HG3051 Corpus Linquistics

Multimodal and Multilingual Corpora

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 3
https://github.com/bond-lab/Corpus-Linguistics

HG3051 (2020)
Overview

➤ Revision of Annotation
  ➤ Mark-up
  ➤ Annotation
  ➤ Regular Expressions
  ➤ The Hinoki Corpus

➤ Multi-modal Corpora

➤ Multi-lingual Corpora
Revision of Annotation
Corpus Annotation vs. Mark-Up

➤ **Mark up** provides objectively verifiable information

➤ Authorship
➤ Publication dates
➤ Paragraph boundaries
➤ Source text (URL, Book, …)
➤ License

➤ **Annotation** provides interpretive linguistic information

➤ Sentence/Utterance boundaries
➤ Tokenization
➤ Part-of-speech tags, Lemmas, Concepts
➤ Sentence structure (syntax, co-reference, roles)
➤ Domain, Genre
Dublin Core Ontology

➤ Goals

➤ Provides a semantic vocabulary for describing the “core” information properties of resources (electronic and “real” physical objects)
➤ Enables intelligent resource discovery systems

➤ Fifteen Elements:

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

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

➤ Instantiation (4)
  ➤ Date, Language, Format, Identifier

Geoffrey Leech’s Seven Maxims of Annotation

1. Annotation should be separable from text, leaving the raw corpus.

2. It should be possible to extract just the annotations from the text.

3. The annotation guidelines should be available.

4. Who did the annotation and how should be made clear.

5. The possibilities of errors should be made clear.

6. Annotation schemes should be theory-neutral

7. Standards emerge through practical consensus.
Types of Corpus Annotation

- Tokenization, Lemmatization
- Part-of-speech
- Syntactic analysis (chunks, parses)
- Semantic analysis (word senses, semantic roles)
- Discourse and pragmatic analysis (co-reference, time)
- Phonetic, phonemic, prosodic annotation
- Error tagging
How is Corpus Annotation Done?

- Mainly semi-automatic (done first by computer programs; post-edited)
  1. An small 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

- Large corpora often fully automatic
  - Segmentation
  - Part-of-speech tagging: accuracy of 97%
  - Lemmatization

- Corpora should indicate reliability of tags
  - Inter-annotator agreement, kappa (human)
  - Tagger accuracy (machine)
How are corpora represented?

- Far too many encoding schemes (TEI is common)
  - Header: for markup
  - Body: for annotation

- Text: one sentence per line, POS affixed can_{VV}

- XML: `<s sid='1'><w wid='1' pos='vv'>can</w></s>`

- XML standoff:
  
  - Text file:
  
  - Corpus file:

- Often stored in a database in applications
Best Practices

➤ XML

➤ Header and documentation

➤ Open license

➤ Maintained (errors corrected, dynamic updates)
Case Study: the Hinoki Corpus

- Grammar-based syntactic annotation using discriminants
  - Parse the corpus and select the best parse
    * discriminant-based selection is efficient
  - Guarantees consistency
  - Loses some trees
Discussant-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
  - \(|\text{decisions}| \propto \log |\text{parses}|\)

- Alternatively reject all parses
  - i.e., the grammar can not parse successfully
Derivations of *kāten* “curtain” (4/6)

**Tree #1 (Correct)**

NP-frag

rel-cl-sbj-gap

hd-complement

N

hd-complement

V

hd-specifier

DET

N

CASE-P

adnominal noun particle verb noun

a certain thing hide thing

**Tree #2**

NP-frag

rel-clause

hd-complement

N

hd-complement

V

hd-specifier

DET

N

CASE-P

adnominal noun particle verb noun

a certain thing hide thing

**Tree #3**

NP-frag

rel-cl-sbj-gap

hd-complement

N

hd-complement

V

rel-cl-sbj-gap

V

N

CASE-P

verb noun particle verb noun

exist thing hide thing

**Tree #4**

NP-frag

rel-clause

hd-complement

N

hd-complement

V

rel-cl-sbj-gap

V

N

CASE-P

verb noun particle verb noun

exist thing hide thing
Hinoki — Summary

- 5,000 sentences are annotated by three different annotators

- average inter-annotator agreement
  - 65.4% (sentence)
  - 83.5% using labeled precision
    - bracket with same label in the same place
  - 96.6% on ambiguous annotated trees
    - most disagreement is in if the tree is good or not

- Hinoki corpus was then extended to another 30,000 trees
## 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>
Wild Cards

- 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\wurd</td>
<td>Any of the words</td>
<td>stunning</td>
<td>gorgeous</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
Multi-Modal Corpora
Multi-modal Corpora

Language is not the only channel for communication: It is often combined with other modalities

- speech
- gesture
- facial expression
- gaze
- body posture
  - ECG (Electrocardiogram), HR (Heart Rate), GSR (Galvanic Skin Response)
- activity: nursing, drawing, building
- orthographic cues (color, size, font choice, ...)

Corpora that include more than one of these are multi-modal
HCRC Map Task

- Early, influential dialog corpus (with maps)

- Task
  - Two speakers sit opposite one another
  - Each has a map which the other cannot see
  - The Instruction Giver has a route marked on their map
  - The Instruction Follower has no route
  - The goal is to reproduce the Instruction Giver’s route on the Instruction Follower’s map
  - The maps are not identical and the speakers are told this

- Conditions
  - familiar (friends) vs non-familiar
  - gaze vs no-gaze

http://groups.inf.ed.ac.uk/maptask/
The Maps
Some design points

Landmarks chosen for phonetic properties

- /t/-deletion eg vast meadow
- /d/-deletion eg reclaimed fields
- glottalisation eg chestnut tree
- nasal assimilation eg broken gate

Making the data maximally useful

Annotation

- POS, parse
- Discourse structure
- Gaze

Now replicated in many languages and dialects (Dutch, Italian, Japanese, Swedish, Occitan, Portuguese, Australian, American and British English)
E-Nightingale: Nursing Task Corpus

- Japanese project to analyze Nursing tasks and dialogs
- Recorder worn all day
- Beeps at ten minute intervals (event-driven recording)
  - Nurse records what they are doing
- Linked to location
- Very hard speech to decode

Hiromi Itoh Ozaku; Akinori Abe; Noriaki Kuwahara; Futoshi Naya; Kiyoshi Kogure; Kaoru Sagara *Building Dialogue Corpora for Nursing Activity Analysis* in LINC-2005

www.aclweb.org/anthology/I05-6005.pdf
Lei Chen (2007) *VACE Multimodal Meeting Corpus* Virginia Polytechnic Institute and State University (Video online at: http://videolectures.net/mlmi04uk_chen_vmmc/ accessed 2010-02-10)
They had me at ‘attacked and occupied by wombats’

Industrial scale coding:

- motion capture
- gaze
- speech automatic; OOV by experts; further checking
- detailed information about the participants
- detailed information about the task

I think there should be newer corpora than this, but could not find one with such complete description.
British Sign Language Corpus

  - 249 deaf signers of BSL from 8 regions around the UK: London (L), Bristol (BL), Cardiff (CF), Birmingham (BM), Newcastle (N), Manchester (M), Glasgow (G) and Belfast (BF).
  - Mixed for gender, age group, age of BSL acquisition, social class and ethnicity
  - Interviews (i); conversation (c); narrative followed by conversation (n-c)

- Marked up with the above information

- Not yet annotated or searchable
  - Some annotation available
  - Done in different sites

http://www.bslcorpusproject.org
Multi-Lingual Corpora
Multilingual corpora are useful for

- Contrastive linguistic analysis
  - Comparing distributions between languages
  - Learning about translations
  - Using one language to describe the other
- Language learning
  - Teaching new phenomena in terms of what you already know
- Machine translation training
  - Learning translations directly
Europarl

- Large automatically aligned corpus of European parliament proceedings
- Translation between EU languages (EU funded project)
- 18-40 million words, .6–1.3 million sentences
- Freely available text in all European Languages
- Used in the Euro Matrix MT project

*Europarl: A Parallel Corpus for Statistical Machine Translation*, Philipp Koehn, MT Summit 2005

# Euro Matrix SMT Results

<table>
<thead>
<tr>
<th>Output Language</th>
<th>Danish</th>
<th>Dutch</th>
<th>German</th>
<th>Greek</th>
<th>English</th>
<th>Finnish</th>
<th>French</th>
<th>Italian</th>
<th>Portuguese</th>
<th>Spanish</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28.79</td>
<td>24.67</td>
<td>27.75</td>
<td>32.15</td>
<td>25.39</td>
<td>35.37</td>
<td>35.37</td>
<td>39.04</td>
<td>40.27</td>
<td>29.77</td>
<td>38.47</td>
</tr>
<tr>
<td></td>
<td>22.22</td>
<td>20.07</td>
<td>21.36</td>
<td>26.84</td>
<td>27.10</td>
<td>35.37</td>
<td>35.37</td>
<td>32.07</td>
<td>32.31</td>
<td>23.94</td>
<td>38.47</td>
</tr>
<tr>
<td></td>
<td>24.32</td>
<td>20.71</td>
<td>23.28</td>
<td>27.67</td>
<td>30.16</td>
<td>38.47</td>
<td>35.37</td>
<td>32.07</td>
<td>35.37</td>
<td>25.95</td>
<td>38.47</td>
</tr>
<tr>
<td></td>
<td>26.49</td>
<td>22.95</td>
<td>25.49</td>
<td>31.26</td>
<td>30.16</td>
<td>38.47</td>
<td>35.37</td>
<td>32.07</td>
<td>35.37</td>
<td>28.66</td>
<td>38.47</td>
</tr>
</tbody>
</table>

**Note:** BLEU scores represent the performance of the SMT models in different languages.
Euro Matrix Discussion

- Linguistic similarity affects the statistical machine translation score:
  - Highest: Spanish $\rightarrow$ French (BLEU = 40.27)
  - Lowest: Italian $\rightarrow$ Finnish (BLEU = 11.08)

- Translation done using the open source SMT System: Moses <statmt.org>

- Creating all $n(n-1)$ language pairs took a week
  - It is easy to train new systems if you have a multi-lingual corpus
Interesting facts

- Also used for lexicon and thesaurus construction

- Several English on-line translations are actually French and no-one had noticed

- Almost entirely constructed automatically

- New languages being added to the EU means more data
OPUS

» On-line collection of multilingual text

» Mainly automatically created

» OPUS multilingual
» Europarl
» OpenSubtitles
» EUconst
» Word Alignment Database

» Slightly hard to use interface

http://opus.lingfil.uu.se/
Opus Downloads & Samples

- EMEA - European Medicines Agency documents (5.0 GB)
- EUconst - The European constitution (67 MB)
- EUROPARL - European Parliament Proceedings (3.6 GB)
- OO - the OpenOffice.org corpus (34 MB)
- OpenSubs - the opensubtitles.org corpus (1.3 GB)
- KDE4 - KDE4 localization files (v.2) (1.4 GB)
- KDEdoc - the KDE manual corpus (35 MB)
- PHP - the PHP manual corpus (172 MB)
- SETIMES - A parallel corpus of the Balkan languages 2.3 GB)
- SPC - Stockholm Parallel Corpora (3.5 MB)
Taoteba

➤ User generated corpus of example sentences
➤ Not authentic at all
➤ Short and well aligned (thus easy to process)

➤ Used for teaching, learning and MT research

(1)

(\textit{jp})

ano ki no eda ni suu hiki no tori ga tomatte iru .
that tree of branch on some wing of bird SBJ stop be .

"Some birds are sitting on the branch of that tree." (en)

"Des oiseaux se reposent sur la branche de cet arbre." (fr)

(also Hebrew, Esperanto, Italian: added since 2009)
Task

➤ Pick a couple of relatively basic words: *dog*, *tree*, *all* for which you know the translation in some language.

➤ Look at the word in the OPUS subtitle corpus and Tatoeba

  ➤ How often is it translated into a word you know?
  ➤ How often is it not translated at all?

➤ Now try one of the more technical corpora

Translations can be very surprising.
Other Large Multilingual Corpora

➤ Canadian Hansard
➤ Hong Kong Hansard
➤ Bible Translation Corpus
➤ Universal Declaration of Human Rights
➤ Swadesh list
➤ GALE Chinese-English, Japanese-English (DoD)
➤ NICT Japanese-English, Japanese-Chinese
➤ NTU Multilingual Corpus
Sentence Alignment

- Various ways to align sentences

- Length-based
  - Gale-Church algorithm (basically match sentence length in characters)

- Other lexical methods are also popular
  - match using a dictionary
  - content words only

- The better the alignment, the easier it is to extract information
Word Alignment

➤ GIZA++ — match words depending on shared position

➤ How many times two words appear in the same sentence pair
➤ Hard to match very free translations, also MWEs

➤ Lexically-based (using dictionaries and thesauruses) are also common

➤ Typically newspapers may have length differences of up to a third

➤ The more direct the translation, the easier it is to align
➤ Many things do not align clearly
Introduce Lab 2

➤ Prepare a short intro to a corpus

➤ Present in class