

# Tools for (Field) Linguists

Gina

April 30, 2011

Concordia

## Mission:

Learn a few key words and a bit of programming so you can be resourceful, and save time (and pull in the big grants).

## Advice:

- Reduce, reuse, recycle,
- Avoid toy world projects

# *Roadmap*

- Git
- Wiki
- GATE
- Groovy

# *Git - 10mins*

- Create a GitHub account
  - <http://github.com/>
- Fork ToolsForFieldLinguistics
  - <https://github.com/cesine/ToolsForFieldLinguistics>
- Clone it onto your computer (if you dont have Git see next slide)
  - `git clone git://github.com/cesine/ToolsForFieldLinguistics.git`
- Look at the files, what is in where?
  - Src?
  - Gen?
  - Doc?
  - Readme?

# *Get Git*

- Mac SnowLeopard
  - <http://help.github.com/mac-set-up-git/>
- Ubuntu
  - <http://help.github.com/linux-set-up-git/>
- Windows:
  - <http://help.github.com/win-set-up-git/>

# *Git 30mins*

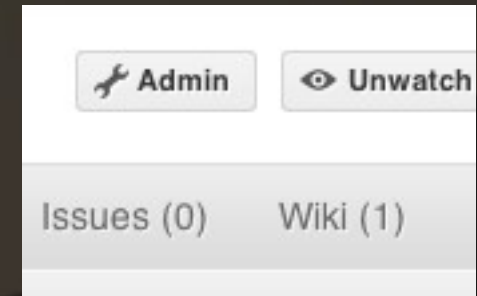
- Now that you know what a repository looks like:
  - Create a new repository in your account (for the mini project you will do today)
  - Follow the instructions to create the repository on your computer and push it to GitHub

A rectangular button with a dark background and a light border, containing the text "New Repository" in a light-colored font.

New Repository

# Wiki 10mins

- Create a wiki for your mini project
- Write a sequence of steps to get you from your data to your goal



h1 h2 h3 🔗 📄 📁 **B** *i* { } ☰ ☷ “ ” HR ?

Edit Mode: **Markdown**

<b>Wiki-Specific Syntax</b>	<b>Links</b>	<p>To create links to other wiki pages, simply wrap the page's name in <code>[[ (square brackets.)</code> For example, <code>[[My Page]]</code> will link to the wiki page titled "My Page".</p> <p>You can also use this syntax to make quick external links; for example, <code>[[http://google.com/]]</code> will make a link <a href="http://google.com/">http://google.com/</a>.</p>
<a href="#">Block Elements</a> <a href="#">Span Elements</a> <a href="#">Miscellaneous</a>	<a href="#">File Links</a> <a href="#">Images</a> <a href="#">Escaping Tags</a> <a href="#">Code Blocks</a> <a href="#">Mathematics</a>	

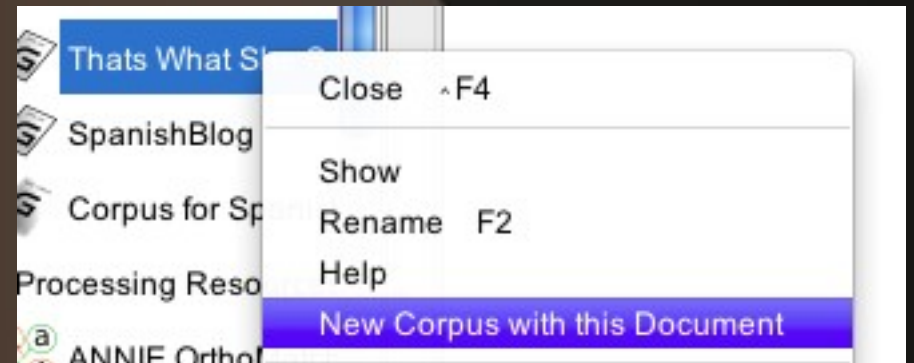
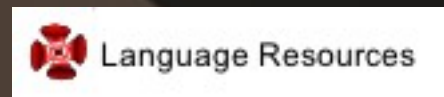
# *GATE*


- Open Source
- Very complex
- Lots of things to do and try
- We will only use GATE to highlight words in context, and to run Groovy scripts on our corpus



# *GATE 30mins*

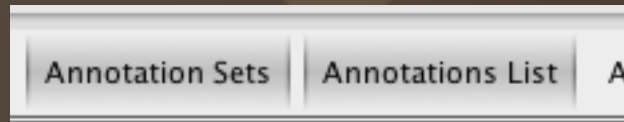
- Lets see what the built-in ANNIE pipeline can do on English text <http://www.cs.washington.edu/homes/brun/pubs/pubs/Kiddon11.pdf>
- Create a new Language Resource
- Create a new Corpus with that document



- Click on the Green Flower 
  - It will load all the processing resources for ANNIE and the Annie pipeline application

# *GATE cont*

- Run the application
- Look around at the information GATE tagged for you
- Notice the Tokens, that's where noun/verb information is stored


A screenshot of the GATE application interface. On the left, a text document is open, showing a paragraph about humor identification. Below the text is a table of annotations. In the center, a 'Token' editor window is open, showing fields for category, kind, length, orth, and string. On the right, a list of annotation types is visible, with 'Token' selected.

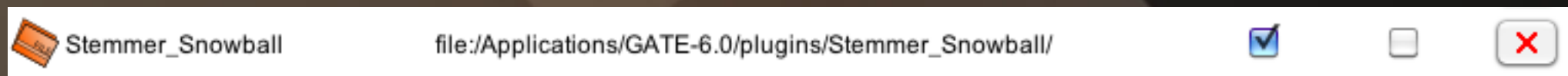
Type	Set	Start	End	Id	Feat
Token		343	350	139	{cate
Token		350	351	140	{cate
Token		352	354	142	{cate
Token		355	363	144	{cate

category	kind	length	orth	string
VBP	word	8	lowercase	identify

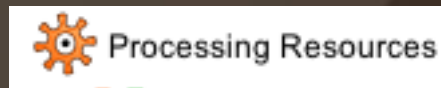
- ☐ Percent
- ☐ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Temp
- ☒ Token
- ☐ Unknown
- ☐ UrlPre
- ☐ Original markups

# *GATE cont*

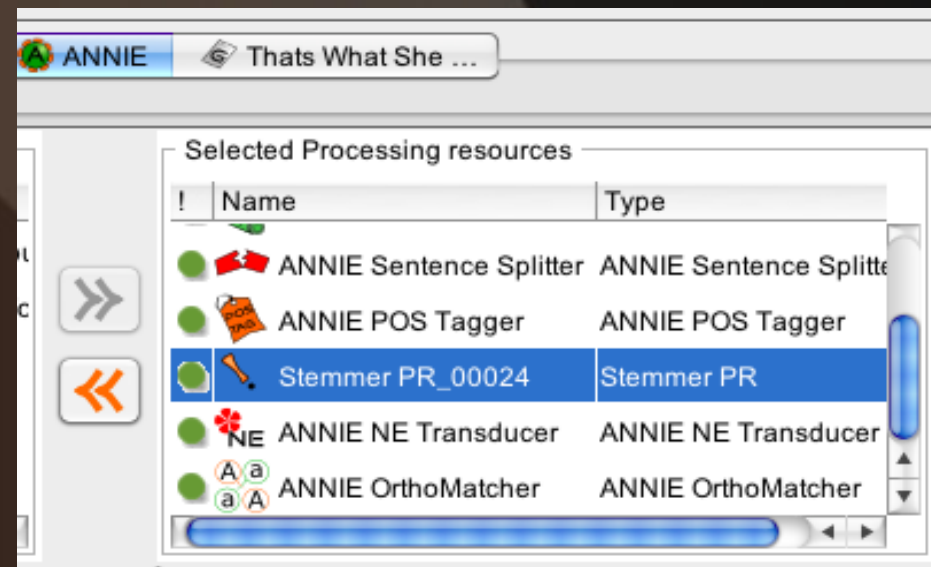
- Add a Snowball stemmer into the mix
  - Load the plugins 



- Load the processing resource (Stemmer)



- Put it into the Pipeline
- Run it



# *GATE 5mins*

- Create a new Language Resource with the
  - URL : <http://blogworkorange.blogspot.com/>
  - Encoding: utf-8
- Right Click on the Document
  - Create a corpus with this document
- Create a new Processing Resource
  - Document reset
  - GATE Unicode Tokenizer
- Create a new Application
  - Put the Document Reset and Unicode Tokenizer to the right pane
  - Run this application

The background consists of several overlapping, semi-transparent shapes in various shades of brown and dark grey. A large, light brown shape is on the left, a medium brown shape is in the center, and a dark grey shape is on the right. The word "Lunch" is centered in the medium brown shape.

Lunch

# *Groovy*

- Click on the black P to load plugins
  - Check “Load now” for Goovy
  - Click OK
- Click on the Tools menu
  - Groovy Tools
    - Groovy Console
- Now you're read to write some Groovy!

# *The End Result*

- Open  
ToolsForFieldLinguistics/src/com/fieldlinguist/groovyInGate/ExtractWordsOrderBySuffix.groovy
- Run it (CTRL+R)
- If it worked, your output will be in your GATE folder
  - Words\_function\_vs\_content.html
    - To see what this looks like, copy it into the  
ToolsForFieldLinguistics/src/com/fieldlinguist/javascript/tabletobarpielinegraph, then open it in a browser
  - Words\_function\_vs\_content.txt
  - Words\_to\_look\_for\_suffixes.txt
  - Words\_function\_vs\_context.jape
    - To see what this does load it as a “Jape Transducer” in your Processing resources and put it at the end of your pipeline



# *Groovy Practice – Maps, loops, output, regex 10mins-1hr*

- Go to your github fork so you can see the visual commit history
- Use tutorial.groovy to build your own spanish “Morpheme Finder”
- Start with commit 1, work your way until the final commit



# *Groovy Practice – Teams of 2*

## *1hr*

- As a team of two, grab some data (you can use one of the corpus in [cesine/CorporaForFieldLinguistics](#) or build your own from a blog, rss feed, twitter, or [opensubtitles.org](#))
- Decide on a Research Question that you want to investigate
  - How can you load that text into GATE
  - What kind of rules would you like to write to highlight the key data?
  - How can you write some Groovy to find what you want?