COR: Corpus Linquistics

Lecture 4
Survey of Available Corpora

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

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

COR (2024)

Overview

- Revision
  - Multi-modal Corpora
  - Multi-lingual Corpora
- Survey of Corpora
  - Corpora that caught your interest
  - Corpora that caught my interest

Revision of Multi-modal and Multi-lingual 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
- Corpora that include more than one of these are multi-modal

HCRC Map Task

- Early, influential dialog corpus (with maps)
  - Two speakers sit opposite one another
  - Each has a (different) map which the other cannot see
  - One explains the route to the other
- Conditions
  - familiar (friends) vs non-familiar
  - gaze vs no-gaze
- Landmarks chosen for phonetic properties
- Annotation: POS, Parse, Discourse structure, Gaze
- Now replicated in many languages and dialects

http://groups.inf.ed.ac.uk/maptask/

Other Multimodal Corpora

- 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)

- Many Meeting Corpora
  - VACE Multimodal Meeting Corpus
    - extra linguistic information very important

Multi-Lingual Corpora

- Multilingual corpora are useful for
  - Contrastive linguistic analysis
  - Language learning
  - Machine translation training
- Well known Corpora
  - Europarl: 20+ languages; 18-40 million words, .6–1.3 million sentences
  - OPUS: On-line collection of multilingual text
  - Taoteba: User generated corpus of example sentences
  - Canadian Hansard
  - Hong Kong Hansard
  - Bible Translation Corpus
  - GALE Chinese-English, Japanese-English (DoD)
  - NICT Japanese-English, Japanese-Chinese

Multilingual Corpus Construction

- Other languages used as annotation
  - Assumed high quality
- Other construction often done automatically
- Article, Sentence and Word alignment
  - Length/Structure based cues
  - Lexically based cues
- Driven by MT research
- Not so much high quality alignable multi-lingual text

Your Corpora

Slides

- You need to learn to follow instructions ($\frac{-1}{10}$/issue)
  - pdf not powerpoint
  - no more than 5 pages
  - Name on page one
  - Deadline is not negotiable — late revisions don't count
- For paper/grant submissions
  — your submission paper will be rejected without review
- In the workplace
  — you will get shouted at and made to do it again
- Why is it so important?
  - The reviewer/boss has to read/process many, many submissions
  - Anything that distracts them/takes extra time is bad

- Even if you forget everything about Corpora, remember this lesson

Presentations

- 5 minutes + 2 minutes for question
  - I will indicate time at 4:00 and 5:00 (and 7:00)
  - You must stop talking at 5:00
- Everyone must ask one (new) question!
- We need to keep time strictly
  - Choose one or two points about your corpus to emphasize

My Corpora

Some corpora of interest (and why)

- Open American National Corpus
- Corpus of Hong Kong Cantonese 香港粵語語料庫 (by KK Luke)
- Hinoki Treebank of Japanese
- Redwoods Treebank of English
- Tatoeba Corpus
- NICT Multilingual Corpus/Kyoto Corpus
- NTU Multilingual Corpus

Open American National Corpus

Name	Domain	No. files	No. words
charlotte	face to face	93	198,295
switchboard	telephone	2,307	3,019,477
911 report	government, technical	17	281,093
berlitz	travel guides	179	1,012,496
biomed	technical	837	3,349,714
eggan	fiction	1	61,746
icic	letters	245	91,318
oup	non-fiction	45	330,524
plos	technical	252	409,280
slate	journal	4,531	4,238,808
verbatim	journal	32	582,384
web data	government	285	1,048,792
Total		8,832	14,623,927

A large collection of freely available data

Creation/Annotation

- OANC Annotation
  - Structural markup (sections, chapters, ..., paragraph, sentence)
  - Words (tokens) with part of speech annotations using the Penn tagset
  - Noun, Verb chunks
- Contributed annotations
  - BBN Named Entities (inline format)
  - Syntactic parses
    - ∗ Charniak constituency-based parser (Charniak & Johnson, 2005)
    - ∗ LTH dependency converter (Johansson & Nugues, 2007)
    - ∗ MaltParser (Nivre et al., 2007).
    - ∗ English Resource Grammar (Flickinger, 2008)
  - Slate coreference (anaphora) annotations
  - CLAWS part of speech tags

Manually Annotated Sub-Corpus (MASC)

- 82,000 words drawn from the OANC
  - WordNet senses
  - FrameNet frame annotations
  - Validated annotations for token and sentence boundaries, part of speech, noun chunks, verb chunks, named entities, and Penn Treebank syntactic annotation
  - Language Understanding Corpus annotations
  - Opinion, PropBank, and TimeML are either included in MASC I or forthcoming
- All annotations are in LAF/GrAF format and can therefore be merged or combined using the ANC Tool and transduced to other formats using ANC2Go.

References

- n-gram search http://www.americannationalcorpus.org/OANC/ngram.html
- Ide, N., Baker, C., Fellbaum, C., Passonneau, R. (2010). The Manually Annotated Sub-Corpus: A Community Resource For and By the People. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.
- Ide, N., Suderman, K., Simms, B. (2010). ANC2Go: A Web Application for Customized Corpus Creation. Proceedings of the Seventh Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta.
- Ide, N. (2008). The American National Corpus: Then, Now, and Tomorrow. In Michael Haugh, Kate Burridge, Jean Mulder and Pam Peters (eds.), Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Cascadilla Proceedings Project, Sommerville, MA.

Corpus of Hong Kong Cantonese 香港粵語語料庫

- 180,000-word corpus
- 52 spontaneous conversations
- 42 radio programmes
- Segmented; POS tagged; Romanized
- Available directly for download (no explicit license)
- Produced by KK Luke

Creation/Annotation

- 30 hours of recordings (March 1997 — August 1998)
- Native speakers of Cantonese
- ordinary settings with family members, friends and colleagues talking with each other freely on everyday topics such as current affairs, work and study, and personal hobbies
- Some parts selected

Usage

- Used to examine the uses of the frequently used sentence final particles wo3 and bo3 in the 1990s in Hong Kong Cantonese by examining speech data.
- Question: are wo (喎) and bo (嚩) variant forms?
- Answer: No

> [...] the two SFPs carry and serve different meanings and functions in modern Hong Kong Cantonese, and thus they are not exactly the same particles and not interchangeable as previously assumed. (Leung, 2010, p21)

References

- Wong, P.-W. (2006). The specification of POS tagging of the Hong Kong University Cantonese corpus. *International Journal of Technology and Human Interaction*, 2(1):21–38. DOI10.4018/jthi.2006010102

- Leung, W.-M. (2010). On the identity and uses of Cantonese sentence-final particles in the late 20th century: The case of wo and bo. *Asian Social Science*, 6(1):13–23. (http://www.ccsenet.org/journal/index.php/ass/article/view/4765)

Hinoki Treebank of Japanese

- Based on an HPSG grammar of Japanese (JACY)
- Parsing dictionary definition sentences (Lexeed)
- Creating a corpus that can be studied (Hinoki)
- Creating an ontology that links senses (Ontology)

We want to combine
structural and **lexical**
semantics

Creation/Annotation

- Grammar Development
  - Lexical acquisition from MRDs, Corpora, hand-built
  - Treebanking (Definition and Example sentences)
- Ontology Development
  - Extracting from MRDs
  - Boot-strapping from a closed world

Corpus: Hinoki

Type	# Sents	Tree	Sense
Def.	81,000		
Ex.	46,000		
News	74,000		

Lexicon: Lexeed

Familiarity, Defs, Exs.

Head words	28,000
Senses	46,000

Grammar: JACY (HPSG)

Lex-types, rules, lex-items

Lex-items	37,000
Types	7,000
Rules	114

Meaning: Ontology

links the Lexeed senses

relation types	11
`hypernym, meronym, …`	
relations	81,300

Usage

- Ontology Extraction (also for English)
- Parse Ranking using Semantics (+3.8%)
- Word Sense Disambiguation using Parsing
- POS tagger training

References

➢ Bond, F., Fujita, S., and Tanaka, T. (2008). The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 42(2):243–251. (Reissue of DOI 10.1007/s10579-007-9036-6 as Springer lost the Japanese text)

➢ Tanaka, T., Bond, F., Baldwin, T., Fujita, S., and Hashimoto, C. (2007). Word sense disambiguation incorporating lexical and structural semantic information. In *The 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing (EMNLP) and the Conference on Natural Language Learning (CoNLL)*, pages 477–485, Prague

➢ Fujita, S., Bond, F., Tanaka, T., and Oepen, S. (2010). Exploiting semantic information for HPSG parse selection. *Research on Language and Computation*, 8(1):1–22

Sketch Engine

- corpus manager and text analysis software

- it provides (among other things) word sketches:
  one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour.

- it supports and provides corpora in 90+ languages

- Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The sketch engine. Paper presented at EURALEX

Google N-grams

- an online search engine that charts the frequencies of any set of search strings using a yearly count of n-grams found in printed sources published between 1500 and 2019
  in Google's text corpora in English, Chinese (simplified), French, German, Hebrew, Italian, Russian, or Spanish

- texts are automatically POS tagged

- you can search with limited wildcards

- text comes from OCRs, has many errors in text and meta-data

- Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. In Zhang,

https://books.google.com/ngrams/

M., editor, *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics

➢ Younes, N. and Reips, U.-D. (2019). Guideline for improving the reliability of Google ngram studies: Evidence from religious terms. *PloS one*, 14(3) 10.1371/journal.pone.0213554

Universal Dependencies (UD)

- is a framework for consistent annotation of grammar
  (parts of speech, morphological features, and syntactic dependencies)
- across different human languages
- an open community effort with over 200 treebanks in over 100 languages
- many useful tools and interfaces
  - Grew Match for matching trees

https://universaldependencies.org/

References

- McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics

- Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA)

- Guibon, G., Courtin, M., Gerdes, K., and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association