Finding the right words in trees and the right trees in a forest

An introduction to *tgrep2* and *TIGERSearch*

Florian Jaeger & Roger Levy
tiflo@stanford.edu & rog@stanford.edu

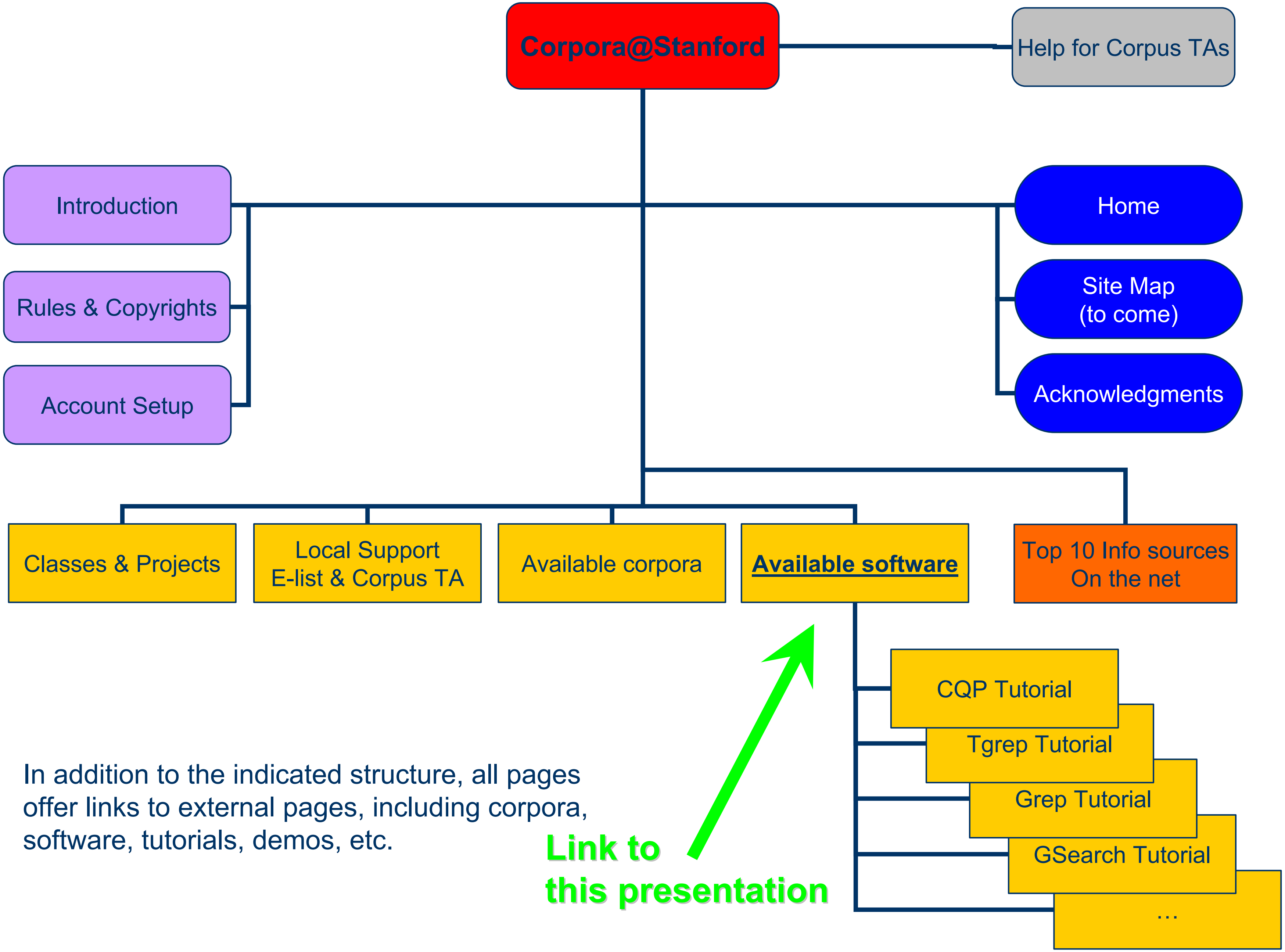
February 6th, 2004

Where will you be able to find this presentation?

- Corpora@Stanford

  http://www.stanford.edu/dept/linguistics/corpora/

  - Goto *Corpus-tools & other useful software*
  - Goto *Locally available corpus-software*
  - There will be links to this presentation in the tgrep and TIGERSearch tutorials, links to which are given in the table



Corpora@Stanford

Help for Corpus TAs

Introduction

Rules & Copyrights

Account Setup

Home

Site Map
(to come)

Acknowledgments

Classes & Projects

Local Support
E-list & Corpus TA

Available corpora

Available software

Top 10 Info sources
On the net

CQP Tutorial

Tgrep Tutorial

Grep Tutorial

GSearch Tutorial

…

In addition to the indicated structure, all pages offer links to external pages, including corpora, software, tutorials, demos, etc.

**Link to
this presentation**

Formats, compatible tools, and where to find them

- Currently we have five different search tools for syntactically annotated corpora
  - **Corpussearch** (needs a grammar fragment)
  - **tgrep & tgrep2** (need specific format)
  - **TIGERSearch** (needs specific format)
  - **Linguist's Search Engine** (doesn't need either, but is slow)
  - **Roger's tgrep – coming soon**

- Syntactically corpora come in their own annotation format (PENN, Negra, TigerXML, etc.)
  - The original corpus files are stored with each corpus on AFS.
  - If you want to use tgrep, tgrep2, or TIGERSearch, the corpora have to be converted into the right format.
  - Here is where you can find which corpus for which search tool (if the corpus you need is not in the right format, ask the corpus TA to convert it for you).

Syntactically annotated corpora of (more or less) contemporary English

- **Parsed Brown corpus, release 3 (~30,000 sentences)**
  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/brown.t2c.gz
  tgrep: /afs/ir.stanford.edu/data/linguistic-data/tgrepable/
  TIGERSearch: on the corpus computer, D:\
- **Parsed Switchboard corpus, (~50,000 sentences [sp])**
  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/swbd.t2c.gz
  tgrep: /afs/ir.stanford.edu/data/linguistic-data/tgrepable/
  TIGERSearch: on the corpus computer, D:\
- **Parsed Penn Wall Street Journal corpus, release 3 (~50,000 sentences, ~1 million words)**
  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/wsj_mrg.t2c.gz
  tgrep: /afs/ir.stanford.edu/data/linguistic-data/tgrepable/
  TIGERSearch: on the corpus computer, D:\
- **Parsed ICE-GB corpus (~84,000 sentences [sp & wr])**
  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/icegb.t2c.gz

Syntactically annotated corpora of historical English

- **The York-Toronto-Helsinki Parsed Corpus of Old English Prose (~110,000 sentences; 1.5 million words)**

  TIGERSearch: /afs/ir.stanford.edu/data/linguistic-data/YCOE/TigerXML/

- **Helsinki Parsed Corpus of Middle English (>100,000 sentences; 1.3 million words)**

  Available but not yet converted

Syntactically annotated corpora of German

- **Parsed NEGRA corpus, v2 (German, ~ 200,000 sentences)**

  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/negra.t2c.gz

  TIGERSearch: on the corpus computer, D:\

- **Parsed TIGER corpus (German, ~40,000 sentences; 700,000 words)**

  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/tiger.t2c.gz

  TIGERSearch: on the corpus computer, D:\

Syntactically annotated corpora of other languages

- **Parsed Penn Chinese Treebank corpus, (~50,000 sentences)**

  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/chtb2.t2c.gz

  TIGERSearch: on the corpus computer, D:\

- **Parsed Arabic Treebank Part 1 - v2.0 (734 stories representing 140,265 words)**

  tgrep2: /afs/ir.stanford.edu/data/linguistic-data/tgrep2able/arabic-treebank-without-vowels.t2c.gz

- **Prague Dependency Treebank v1.0 (1.8 million words)**

  Available but not yet converted

- **Samples of: Korean, Chinese, and many other corpora come with TIGERSearch 2.1**

Where to find the available search tools

- Tgrep, tgrep2, and TIGERSearch are installed on AFS

  /afs/ir.stanford.edu/data/linguistic-data/bin/sun4x_57/tgrep

  /afs/ir/data/linguistic-data/bin/linux_2_4/tgrep2

  /afs/ir/data/linguistic-data/TIGERSearch/bin/TIGERSearch

  - Instructions on how to set up your account to use tgrep:

    http://www.stanford.edu/dept/linguistics/corpora/cas-tut-tgrep.html#TGrep2

    http://www.stanford.edu/dept/linguistics/corpora/cas-tut-tgrep.html#TGrep

- TIGERSearch is also installed on the corpus computer
  - See also downloadable installation files on AFS for Linux, Unix, and Windows:

    /afs/ir.stanford.edu/data/linguistic-data/downloadables/

Tgrep2 vs. TIGERSearch (1)
Shared features

- Search power (almost the same?)
  - full regular expression power within nodes,
  - Boolean expressions over nodes
  - Multiple simultaneous inter-node constraints on one node
  - Syntactic relations: dominance, siblings, lin. Precedence
  - Underspecification: e.g. direct vs. indirect dominance
  - POS & category searches,
  - Word searches
- Output of searches can be formatted in various ways
- Detailed help and manuals
- Full Penn Treebank support
- Work on zipped files (saves space)

Tgrep2 vs. TIGERSearch (2) Features not shared

- Slightly different query languages
  - **Tgrep:** less predefined structure – basically regular expression on nodes plus a syntax to combine those nodes.
  - **TIGERSearch:** more predefined structure, so that - prior to everything else - the search expression involve more typing.
- TIGERSearch has a GUI (more comfort, longer loading time)

Tgrep2 vs. TIGERSearch (3) Advantages of TIGERSearch

- Platform independent – runs on Windows, Linux, Unix, Apple
- Queries can be typed or drawn
- Categories and POS available in the selected corpus are displayed in drop down menu while typing
- Online query syntax check and suggestions during query construction
- Bookmarks to searches (and the results) can be stored
- Template definition (shortcut for searches used over and over again)
- UNICODE fonts allow you to read the corpus “as is” even for non-latin alphabets (e.g. for Chinese, Korean)
- Trees browseable in GUI
- Graphical output of searches (trees) can be saved in many formats (SVG, TIF, JPG, XML etc).
- Includes easy handling of secondary and primary edge labels in search (e.g. that a PP is a ‘dative’ or a ‘directional PP’)
- TIGERSearch seems to be faster (not tested)
- Works on trees with discontinues constituents.

An introduction to Tgrep2

- TGrep2
  - Tutorial
  - Basic unit for searches
    - A node: e.g. ‘V*’
  - Examples:
    - tgrep2 -c wsj_mrg.t2c.gz -l 'VP < (NP $. NP)‘
    - tgrep2 -c wsj_mrg.t2c.gz -l 'VP < (NP $. PP-DTV)‘
    - tgrep2 -c wsj_mrg.t2c.gz -l 'VP=foo < (/VB*/ < gave) & < (NP $ NP)‘
    - tgrep2 -c wsj_mrg.t2c.gz -l 'VP=foo < (/VB*/ < gave) & < (NP $ PP-DTV)'

Tgrep2 syntax

- A < B A is the parent of (immediately dominates) B.
- A > B A is the child of B.
- A <N B B is the Nth child of A (the first child is <1).
- A >N B A is the Nth child of B (the first child is >1).
- A <, B Synonymous with A <1 B.
- A >, B Synonymous with A >1 B.
- A <-N B B is the Nth-to-last child of A (the last child is <-1).
- A >-N B A is the Nth-to-last child of B (the last child is >-1).
- A <- B B is the last child of A (synonymous with A <-1 B).
- A >- B A is the last child of B (synonymous with A >-1 B).
- A <` B B is the last child of A (also synonymous with A <-1 B).
- A >` B A is the last child of B (also synonymous with A >-1 B).
- A <: B B is the only child of A
- A >: B A is the only child of B
- A << B A dominates B (A is an ancestor of B).

TGrep2 syntax

- A >> B A is dominated by B (A is a descendant of B).
- A <<, B B is a left-most descendant of A.
- A >>, B A is a left-most descendant of B.
- A <<` B B is a right-most descendant of A.
- A >>` B A is a right-most descendant of B.
- A <<: B There is a single path of descent from A and B is on it.
- A >>: B There is a single path of descent from B and A is on it.
- A . B A immediately precedes B.
- A , B A immediately follows B.
- A .. B A precedes B.
- A ,, B A follows B.
- A $ B A is a sister of B (and A 6= B).
- A $. B A is a sister of and immediately precedes B.
- A $, B A is a sister of and immediately follows B.
- A $.. B A is a sister of and precedes B.
- A $,, B A is a sister of and follows B.
- A = B The node matched by A is also matched by B.

TIGERSearch 2.1 syntax

- The TIGERSearch query language tutorial
- Query language quick references
- Basic unit for searches
    - A node: e.g. ‘[pos=‘..’ cat=‘..’ word=‘..’ lemma=‘..’]’
- Examples:
    - ([cat="VP"] > #n1:[cat="NP"]) & (#n1 $ [cat="NP"])
    - ([cat="VP"] > #n1:[cat="NP"]) & (#n1 $ [cat=“PP"])
    - (#v1:[cat="VP"] > [cat="NP"]) & (#v1 >DTV [cat="PP"])

Finally:
Little buggers hiding in trees

- TIGERSearch and tgrep/tgrep2 use exactly the opposite syntax for dominance:
  - Tgrep: A < B ‘A dominates B’
  - TIGERSearch: A > B ‘A dominates B’

- Tgrep is right-headed!
  - The following pattern matches an S which has a child A and another child that is a C and that the A has a child B:

    S < (A < B) < C
  - However, this pattern means that S has child A and that A has children B and C:

    S < ((A < B) < C) equivalent to: S < (A < B < C)

Potential problems (2)

- Display of non-Latin characters (e.g. with the Chinese, Arabic, and Korean treebanks)

- Your questions & comments