HG351 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
http://compling.hss.ntu.edu.sg/courses/hg3051/

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

- OLAC Lang. Resource Catalog: `http://search.language-archives.org/`

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 mark up
  - 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:

  can (text file)

  `<w pos='vv' cfrom='0' cto='3' />` (corpus file)
- Often stored in a database in applications

Best Practices

- XML
- Header and documentation
- Open license
- Maintained (errors corrected, dynamic updates)

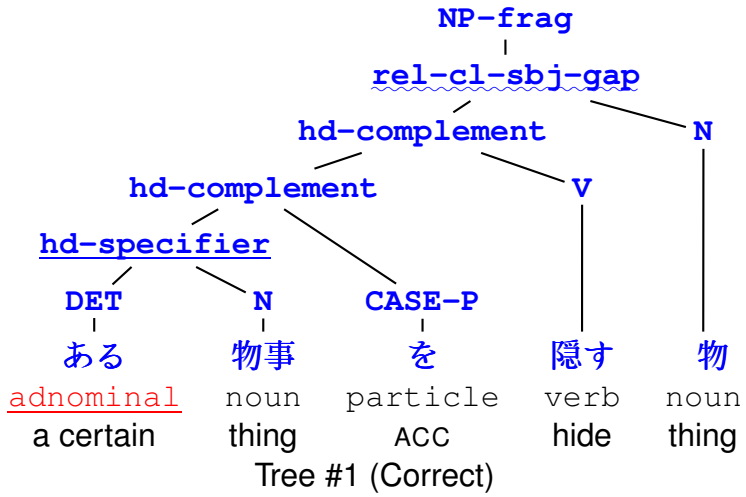
Case Study: the Hinoki Corpus

- Grammar-based syntactic annotation using discriminants
  - Parse the corpus and select the best parse
    - discriminant-based selection is efficient
  - Guarantees consistency
  - Loses some trees

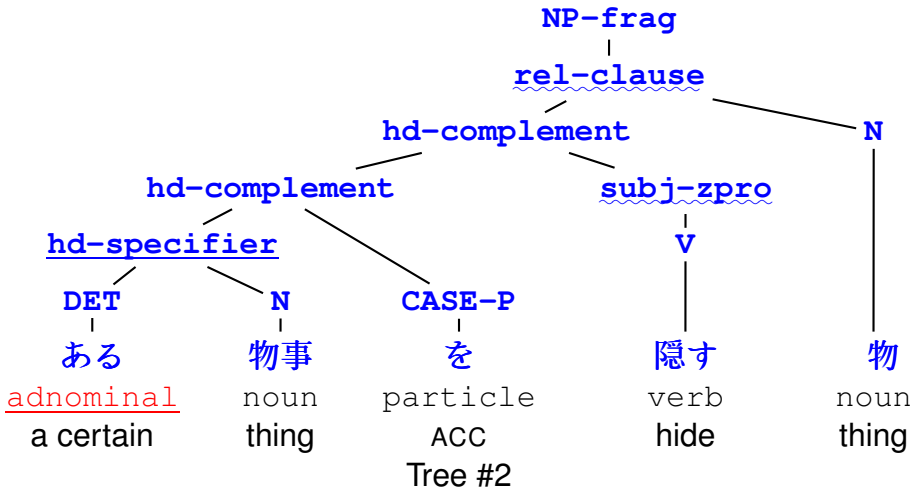
Discriminant-based Treebanking

- 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
  - $|\text{decisions}| \propto \log |\text{parses}|$
- Alternatively reject all parses
  - i.e, the grammar can not parse successfully

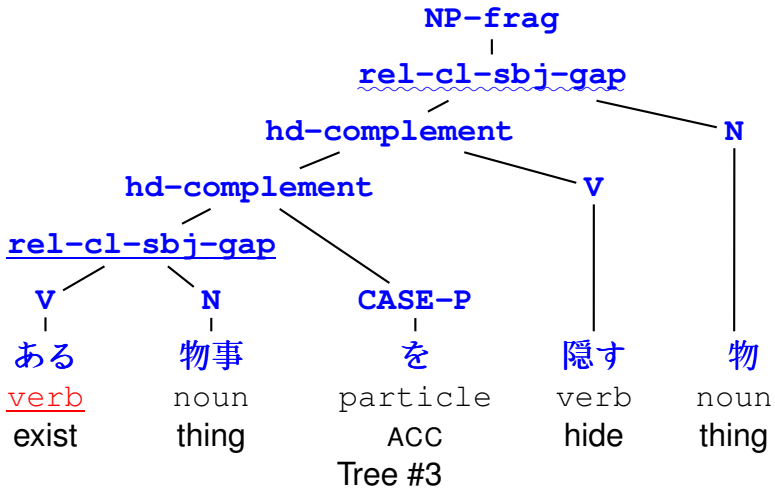
Derivations of *kāten* “curtain”$_2$ (4/6)



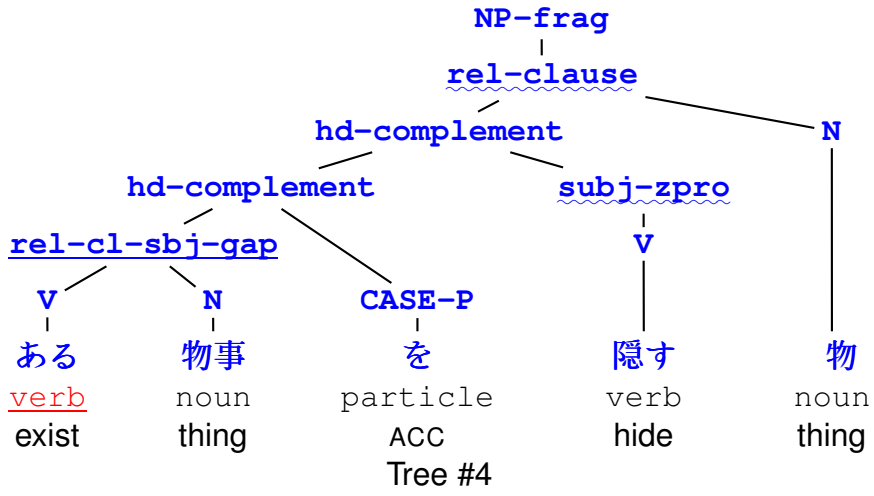
Tree #1 (Correct)



Tree #2



Tree #3



Tree #4

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.

Symbol	Matches
.	any single character
[ ]	a single character that is contained within the brackets.
	[a-z] specifies a range which matches any letter from ”a” to ”z”.
[^]	a single character not in the brackets.
^	the starting position within the string/line.
$	the ending position of the string/line.
*	the preceding element zero or more times.
?	the preceding element zero or one time.
+	the preceding element one or more times.
|	either the expression before or after the operator.
\	escapes the following character.

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

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

Multi-Modal Corpora

Multi-modal Corpora

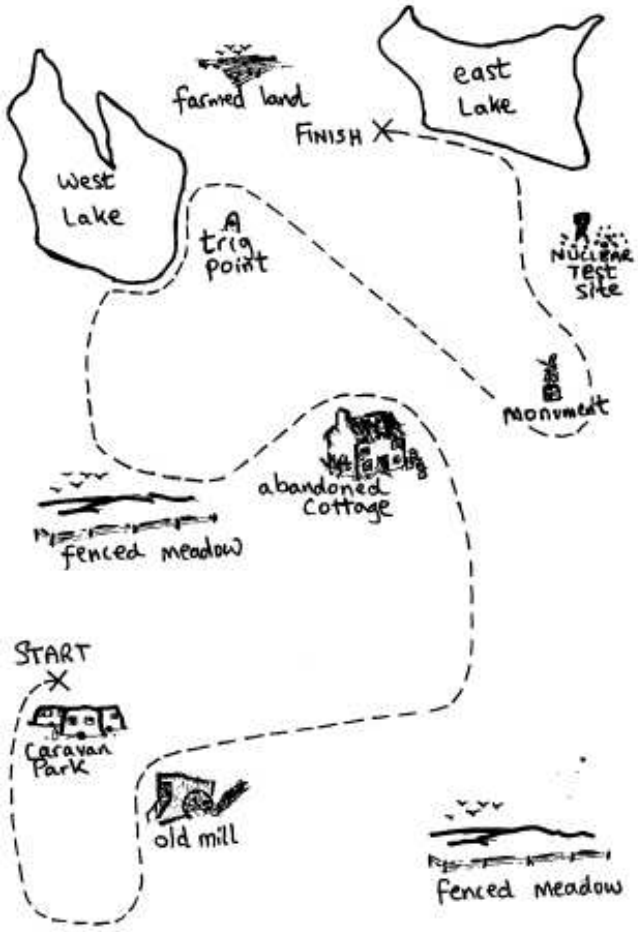
- Language is not the only channel for communication: It is often combined with other modalities
  - speech
  - gesture
  - facial expression
  - gaze
  - body posture
    ECG (Electrocardiogram), HR (Heart Rate), GSR (Galvanic Skin Response)
  - activity: nursing, drawing, building
  - orthographic 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

The Maps


farmed land
east Lake
FINISH
West Lake
A trig point
NUCLEAR TEST SITE
Monument
abandoned cottage
fenced meadow
START
Caravan Park
old mill
fenced meadow


farmed Land
east Lake
West Lake
golf course
Monument
abandoned cottage
START
forest
Caravan Park
mill wheel
pricket fence
fenced meadow

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

VACE Multimodal Meeting Corpus

Lei Chen (2007) *VACE Multimodal Meeting Corpus* Virginia Polytechnic Institute and State University (Video online at: `http://videolectures.net/mlmi` accessed 2010-02-10)

VACE comments

- 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

British Sign Language Corpus

- Collection of sign language recordings (2008–2011)
  - 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

Multi-Lingual Corpora

Bitexts and more

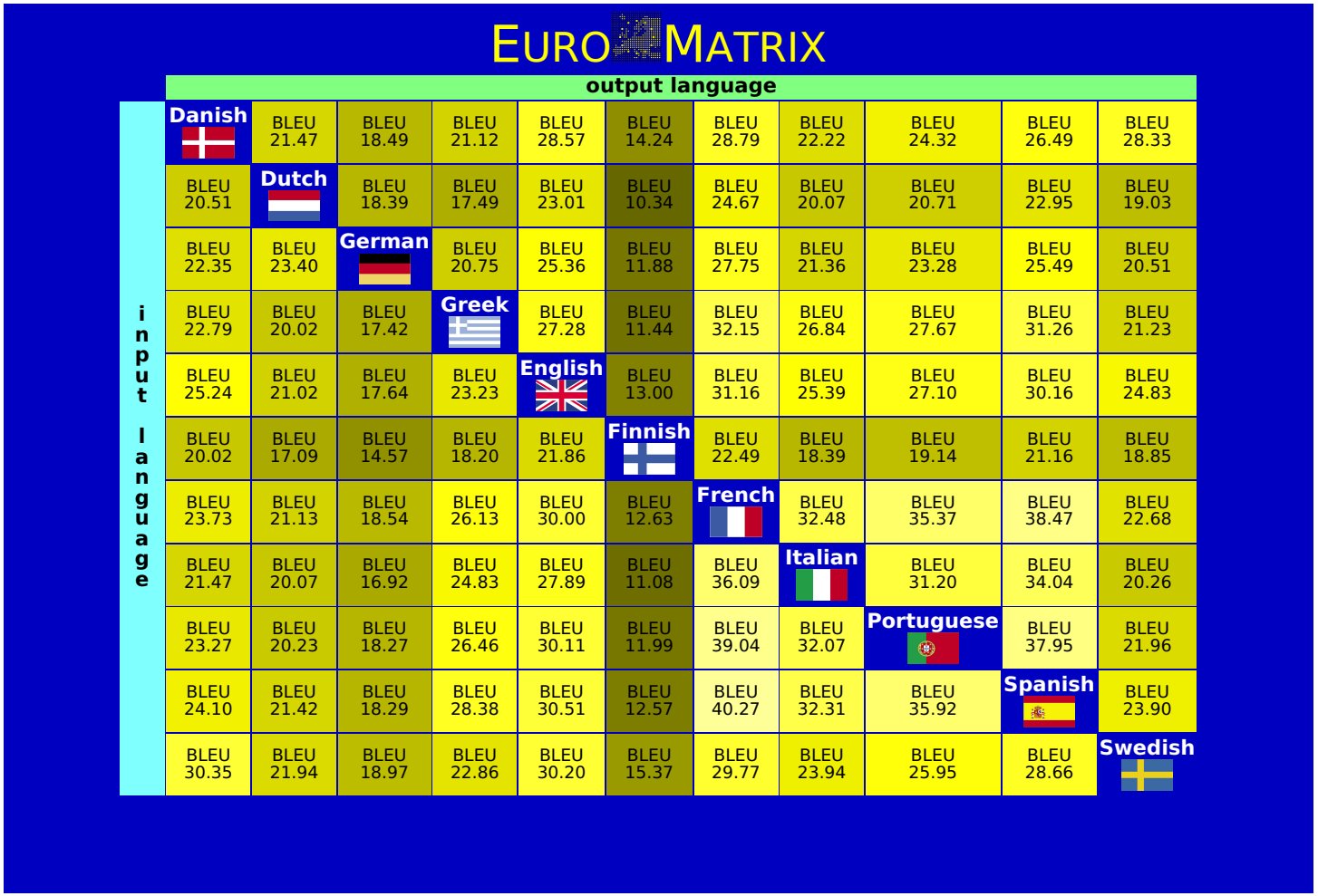
- Multilingual corpora are useful for
  - Contrastive linguistic analysis
    - Comparing distributions between languages
    - Learning about translations
    - using one language to describe the other
  - Language learning
    - Teaching new phenomena in terms of what you already know
  - Machine translation 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



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

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) あの木の枝に数羽の鳥がとまっている。

あの	木	の	枝	に	数	羽	の	鳥	が	とまって	いる	。	*(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

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

Word Alignment

- GIZA++ — match words depending on shared position
  - How many times to 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