Logging on to UGA corpus server (for both on-campus and off-campus)

**For access to the corpus sever, contact:** linglab@uga.edu

I. **On-campus**

1. If you are on campus, you can open terminal or PuTTY. You will NOT need to use VPN in order to connect to the UGA corpus sever.
   a. For windows users, you will need to install PuTTY: [https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html](https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html). Mac users use terminal, which is under “utilities” in “Applications”.
   b. For both Windows and Mac users: You will need to check to see if accented characters are displayed properly.
   c. For PuTTY: Under translation set the remote character set to (Latin-1, West Europe).
   d. For Mac users: Go to Preferences>Profiles>Advanced>International. Make sure to set encoding to “western (ISO Latin 1)” *depending on MacOS you may have to do Preferences>Settings>Advanced>International.
2. Your login credentials for the server are the same as your UGA MyID login credentials.
   a. Type ssh username@corpus.uga.edu OR username@kucera.uga.edu
   b. Then type in your password.

```
Last login: Sat Feb 16 09:08:42 on tty5000
William's MacBook Pro:~ arielle$ ssh afl18818@corpus.uga.edu
afl18818@corpus.uga.edu's password: 
```

NOTE: For both Mac and Windows users, the first time you access the server, you will be asked if you are sure that you want to connect. Enter “Yes”.

3. Once logged on, you will see [username@kucera ~]$
4. To access the corpus sever, enter “cqp -eC;”

5. Cqp should look like the photo above.

6. There are two ways to access a corpus
   a. Type “show corpora;” to show the list of corpora that are available on the sever and then type “corpusname;”, for example “FRWAC;”.  
   b. Or just type the name of the corpus that you wish to use.

II. Off-campus

1. If you try to use terminal or PuTTY without the VPN connection, a message, like the one below will appear:

2. VPN is required in order connected to UGA corpus sever remotely.
   a. Cisco AnyConnect VPN from EITS will need to be downloaded in order to connect to UGA remotely.  
      (https://eits.uga.edu/access_and_security/infosec/tools/vpn/)
   b. Click download Cisco AnyConnect VPN.
   c. Once downloaded, Open Cisco AnyConnect. Type in “remote.uga.edu”
VPN:
Ready to connect.
remote.uga.edu
Connect

Web Security:
No License Key.

System Scan:
No policy server detected.
Default network access is in effect.

Roaming Security:
You are not currently protected by Umbrella.
Profile is missing.

AMP Enabler:
Waiting for configuration...
d. DOU is required as well as UGA MyID. Username and Password are your MyID credentials and “Second Username” is for DOU. “push”, “sms”, or “phone” can be used in order to log-in.

You will need ArchPass, powered by Duo, to complete the log-in process. Enroll at archpass.uga.edu/enroll.

Please enter your MyID and password.

Group: 01 Default
Username: afl18838
Password: ************
Second Password: ****

3. Once UGA is accessed remotely, you will be able to access the corpus as if you are on campus.
Corpus resources at UGA

The Linguistics Lab maintains a collection of natural language corpora. Some highlights are listed below. Many of these have been released through the Linguistic Data Consortium (LDC). UGA’s membership in the LDC is funded annually by contributions from cooperating academic units such as Departments, Colleges and Institutes. Faculty and students in these cooperating units enjoy unrestricted access to the entire collection.

For information about becoming a cooperating academic unit, please contact John Hale via <linglab@uga.edu>

The Brown Corpus
This was the first million-word electronic corpus of English, created in 1961 at Brown University. It spans about fifteen different categories of text.

The Penn Treebank
Manually-corrected phrase structure trees for English, including 1.2 million words of newspaper text from the Wall Street Journal.

COCA: Corpus Of Contemporary American English
531 million tokens of American English sampled from 1990--2017 across categories such as Academic, Fiction, Magazine, Newspaper, and Spoken. This version is annotated with lemmas and parts of speech. Provided courtesy of the UGA Library.

COHA: Corpus Of Historical American English
400+ million words from the period 1810--2008 facilitate diachronic investigation. Provided courtesy of the UGA Library.

British National Corpus
100 million words of text from a variety of British sources that are annotated with parts of speech and lemmas as well as sociolinguistic variables such as speaker age, social class and geographical region. 91% was published between 1985 and 1993.

AudioBNC
Audio and all available transcriptions of the 7.5 million words of the spoken portion of the British National Corpus, including TextGrids aligned at the word and phone level, with associated speaker metadata. For more details see Coleman et al. 2012.

SpokenBNC 2014
11.4 million tokens, orthographically transcribed from smartphone recordings made between 2012 and 2016. Substantial speaker metadata is included. Annotations include parts of speech, lemmas and a system of semantic tags. For more details see Brezina et al 2018.
**Arabic Treebank**
Approximately 800 thousand words of newswire text from Agence France-Presse annotated with parts of speech, morphology and phrase structure.

**DEFT Spanish Treebank**
About 100 thousand words from both Spanish newswire and discussion forums, with extensive morphological and syntactic annotations.

**CETEMPúblico**
180 million words from the Portuguese newspaper "Publico" 1991--1998 with morphological and syntactic annotations.

**French Treebank**
This corpus is drawn from the newspaper Le Monde 1989-1994 annotated with syntactic constituents, syntactic categories, lemmas and compounds and totals about 650 thousand words.

**SPMRL2014**
Dependency, constituency and morphology annotations for Arabic, Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish.

**NEGRA corpus**
355 thousand tokens of the German newspaper Frankfurter Rundschau annotated with syntactic structures.

**EuroParl**
About 40 million words of European Parliamentary proceedings aligned across translations into English, German, Spanish, French, Italian and Dutch.

**CALLHOME**
This corpus consists of 5-10 minute snippets from 120 phone calls, each 30 minutes each in length. They are made from native English speakers from various places in North America; mostly made to family members or close friends. Roughly 90 of the phone calls are placed to persons living outside of North America but all are in English. Holdings include both audio and transcripts.

**The Buckeye Speech Corpus**
This corpus is comprised of 40 speakers from Columbus Ohio, recorded from 1992-2000. Each interview lasted about 60 minutes and the corpus totals more than 300,000 words of speech. Created primarily for studying phonological variation in American English, it was gathered via a "modified sociolinguistic interview format" to target a representative sample of forms and frequency of phonological variants. The corpus' .wav files are transcribed and force-aligned at the segment level.
CELEX2
Orthography, phonology, morphology and attestation frequency information for words in English, German and Dutch.

Concretely Annotated New York Times

WaCky corpora
Between 1.2 and 1.9 billion tokens each of French, German and Italian as crawled from the world wide web. Also includes about 800 million tokens of English Wikipedia as it was in 2009. These corpora are annotated with lemmas and parts of speech. For more details see Baroni et al 2009.