.</td>
<td>eng (English)</td>
</tr>
<tr>
<td>Resource Creator</td>
<td>The person or organization primarily responsible for making the resource.</td>
<td>Ah Lian; NTU; <a href="lian@ntu">lian@ntu</a>; Singapore</td>
</tr>
<tr>
<td>Distributor</td>
<td>The person or organization responsible for making the resource available.</td>
<td>Ah Beng; NUS; <a href="Beng@nus">Beng@nus</a>; Singapore</td>
</tr>
<tr>
<td>Rights Holder</td>
<td>The person or organization owning or managing rights over the resource.</td>
<td>Lee, KY; Gov; <a href="LKY@gov">LKY@gov</a>; Singapore</td>
</tr>
<tr>
<td>Relation</td>
<td>A related resource.</td>
<td></td>
</tr>
</tbody>
</table>

[http://www.islrn.org/basic_metadata/](http://www.islrn.org/basic_metadata/)
Annotation
Geoffrey 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. This is the flip side of maxim 1. Taking points 1. and 2. together, the annotated corpus should allow the maximum flexibility for manipulation by the user.

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.

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.
Types of Corpus Annotation

➤ Tokenization, Lemmatization

➤ Parts-of-speech

➤ Syntactic analysis

➤ Semantic analysis

➤ Discourse and pragmatic analysis

➤ Phonetic, phonemic, prosodic annotation

➤ Error tagging
How is Corpus Annotation Done?

➢ Three ways:

1. Manual (done entirely by human annotators)
2. Semi-automatic (done first by computer programs; post-edited)
3. Automatic (done entirely by computer programs)

➢ Labor intensive: 1 > 2 > 3

➢ Some types of annotation can be reasonably reliably produced by computer programs alone

➢ Part-of-speech tagging: accuracy of 97%
➢ Lemmatization

Computer programs for other annotation types are not yet good enough for fully automatic annotation
Lemmatization

- *unit* and *units* are two word-forms belonging to the same lemma *unit*.

- Same lemmas are shared for:
  - Plural morphology: *unit/units*: *unit*, *child/children*: *child*
  - Verbal morphology: *eat/eats/ate/eaten/eating*: *eat*
  - Comparative/superlative morphology of adjectives
    - *many/more/most*: *many*, *slow/slower/slowest*: *slow*
    - but also *much/more/most*: *much*

- Lemmatization does not affect:
  - derived words that belong to different part-of-speech groups
    - *quick/quicker/quickest*: *quick*, *quickly*: *quickly*
    - Korea: *Korea*, Korean: *Korean*
Lemmatization for English can be performed reliably and accurately using automated programs.

Why do we need lemmatization?
Part-of-Speech (POS) Tagging

➢ (POS) Tagging: adding part-of-speech information (tag) to words. *Colorless/JJ green/JJ ideas/NNS sleep/VBP furiously/RB ./.*

➢ Useful in searches that need to distinguish between different POSs of the same word (ex: *work* can be both a noun and a verb)

➢ In the US, the Penn Treebank POS set is de-facto standard:

  ➢ [http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html](http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html)
  ➢ 45 tags (including punctuation)

➢ In Europe, CLAWS tagset is popular (also used by BYU Corpora):

  ➢ [http://ucrel.lancs.ac.uk/claws7tags.html](http://ucrel.lancs.ac.uk/claws7tags.html)
  ➢ 137 tags (without punctuation)
For English, many POS taggers with good performance are available for automated corpus annotation:

- CLAWS (Lancaster University, 96-97% accuracy)
- TnT tagger/Hunpos (Saarland University, 94-97% accuracy)
- Averaged Perceptron (NLTK)/RNNs do a little better
# Penn Treebank Examples

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
<td>VB</td>
<td>Verb, base form</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td>VBD</td>
<td>Verb, past tense</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>VBG</td>
<td>Verb, gerund or present participle</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>VBN</td>
<td>Verb, past participle</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal pronoun</td>
<td>VBP</td>
<td>Verb, non-3rd person singular present</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition</td>
<td>VBZ</td>
<td>Verb, 3rd person singular present</td>
</tr>
<tr>
<td>TO</td>
<td><em>to</em></td>
<td></td>
<td>Sentence Final punct (. , ?, !)</td>
</tr>
</tbody>
</table>

- The tags include inflectional information

- If you know the tag, you can generally find the lemma

- Some tags are very specialized: I/PRP wanted/VBD to/TO go/VB ./
## Universal Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERB</td>
<td>verbs (all tenses and modes)</td>
<td>twenty-four, fourth, 1991, 14:24</td>
</tr>
<tr>
<td>NOUN</td>
<td>nouns (common and proper)</td>
<td></td>
</tr>
<tr>
<td>PRON</td>
<td>pronouns</td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>adjectives</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>adverbs</td>
<td></td>
</tr>
<tr>
<td>ADP</td>
<td>adpositions (prepositions and postpositions)</td>
<td></td>
</tr>
<tr>
<td>CONJ</td>
<td>conjunctions</td>
<td></td>
</tr>
<tr>
<td>DET</td>
<td>determiners</td>
<td></td>
</tr>
<tr>
<td>NUM</td>
<td>cardinal numbers</td>
<td></td>
</tr>
<tr>
<td>PRT</td>
<td>particles or other function words</td>
<td>at, on, out, over per, that, up, with</td>
</tr>
<tr>
<td>X</td>
<td>other: foreign words, typos, abbreviations</td>
<td>ersatz, esprit, dunno, gr8, univeristy</td>
</tr>
<tr>
<td>.</td>
<td>. , ; ! punctuation</td>
<td></td>
</tr>
</tbody>
</table>

**Task:** Find a tagset in a language you speak — map it to the universal tagset. (format tag—utag e.g. NNS—Noun).
Parsing (Syntactic Annotation)

➢ Parsing: adding phrase-structure information (parse) to sentences

(S (NP (N Claudia))
 (VP (V sat)
 (PP (P on)
 (NP (AT a)
 (N stool)))))

➢ Useful for corpus investigation of grammatical structures
➢ Parsed corpora are sometimes known as treebanks
  ➢ The Penn Treebank, Hinoki Treebank, . . .
➢ Parsed corpora are used for “training” automated parsers
➢ Stanford Parser and CMU parser were trained on the Penn Treebank corpora
Corpus vs. Annotation Software

➢ The two help each other. How?

1. An annotated corpus is built, entirely by humans
2. Then a computer program is *trained* on this corpus
3. Now new corpora can be automatically annotated using this program

➢ In practice, normally start off with a simple program and then correct its output
A corpus can be used to **train** a computer program. A program **learns** from corpus data. What does this mean?

* **work** is a noun (NN) in some contexts, and a verb (VB) in some others.
* When **work** follows an adjective (ADJ), it is likely to be a noun.
* When **work** follows a plural noun (NNS), it is likely to be a verb.
  * *nice/ADJ work/NN, beautiful/ADJ work/NN*
  * *they/NNS work/VB at a hospital*
  * *my parents/NNS work/VB too much*

These patterns can be extracted from a corpus, and the “trained” computer program makes a statistical model with them to predict the POS of **work** in a new text.
Semantic Annotation

➢ Word sense disambiguation between homonyms

➢ ex. *lie* in:

➢ *The boy lied*₁ to his parents.
➢ *Mary lied*₂ down for a nap.

➢ ex. *share* in:

➢ *Mary did not share*₁ her secret with anyone.
➢ *The share*₂ holders of Intel were disappointed.
A semantic role is the relationship that a syntactic constituent has with a predicate: e.g., Agent, Patient, Instrument, Locative, Temporal, Manner, Cause, . . .

An example from the PropBank corpus:

[A0 He ] [AM-MOD would ] [AM-NEG n’t ] [V accept ] [A1 anything of value ] from [A2 those he was writing about ] .

V: verb
A0: acceptor
A1: thing accepted
A2: accepted-from
A3: attribute
AM-MOD: modal
AM-NEG: negation
Temporal expressions tell us:

- When something happened
- How long something lasted
- How often something occurs

Examples

- He wrapped up a three-hour meeting with the Iraqi president in Baghdad today.
- The king lived 4,000 years ago.
- I’m a creature of the 1960s, the days of free love.
Annotating for discourse information (usually spoken dialogue corpora):

- speech acts (ex. accept, acknowledge, answer, confirm, correct, direct, echo, exclaim, greet)
- speech act forms (ex. declarative, yes-no question, imperative, etc.)

Coreference annotation:

- keep track of an entity that is mentioned throughout a text
  - ex: Kim\textsubscript{4} said ... Sandy\textsubscript{6} told him\textsubscript{4} that she\textsubscript{6} would ...
  - ex: \textless COREF ID='100'\textgreater The Kenya Wildlife Service\textless /COREF\textgreater estimates that \textless COREF ID='101' TYPE=IDENT REF='100'\textgreater it\textless /COREF\textgreater loses $1.2 million a year in park entry fee...
Error Tagging

- Error tagging is often done on learner corpora

- Cambridge Learner Corpora (CLC) and the Longman Learner’s Corpus are tagged for errors, as well as numerous other learner corpora

- Error types used for CLC include:
  
  ➢ wrong word form used
  ➢ something missing
  ➢ word/phrase that needs replacing
  ➢ unnecessary word/phrase
  ➢ wrongly derived words
  ➢ ex:

    For example, my friend told me if I knew about Shakespeare. But, <TIP id=17-56 etype=24 tutor="I knew">I know</TIP>
about him <TIP id=17-57 etype=10 tutor="a little bit">little bit</TIP>, so I couldn ’t <TIP id=17-56679-5 etype=15 tutor="explain it">explain</TIP> fairly to her.
For most annotation we don’t know the answer

Q How can we test whether the annotation is correct (and reproducible)?

A Tag with multiple annotators, measure the agreement
Approaches to Annotation

- Multiple annotators, discard outliers
- Multiple annotators, majority tag
- Two annotators, adjudication for disputes
- Single annotator, adjudication vs model
- Single annotator
Cohen’s Kappa Coefficient

Cohen’s kappa coefficient, also known as the Kappa statistic, is a better way of measuring agreement that takes into account the probability of agreeing by change. $\kappa$ is defined as:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- $\Pr(a)$ is the relative observed agreement among raters
- $\Pr(e)$ is the probability of chance agreement, calculated using the annotated data to estimate the probabilities of each observer randomly saying each category
- If the raters are in complete agreement then $\kappa = 1$
A project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, and the Association for Computers in the Humanities encoding guidelines: [link:http://www.tei-c.org]

It defines how documents should be marked-up with the mark-up language SGML (or more recently XML)
XML

- XML: Extensible Markup Language similar to HTML has no fixed semantics: user defines what tags mean

- recognized as international ISO standard

- formally verifiable via document type definitions (DTD)

- tools available for editing, displaying,
<CATALOG>
  <CD>
    <TITLE>Empire Burlesque</TITLE>
    <ARTIST>Bob Dylan</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>Columbia</COMPANY>
    <PRICE>10.90</PRICE>
    <YEAR>1985</YEAR>
  </CD>
  <CD>
    <TITLE> Greatest Hits</TITLE>
    <ARTIST>Dolly Parton</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>RCA</COMPANY>
    <PRICE>9.90</PRICE>
    <YEAR>1982</YEAR>
  </CD>
</CATALOG>
Each text that is conformant with the TEI guidelines consists of two parts

➢ **Header**

➢ author
➢ title
➢ date the edition or publisher used in creating the machine-readable text
➢ information about the encoding practices adopted

…

➢ **Body**

➢ The actual annotated text
Subject: The staffing in the Commission of the European Communities

Can the Commission say:

1. how many temporary officials are working at the Commission?
2. who they are and what criteria were used in selecting them?

Answer given by Mr Cardoso e Cunha on behalf of the Commission (22 September 1992)

1 and 2. The Commission will send tables showing the number
staff working for the Commission directly to the Honourable Member's Secretariat.</p></div>
Stand off Annotation

➤ separate the tags and text

➤ link back to character (or byte positions)

➤ This is a pen

➤ <pos='noun' cfrom='0' cto='4'>
➤ <pos='verb' cfrom='5' cto='7'>
Dynamic Annotation

➤ Update the tagging as the theory develops!
➤ Annotation is a link between text and a model
➤ As the model changes, update the annotation

➤ Idea developed in the Redwoods project for HPSG annotation

Case Study: the Hinoki Corpus

➢ Grammar-based syntactic annotation using discriminants
Approaches to Treebanking

➢ Manual Annotation

➢ Semi-Automatic

➢ Parse and repair by hand: Penn WSJ, Kyoto Corpus
  ↑ 100% cover, reasonably fast
  ↓ Often inconsistent, Hard to update,
      Simple grammars only (prop-bank is separate)

➢ Parse and select by hand: Redwoods, Hinoki
  ↑ All parses grammatical, Feedback to grammar
      Both syntax and semantics, Easy to update
  ↓ Cover restricted by grammar
Discriminant-based Treebanking

➢ Calculate elementary discriminants (Carter 1997)
  ➢ Basic contrasts between parses
  ➢ Mostly independent and local
  ➢ Can be syntactic or semantic

➢ Select or reject discriminants until one parse remains
  ➢ \[|\text{decisions}| \propto \log |\text{pares}|\]

➢ Alternatively reject all parses
  ➢ i.e, the grammar can not parse successfully
Our Contribution

- Evaluation of discriminant-based treebanking
  - Evaluated inter-annotator agreement
    - 5,000 sentences with 3 annotators
  - Measured number of decisions needed

- Improved Efficiency of Treebanking (Blazing)
  - Use Part-of-Speech tags to pre-select discriminants
Japanese Semantic Lexicon (Lexeed)

➤ All words from the NTT Goitokusei with familiarity $\geq 5.0$

➤ Japanese Lexicon defining word familiarity for each word
➤ Familiarity is estimated by psychological experiments

➤ 28,000 words and 46,347 senses

➤ Covers 75% of tokens in a typical newspaper

➤ Rewritten so definition sentences use only basic words
(and function words): 4 candidates/definition
Syntactic View of *kāten* “curtain”

```
UTTERANCE
  NP
    VP
      PP
        NP
          DET
          N
          CASE-P
          a certain
          monogoto
          to
          hide
    N
      mono
```

The diagram illustrates the syntactic structure of the sentence “あるあるある物事を隠すあるあるある物を隠すあるあるある物” (aru monogoto o kakusu mono aru monogoto o kakusu mono aru monogoto o kakusu mono), which translates to “hide a certain stuff.”
Semantic View of *kāten* “curtain”

\[
\langle h_0, x_2 \{ h_0 : \text{prpstn}_{\text{rel}}(h_5) \\
\quad h_1 : \text{aru}(e_1, x_1, u_0) \quad \text{“a certain”} \\
\quad h_1 : \text{monogoto}(x_1) \quad \text{“stuff”} \\
\quad h_2 : \text{u_def}(x_1, h_1, h_6) \\
\quad h_5 : \text{kakusu}(e_2, x_2, x_1) \quad \text{“hide”} \\
\quad h_3 : \text{koto}(x_2) \quad \text{“thing”} \\
\quad h_4 : \text{u_def}(x_2, h_3, h_7) \rangle
\]
Derivations of *kāten* “curtain” \(_2\) (4/6)
## Discriminants

<table>
<thead>
<tr>
<th>$DA$</th>
<th>rules / lexical types</th>
<th>subtrees / lexical items</th>
<th>parse trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rel-cl-sbj-gap</td>
<td>ある物事を隠す</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rel-clause</td>
<td>ある物事を隠す</td>
<td></td>
</tr>
<tr>
<td>- ?</td>
<td>rel-cl-sbj-gap</td>
<td>ある</td>
<td></td>
</tr>
<tr>
<td>- ?</td>
<td>rel-clause</td>
<td>ある</td>
<td></td>
</tr>
<tr>
<td>+ ?</td>
<td>hd-specifier</td>
<td>ある</td>
<td></td>
</tr>
<tr>
<td>? ?</td>
<td>subj-zpro</td>
<td>隠す</td>
<td>2,4,6</td>
</tr>
<tr>
<td>- ?</td>
<td>subj-zpro</td>
<td>ある</td>
<td>5,6</td>
</tr>
<tr>
<td>- ?</td>
<td>aru-verb-lex</td>
<td>ある</td>
<td>3–6</td>
</tr>
<tr>
<td>+ +</td>
<td>det-lex</td>
<td>ある</td>
<td>1,2</td>
</tr>
</tbody>
</table>

+: positive decision  
-: negative decision  
?: indeterminate / unknown
Blazing POS tags reduced discriminants shown by 20%
## Utterance Inter-annotator Agreement

<table>
<thead>
<tr>
<th></th>
<th>$\alpha - \beta$</th>
<th>$\beta - \gamma$</th>
<th>$\gamma - \alpha$</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parse Agreement</td>
<td>63.9%</td>
<td>68.2%</td>
<td>64.2%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Parse Disagreement</td>
<td>17.5%</td>
<td>19.2%</td>
<td>17.9%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Reject Agreement</td>
<td>4.8%</td>
<td>3.0%</td>
<td>4.1%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Reject Disagreement</td>
<td>13.7%</td>
<td>9.5%</td>
<td>13.8%</td>
<td>12.4%</td>
</tr>
</tbody>
</table>

➢ Total Agreement 69.4%

➢ *cf* German NeGra corpus: 52% (Newspaper text)

➢ Average of 190 parses/sentence (10 words)
## Mutual Labeled Precision F-Score

<table>
<thead>
<tr>
<th>Test Set</th>
<th>(\alpha - \beta)</th>
<th>#</th>
<th>F</th>
<th>(\beta - \gamma)</th>
<th>#</th>
<th>F</th>
<th>(\gamma - \alpha)</th>
<th>#</th>
<th>F</th>
<th>Average</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>507</td>
<td>96.03</td>
<td>516</td>
<td>96.22</td>
<td>481</td>
<td>96.24</td>
<td>96.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>505</td>
<td>96.79</td>
<td>551</td>
<td>96.40</td>
<td>511</td>
<td>96.57</td>
<td>96.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>489</td>
<td>95.82</td>
<td>517</td>
<td>95.15</td>
<td>477</td>
<td>95.42</td>
<td>95.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>454</td>
<td>96.83</td>
<td>477</td>
<td>96.86</td>
<td>447</td>
<td>97.40</td>
<td>97.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>480</td>
<td>95.15</td>
<td>497</td>
<td>96.81</td>
<td>484</td>
<td>96.57</td>
<td>96.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2435</td>
<td>96.32</td>
<td>2558</td>
<td>96.28</td>
<td>2400</td>
<td>96.47</td>
<td>96.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Blazing Using POS scores

(verb-aux      v-stem-lex −1.0)
(verb-main     aspect-stem-lex −1.0)
(noun          verb-stem-lex −1.0)
(adnominal     noun_mod-lex-1 0.9
det-lex 0.9)
(conjunction   n_conj-p-lex 0.9
v-coord-end-lex 0.9)
(adjectival-noun noun-lex −1.0)
Blazing POS tags reduced decisions by 19.5%
## Number of Decisions with Blazing

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Annotator Decisions</th>
<th>Blazed Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>A</td>
<td>2,659</td>
<td>2,606</td>
</tr>
<tr>
<td>B</td>
<td>2,848</td>
<td>2,939</td>
</tr>
<tr>
<td>C</td>
<td>1,930</td>
<td>2,487</td>
</tr>
<tr>
<td>D</td>
<td>2,254</td>
<td>2,157</td>
</tr>
<tr>
<td>E</td>
<td>1,769</td>
<td>2,278</td>
</tr>
</tbody>
</table>
Current Status

- Finished treebanking

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentences</th>
<th>Words</th>
<th>Content Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>75,000</td>
<td>690,000</td>
<td>320,000</td>
</tr>
<tr>
<td>Example</td>
<td>46,000</td>
<td>500,000</td>
<td>220,000</td>
</tr>
</tbody>
</table>

- But data not released by NTT, so I left them

- Currently treebanking the Wordnet Definitions
Conclusions

- 5,000 sentences are annotated by three different annotators

- average inter-annotator agreement
  - 65.4% (sentence)
  - 83.5% using labeled precision
  - 96.6% on ambiguous annotated trees

- blazing POS tags reduced decisions by 19.5%
Lab 1
Non-exact matching

- Often you will want to match not just a word or sequence of words but some kind of pattern

- Various corpus interface tools make it easy to do this

- A standard way is to match regular expressions

- A good text editor should allow regular expression matching e.g., EMACS, notepad++
Regular Expressions
## Regular Expressions

> Regular expressions: a formal language for matching things.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>any single character</td>
</tr>
<tr>
<td>[ ]</td>
<td>a single character that is contained within the brackets.</td>
</tr>
<tr>
<td></td>
<td>[a-z] specifies a range which matches any letter from &quot;a&quot; to &quot;z&quot;.</td>
</tr>
<tr>
<td>[^ ]</td>
<td>a single character not in the brackets.</td>
</tr>
<tr>
<td>^</td>
<td>the starting position within the string/line.</td>
</tr>
<tr>
<td>$</td>
<td>the ending position of the string/line.</td>
</tr>
<tr>
<td>*</td>
<td>the preceding element zero or more times.</td>
</tr>
<tr>
<td>?</td>
<td>the preceding element zero or one time.</td>
</tr>
<tr>
<td>+</td>
<td>the preceding element one or more times.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>\</td>
<td>escapes the following character.</td>
</tr>
</tbody>
</table>
Regular Expression Examples

- `at` matches any three-character string ending with "at", including "hat", "cat", and "bat".

- `[hc]at` matches "hat" and "cat".

- `[^b]at` matches all strings matched by `.at` except "bat".

- `^[hc]at` matches "hat" and "cat", but only at the beginning of the string or line.

- `[hc]at$` matches "hat" and "cat", but only at the end of the string or line.

- `\[.\]` matches any single character surrounded by "[" and "]" since the brackets are escaped, for example: "]a[" and "]b[".

http://www.regexplanet.com/simple/
Wild Cards

 temperatura a wildcard character substitutes for any other character or characters in a string.

➢ Files and directories (Unix, CP/M, DOS, Windows)
  * matches zero or more characters
  ? matches one character
  [ ] matches a list or range of characters
  * E.g.: Match any file that ends with the string “.txt” or “.tex”.
    ls *.txt *.tex

➢ Structured Query Language (SQL)
  % matches zero or more characters
  _ matches a single character
### BYU Interface Specialties

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Explanation</th>
<th>Example</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>One exact word</td>
<td>mysterious</td>
<td>mysterious</td>
</tr>
<tr>
<td>[pos]</td>
<td>Part of speech</td>
<td>[vvg]</td>
<td>going, using</td>
</tr>
<tr>
<td>[pos*]</td>
<td>Part of speech</td>
<td>[v*]</td>
<td>find, does, keeping, started</td>
</tr>
<tr>
<td>[lemma]</td>
<td>Lemma</td>
<td>[sing]</td>
<td>sing, singing, sang</td>
</tr>
<tr>
<td>[=word]</td>
<td>Synonyms</td>
<td>[=strong]</td>
<td>formidable, muscular, fervent</td>
</tr>
<tr>
<td>word</td>
<td>wurd</td>
<td>Any of the words</td>
<td>stunning</td>
</tr>
<tr>
<td>x?xx*</td>
<td>wildcards</td>
<td>on*ly</td>
<td>only, ontologically, on-the-fly,</td>
</tr>
<tr>
<td>x?xx*</td>
<td>wildcards</td>
<td>s?ng</td>
<td>sing, sang, song</td>
</tr>
<tr>
<td>-word</td>
<td>negation</td>
<td>−[nn*]</td>
<td>the, in, is</td>
</tr>
<tr>
<td>word.[pos]</td>
<td>Word AND pos</td>
<td>can.[v*]</td>
<td>can, canning, canned (verbs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>can.[n*]</td>
<td>can, cans (nouns)</td>
</tr>
</tbody>
</table>

http://corpus.byu.edu/bnc/help/syntax_e.asp
Acknowledgments

➤ Thanks to Na-Rae Han for inspiration for some of the slides (from *LING 2050 Special Topics in Linguistics: Corpus linguistics*, U Penn).

➤ Thanks to Sandra Kübler for some of the slides from her *RoCoLi Course: Computational Tools for Corpus Linguistics*

➤ Definitions from WordNet 3.0

---

1 *Romania Computational Linguistics Summer School*