

Using Corpora at IU

L615

Spring 2009

1 Server

1. I'll give you all accounts on the “jones” server housed in the Linguistics Department here at IU.
 - The full name is `jones.ling.indiana.edu`
2. Permissions & copyrights
 - You will generally not have permission to add or change the files in the `Corpora` directory (see below)
 - According to the copyrights for corpora, you are also not permitted to download the corpora onto your local machines ... See section 3 for ways to handle this.
3. Using `ssh` is a major way to connect to `jones`
 - For much of what you want to do, this will likely be sufficient.
 - But what if you want to run a program on your (local) computer on a corpus? See section 3.
4. Using `unix`
 - We'll walk through the slides at: <http://jones.ling.indiana.edu/~cl/unix.pdf>

2 Corpora directory

1. Basic structure
 - Location: `/Volumes/Data/Corpora/`
 - Subdirectories are generally organized by language, using 2-letter language codes (see <http://www.iana.org/assignments/language-subtag-registry>)
2. More thorough information found at: <http://jones.ling.indiana.edu/wiki/JonesCorpora>

3 Processing corpora

1. The need for Perl (or something similar)
 - Having a big corpus file is nice, but what do we do with it? How do we navigate through it?
 - You can run it through pre-built software ...
 - and/or: you can write your own programs to extract the information you want.
 - Knowing how to write scripts in a language like Perl will improve your corpus-navigating capabilities
2. Handling different character sets: on a terminal, you may need to add these lines to your `.profile` file:

```
# for better unicode support:
export LC_ALL=en_US.UTF-8
export LC_CTYPE=en_US.UTF-8
export LANG=en_US.UTF-8
```

3. Mapping a network drive
 - To map a network drive means that you make a directory on a remote server available *locally*.
 - So, you can, e.g., run software on your machine on a corpus on jones *as if* it were on your own computer

For the jones corpora, you can do this from a mac (available for student use in MM401, if you don't have your own):

- Finder → Go → Connect to Server
- Server address: `afp://jones.ling.indiana.edu/Corpora`
- Enter your `jones` username and password
- The location of the corpora drive on your computer is now: `/Volumes/Corpora/`
- When you're done, "eject" the drive

4 Editing files

The important point for editing files is that you save them as text files. I'm not going to provide much in the way of editing support this semester: remember that google is your friend (as are other, experienced students)

1. Editing locally
 - Some editors you might find useful:

- Notepad/WordPad
- TextEdit
- Aquamacs/Emacs
- Vim
- You can also look for Perl IDEs (Integrated Development Environments): these usually have nice debugging features, do text coloring, etc.
- Be sure that you are writing *plain text* files (and not, e.g., rich text)
- Unix and DOS files end lines differently. If you upload a file onto jones and have strange errors, you might want to make sure it's properly converted to a unix format. See: <http://kb.iu.edu/data/acux.html>

2. Editing on jones directly: you can use `pico`, `emacs`, or `vi`

5 Web corpora and more

You'll often find web interfaces to corpora, which allow for searching (e.g., the different corpus search interfaces Mark Davies has made available, <http://davies-linguistics.byu.edu/personal/>)

1. Pros: nice interfaces, useful for generating basic statistics, etc. access to corpora you might not otherwise have access to
2. Cons: You're limited by what the designer wants you to search for; you can't run a tagger or parser on it; etc.

Final note: More information is generally available from the `corpora-list`: <http://gandalf.aksis.uib.no/corpora/>

- Be sure to search their archives and possibly ask me before posting a question.
- Another good website is: <http://devoted.to/corpora>