Corpora

What corpora do we have and where?

General location

Unpack into /data3/data on the cluster. Use the following permissions to make them nicely readable but difficult to accidentally overwrite (unless you run stuff as root, which can overwrite anything):

```
chown -R nobody:nobody YOURCORPUSDIR

find YOURCORPUSDIR -type d -exec chmod 755 {} \;

find YOURCORPUSDIR -type f -exec chmod 644 {} \;
```

Specific corpora

Google Web1T

Cluster: /data3/data/web1t

DVDs: ??? (Robert?)

Wall Street Journal

Cluster: /data3/data/BLLIP99

CDs: ???

Workshop on Statistical Machine Translation

Cluster: /data3/data/wmt[08|09|10]

Penn Treebank 3

Cluster: /data3/data/Penn3

British National Corpus

Cluster: /data3/data/BNC-world

TIGER corpus

Cluster: /data3/data/tigercorpus1