



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
- Iterative development

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

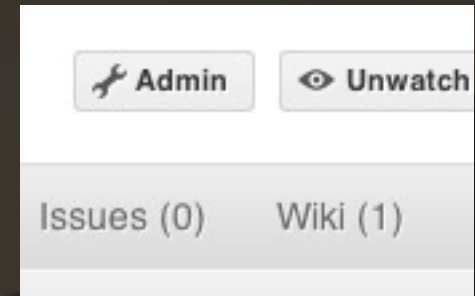
- Now that you know what a repository looks like:
 - Create a new repository in your account (for the mini project you will do today)
 - Follow the instructions to create the repository on your computer and push it to GitHub

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**

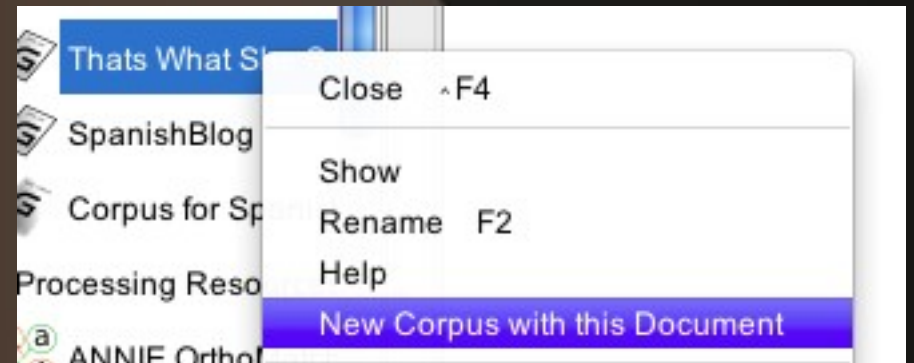
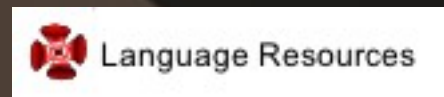
Wiki-Specific Syntax	Links	<p>To create links to other wiki pages, simply wrap the page's name in <code>[[(square brackets.)</code> For example, <code>[[My Page]]</code> will link to the wiki page titled "My Page".</p> <p>You can also use this syntax to make quick external links; for example, <code>[[http://google.com/]]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

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

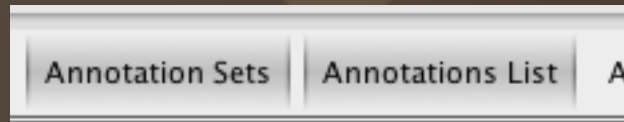
- Lets see what the built-in ANNIE pipeline can do on English text <http://www.cs.washington.edu/homes/brun/pubs/pubs/Kiddon11.pdf>
- Create a new Language Resource
- Create a new Corpus with that document



- 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 software interface. The main window displays a text document with the following content: "Seattle WA 98195-2350", "fchloe,brung@cs.washington.edu", "Abstract", "Humor identification is a hard natural language understanding problem. We", "a subproblem — the 'that's what's", "problem — with two distinguishing". Below the text is a table of annotations. A pop-up window titled 'Token' is open, showing a list of token features and their values. To the right of the main window is a list of annotation sets with checkboxes.

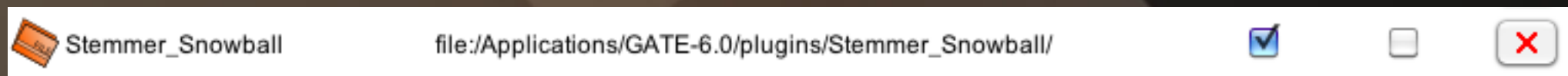
Type	Set	Start	End	Id	Feat
Token		343	350	139	{cate
Token		350	351	140	{cate
Token		352	354	142	{cate
Token		355	363	144	{cate

category	kind	length	orth	string
VBP	word	8	lowercase	identify

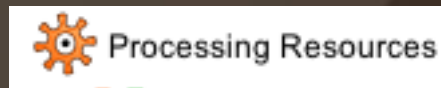
- ☐ Percent
- ☐ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Temp
- ☒ Token
- ☐ Unknown
- ☐ UrlPre
- ☐ Original markups

GATE cont

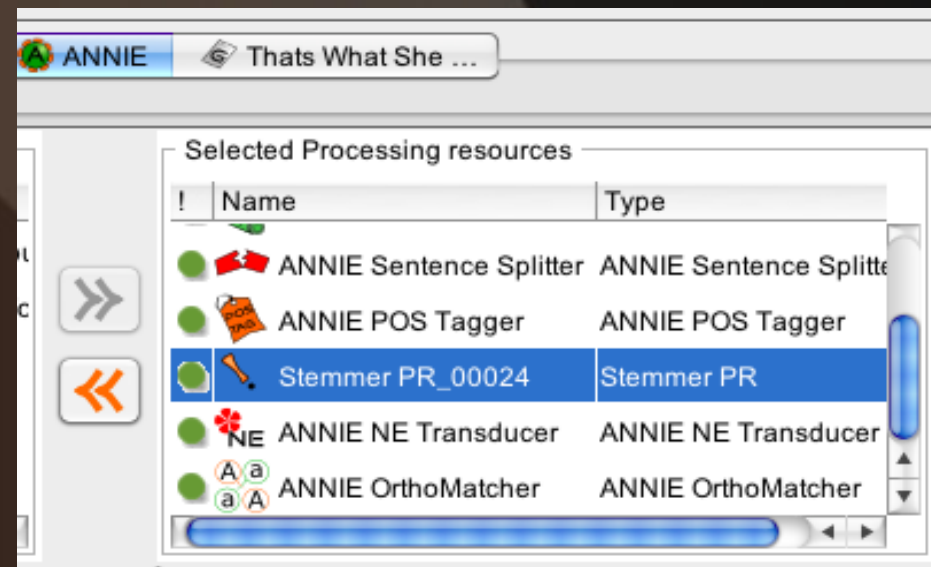
- Add a Snowball stemmer into the mix
 - Load the plugins 



- 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, overlapping with a darker brown shape in the center. To the right, there is a dark grey shape that overlaps with the central brown one. The overall effect is a layered, organic composition.

Lunch

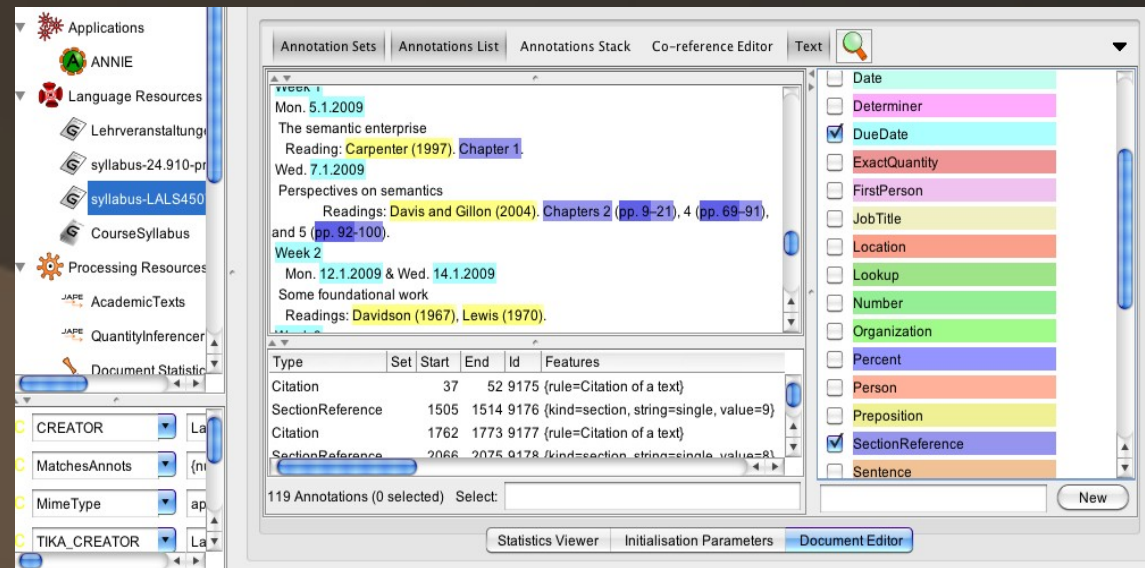
Groovy

- 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

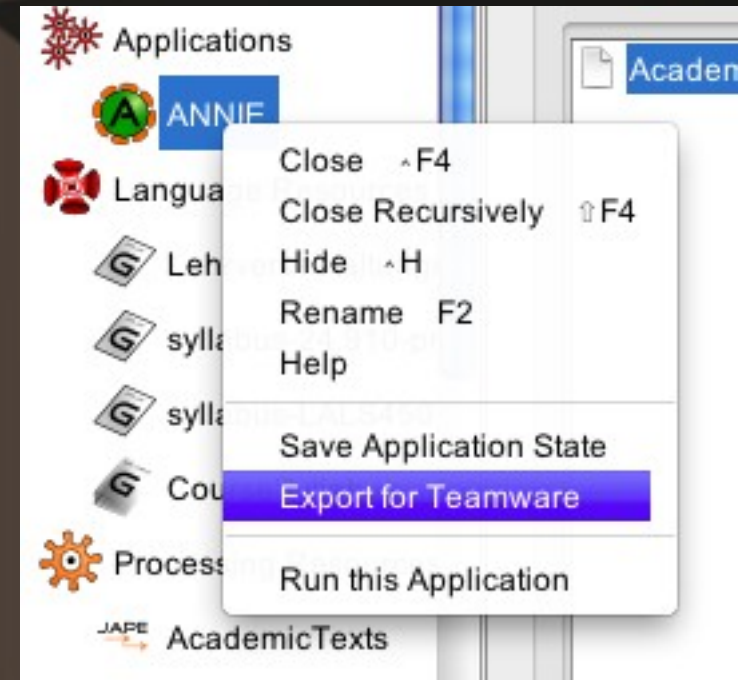
Jape Grammars



- As linguists, your favourite part of GATE will be Jape
- Jape are essentially rules that let you colour code the words in your text so you can find your morphemes in context.
 - Reading lists example of an xgapp:
<https://github.com/cesine/GATEinSpring/tree/master/gate/WEB-INF/gate-files>

Saving your GATE application

- Its easy to save and send your GATE application to your colleagues
- Right click on the Application and choose
 - Export for Teamware
- This creates a zip with everything inside



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 the first commit, work your way until the final commit
- Create your own groovy script in the same directory
 - myTutorialScript.groovy
- Build your script, committing as you go to match the sample tutorial.groovy

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?