CS 224N / Ling 237: Natural Language Processing
Corpora at Stanford
Spring 2004

1 Introduction
Every statistical method in NLP (and some non-statistical ones) requires data to train on, and more often than not, the training data is a linguistic corpus. In addition to their use in NLP, corpora are excellent resources for linguistic research. This is an introduction to a few of the corpora available at Stanford, and to some tools for using them.

One key place to find out more about the various corpora available at Stanford and the tools you can use to work with corpora is at the Corpora@Stanford website; there are many great links from this page:

http://www.stanford.edu/dept/linguistics/corpora

2 Types of Corpora
Stanford currently has a large selection of corpora available for students to work with. Not only are there many different kinds of corpora (which we'll go over below), but Stanford has many foreign language corpora, including corpora in Arabic, Chinese, German, Japanese, Korean, Portuguese, Spanish, and more.

2.1 Untagged corpora
This is generally your basic collection of text – basic, but still very useful. An example of such a corpus is the ICAME version of the Brown corpus [the whole Brown corpus has been tagged, but there are untagged versions lying around too]. Here's a snippet of the untagged Brown:

```
L01 0050 dirty West Side. It seemed incredible, as
I listened to the monotonous
L01 0060 drone of voices and smelled the fetid odors
coming from the patients,
L01 0070 that technically I was a ward of the state
of Illinois, going to
L01 0080 a hospital for the mentally ill. I suddenly
thought of Mary
```

2.2 SGML-marked corpora
Some corpora have SGML-style markup. This is useful for Information Retrieval-type tasks. The North American News Text Corpus is an example; it has fairly sparse markup, mostly pertaining to where the article appeared:

```
<DOC>
<DOCID> reute9501_001.0383 </DOCID>
<STORYID cat=i pri=r> a0034 </STORYID>
<FORMAT> &D3; &D1; </FORMAT>
<KEYWORD> BC-AUSTRALIA-KOALAS </KEYWORD>
<HEADER> reute 01-04 0247 </HEADER>
<SLUG> BC-AUSTRALIA-KOALAS </SLUG>
<HEADLINE>
Australian state moves to protect koalas from humans
</HEADLINE>
<TEXT>
<p>
SYDNEY, Australia (Reuter) - Australia's most populous state
Wednesday said it would stop urban development that has
endangered the habitat of the koala, now facing possible
extinction within 40 years as urban encroachment destroys
bushland.
<p>
New South Wales Planning Minister Robert Webster said a new
development policy effective from Feb. 13 would prevent local
governments approving developments until they determine whether
the proposed sites contain koalas.
```

```
<p>
```

The British News Corpus (BNC) is another SGML-marked corpus (it is tagged, too); its markup scheme is considerably more detailed.

2.3 POS-Tagged Corpora

Since statistical information on parts of speech in text is such a fundamentally useful kind of information, there has been quite a bit of work in putting out tagged corpora in recent years. A couple of the POS-tagged corpora available at Stanford include the Brown corpus and the British News Corpus. Here's a selection from the British News Corpus:

```
<s n="11"><w PNP>We <w VDB>do<w XX0>n't <w VHI>have
<unclear> <w PRP>from <w NP0>Hockerill <w NN1>school
<w VDB>do <w PNP>we<c PUN>?
```

A couple of things to keep in mind regarding POS-tagged corpora: different corpora may have subtly different tagsets, and this may affect the performance of taggers and other applications trained on the tagset. The BNC, for example, uses the Claws c5 tagset, which is generally coarser than the Brown tagset in adverbs, determiners, and pronouns, but makes a couple of distinctions that the Brown conflates, such as giving *of* its own part of speech. The Penn tagset is a strategically simplified version of the Brown: in the Penn, for example, the infinitive verb-marking and prepositional uses of *to* are conflated. Manning & Schütze section 4.3.2 has information on various English tagsets. Other languages, of course, can do without certain English tags, and may need other tags for things like classifiers that English speakers don't have to worry about.

2.4 Parsed Corpora

The gold standard for parsed corpora is the Penn Treebank, but now we have other parsed corpora such as ICE-GB and TIGERcorpus. This is a subset of the Brown corpus, plus pieces of the Wall Street Journal, ATIS, and Switchboard corpora, that have been hand-parsed. It looks like this:

```
( (S (S-SBJ (NP-SBJ *)
            (VP To
               (VP have
                    (NP the Greek paper))))
     (VP is not
          (NP-PRD (NP the great help)
                  (SBAR (WHNP-1 that)
                        (SBAR  (WHNP-1 that)
                               (S (PP-TMP at
                                          (NP first flush))
                                  (NP-SBJ it)
                                  (VP seemed
                                       (NP-PRD *T*-1))))))
     .))
```

This format takes some getting used to. It's good to know in advance that the Treebank, like any tagged or parsed corpus, is a bit idiosyncratic. The Treebank in general has a much flatter structure (= more daughters per mother) than the typical tree structure you'll see linguists write on the board. To some extent that's the difference between the toy sentences linguists come up with, but to some extent it's also the way the corpus was parsed – the rule seems to have been if in doubt, flatten it out!

2.5 Multilingual Corpora

Also coming into prominence are Multilingual corpora – that is, parallel texts in at least two languages. These corpora can be used for things like statistical machine translation, translation memory, or POS tagger development in new languages. Many uses are being found for these when one of the languages is well-analyzed with NLP techniques, and the other is relatively unknown – the knowledge of the well-analyzed language can be transferred over to the poorly-understood language. The multilingual corpora at Stanford

are the Hansards and the Hong Kong corpora – these are *aligned* corpora (that is, the mapping between sentences in the texts are given), and are French/English and Chinese/English respectively.

2.6 Transcripts

There are also many corpora available of transcribed speech. At Stanford these include the Air Traffic Control Corpus; the TDT2 corpus (American news broadcast transcripts); Verbmobil dialogues (German); and the Callhome transcripts, which are pretty interesting: transcripts of international phone calls from the US, available in several languages. Transcribed speech corpora are very useful for both linguistic research and speech-related NLP; a quick look at a bit of the Callhome corpus may illustrate the difference from written text:

```
203.79 209.22 B: I ha- b- the phone is I left the portable
phone upstairs last night so the battery ran out
209.42 209.90 A: okay
209.68 215.07 B: So I have I have the phone from the
living room in the bedroom with a s- cord stretched
across so I can lay on the bed of course
215.46 215.97 A: yeah
5
216.55 218.53 B: so that's why &Lee almost tripped
over the phone
218.52 219.78 A: laugh
219.11 220.58 B: oh it's only the second quarter
221.25 222.58 B: I thought it was the fourth quarter
222.41 224.63 A: yeah see that's what I'm watching
too the second quarter
225.14 226.14 B: oh bummer
```

Another interesting transcript corpus is the HCRC Maptask Corpus, which is a corpus of people finding their way around strange places carrying maps (presumably for direction-giving application development). It's SGML-marked by speaker and for "events" like laughter.

2.7 Speech Corpora

There are also many speech corpora, where the data files consist of recorded speech. At Stanford these include the Boston University Radio Speech Corpus, the Santa Barbara Corpus of Spoken American English, and more. With proper software you can get these to play back, and do interesting things like segment them.

3 Accessing Corpora at Stanford

The most widely-used corpora are all on AFS in the directory **`/afs/ir/data/linguistic-data/`**. Because there are many copyright issues associated with many of the corpora, the directory is protected. However, everyone in the class is automatically given access to all the corpora in the directory so go ahead and see what there is to offer. If you want to use corpora that are available at Stanford but are not on AFS, please check out the protocol listed on the Corpora@Stanford website for accessing these corpora.

4 Using Corpora – Some Simple Tools

- **grep**

  We have already used seen some use of grep in Ken Church's unix tutorial and homework #1. The **grep** family of Unix utilities is extremely useful when working with corpora. Suppose I want to find out something useful for an NLP application, like what kinds of adverbs come between auxilary verbs (which could be used, for example, to help disambiguate auxilary/noun combinations like *might* or *can*). I'll try it here on part of the Brown:

```
elaine7: /afs/ir/data/linguistic-data/ICAME/brown1 > egrep 'might [^ ]* have' *

   brown1 b.txt:B09 1350 would never take active part
   but might once have shown a tolerant
   brown1 c.txt:C16 0370 automobile racing as life in
   microcosm, one might reasonably have
   brown1 f.txt:F12 0270 might not have been really aware
   of his own mood; it had been latent,
   brown1 g.txt:G72 0470 Plan. The United States might
   well have exploited the opportunity
   brown1 h.txt:H24 1670 law, have the right to them and
   might actually have received the money.
   brown1 j.txt:J45 1320 The fact that AIA lists might
   not have been selected on a random
   brown1 j.txt:J62 0060 possible. This letter might
   not have been necessary had our efforts
   brown1 k.txt:K14 0120 mother and son alone in the universe.
   When might Mary have had that
   brown1 k.txt:K24 0600 of Pa's door down before he stopped.
   He might not have gone that
   brown1 l.txt:L09 0650 had been still hot, she might
   even have drunk some of it, she wouldn't
   brown1 l.txt:L14 1000 Payne luggage.) She might now
   have taken it away again. Motive- her
   brown1 l.txt:L15 1210 might well have been included,
   he felt. Mrs& Meeker had
```

and I get a sample of inter-modal adverbs, and even a noun.

- **tgrep**

As you can imagine, searching a parsed Treebank corpus is a bit more complicated than your average grep. But there is a great utility called tgrep that you can use for Treebank searches. Actually, the most up-to-date version of tgrep is now **tgrep2.**

To use tgrep2, you need to be working on a firebird or raptor computer – this doesn't work on the elaines! If you want to be able to use the command 'tgrep2' without typing its full path name each time you need to set your PATH variable by typing

```
setenv PATH /afs/ir/data/linguistic-data/bin/linux_2_4:$PATH
```

The corpora that are available in 'tgrep-able' format are versions of the Brown, Switchboard, WSJ, NEGRA, and Chinese Treebank corpora. These corpora are in

```
/afs/ir/data/linguistic-data/Treebank/tgrep2able
```

You can also find a manual for tgrep2 in this directory.

Here is a simple example of using tgrep2:

```
tgrep2 -c wsj_mrg.t2c.gz 'NP < VP'
```

This command looks for all instances of an NP that dominates a VP in the corpus 'wsj_mrg.t2c.gz' There is a lot more information on the syntax of tgrep2 and how to set it up and use it on the Corpora@Stanford website.

- **more!**

There are many more corpus tools available for all types of corpora. Stanford has software and scripts for playing with corpora and all of these tools are listed on the Corpora@Stanford website. Check it out!

5 Useful Links

- Corpora@Stanford website at
  `http://www.stanford.edu/dept/linguistics/corpora`
  has information about the corpora available at Stanford as well as information about many different useful corpora tools.
- Chapter 4 of Manning & Schütze is a good reference for information about corpora.
- Chris Manning's **Corpora** section on his website
  `http://www-nlp.stanford.edu/links/statnlp.html`
  is another good place to start looking for any specific corpus tool you might want.