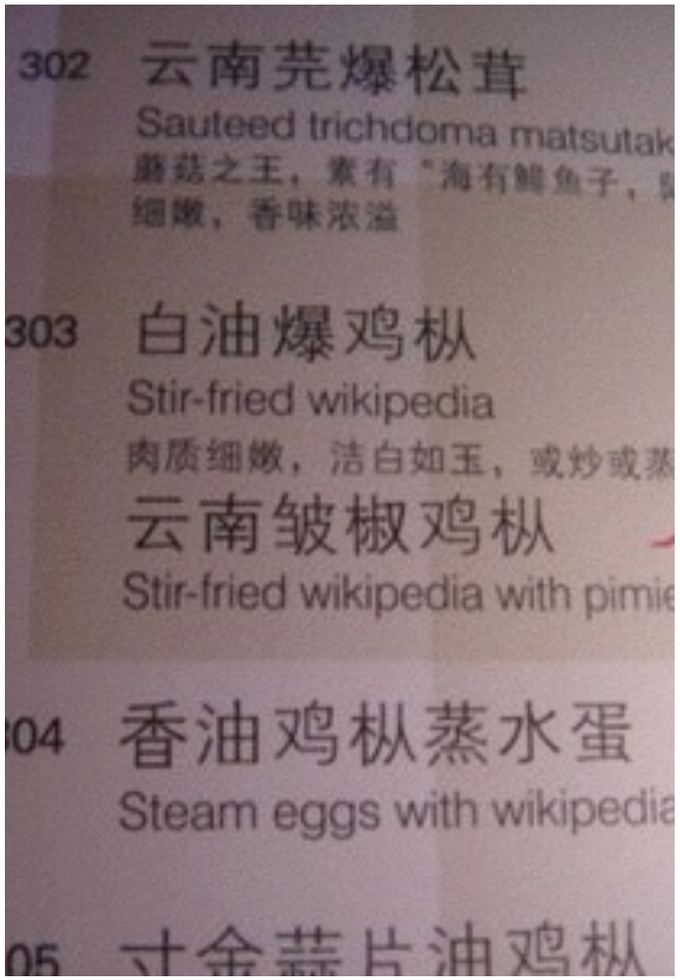
Linguistics & Corpora

Wednesday, January 21, 2015

Plan for Today:

- Language in 10 (Mai)
- Existing Corpora
- Hands on with Corpora



Language in Ten: Mai

Getting Language Data

Linguistic Data Consortium

- http://catalog.ldc.upenn.edu/

Dictionaries

- WordNet – 206K English words
- CELEX2 – 365K German words

Monolingual text

- Gigaword corpus
  - 4M documents (mostly news articles)
  - 1.7 trillion words
  - 11GB of data (4GB compressed)
- Enron e-mails: 517K e-mails

More Corpora

Monolingual text continued

- Twitter
- Chatroom
- Many non-English resources

Parallel data

- ~10M sentences of Chinese-English and Arabic-English
- Europarl: ~1.5M sentences English with 10 different languages

Annotated Corpora

Brown Corpus

- 1M words with part of speech tag

Penn Treebank

- 1M words with full parse trees annotated

Other treebanks

- Treebank refers to a corpus annotated with trees (usually syntactic)
- Chinese: 51K sentences
- Arabic: 145K words
- many other languages…
- BLIPP: 300M words (automatically annotated)

Other types of corpora

Many others…

- Spam and other text classification
- Google n-grams
  - 2006 (24GB compressed!)
  - 13M unigrams
  - 300M bigrams
  - ~1B 3,4 and 5-grams
- Speech
- Video (with transcripts)

Getting our own text data

Using your own laptop or one from CIS…

Practice downloading, parsing, and exploring data