UiT

NORGES
ARKTISKE
UNIVERSITET

TROLLing

Tromsø Repository of Language and Linguistics

Leif Longva
UiT The Arctic University of Norway
The Library

What TROLL?

- A place for linguists to archive and share:
  - Data (from text mining)
  - Statistical analyses
  - R scripts used
  - Not the text or corpus itself

Quantitative analysis

- Linguistics has taken a quantitative turn in recent years
- Access to vast quantities of linguistic corpus for analysis
- Statistical software have become widely available
- -> Statistical methods increasingly used

Sharing data and code

- The community should make a commitment to publicly archive both our data and the statistical code used to analyze it
- Goal to create an ethical standard for sharing data and code
- Need a designated archive for public access to data

Best practices

- Establish best practices in quantitative approaches to theoretical questions
  - Include statistical methods and significance measures
  - Analysis, management, and sharing of data and statistical code
- Data sharing and best practices can also help us to reduce the risk of fraud

A collective learning experience

- Access to examples of datasets and corresponding models will help in choosing the right models for our data
  - A collective learning experience

Requirements from funders

- Agencies also require researchers to share their data with any colleagues who ask for it – particularly common in medicine
- Such conditions will likely be placed upon grant funding for linguistics as well
- Both public archiving and submission of data can be accomplished via the same task, preparing annotations for datasets and code that facilitate the work of peer reviewers and colleagues

Data used in publishing

- In natural sciences, medicine, and psychology authors are routinely requested to submit their data along with their manuscripts when seeking publication in a journal
- We can expect similar requests to become more common in connection with submissions to linguistics journals in the future

Data and publication cross-referring

- Archiving data and code on TROLLing will supplement and not replace or diminish the opportunity to publish articles in scholarly journals
- Ideally scholars will publish their articles and simultaneously put their data and code on TROLLing with cross-references in both directions
- TROLLing will ultimately facilitate more ambitious metastudies by making it possible to do comparisons across datasets, as is common in biostatistics, for example

Example

R. Harald Baayen, Anna Endresen, Laura A. Janda, Anastasia Makarova, Tore Nesset. Forthcoming. "Making Choices in Russian: Pros and Cons of Statistical Methods for Rival Forms" In a special issue of *Russian Linguistics* entitled *Space and Time in Russian Temporal Expressions*, guest edited by Stephen M. Dickey, Laura A. Janda, and Tore Nesset

This website provides data and R scripts for the analyses in our article.

NOTE: If you are already a proficient R user, skip down to the next horizontal line to get the data and R scripts.

How to download R

You can download the R statistical software package to your computer from the R project webpage. We recommend that you use the Austrian CRAN mirror since not all CRAN mirrors include the packages needed to run our scripts.

Once you have downloaded R, you will need to install the following packages: rms, Hmisc, party, modeltools, coin, mvtnorm, zoo, sandwich, strucchange, vcd, colorspace, ndl, lme4, languageR, multcomp. Use the Package Installer in the Menu and Get List to search for these packages.

How to download and run the files from this website

On this webpage we offer you two types of files that you can download to your computer. You can download these files by right-clicking on the links on this pa

Alternative methods for running R scripts

If you simply click on the links with the R scripts, you can then copy and paste all of the code into the R window and R will run the commands and give you the same results. Another option is to download the R script to any location in your computer you want to and provide the path to the file when you use the source command. For example, you can enter a line that looks like this: > source("/Users/janedoe/Downloads/LOAD.R") for Mac users or > source("C://Documents /LOAD.R") for PC users. If you do not know the path, you can open your finder to where the R script is and then drag and drop that file into an open R windo placing it after the cursor prompt ">". When you do this, R will tell you what the path to the file is and you can copy and paste that into the source command.

3.1 Грузить 'load' and its perfectives in the theme-object vs. goal-object constructions

This is the LOAD data: datLOAD.csv

This is the LOAD R script: LOAD.R

3.2 Пере- vs. пре-

This is the PERE data: datPERE.csv

This is the PERE R script: PERE.R

3.3 О- vs. Об-

This is the OB data: datOB.csv

This is the OB R script: OB.R

3.4 -Ну vs. Ø

This is the NU data: datNU.csv

This is the NU R script: NU.R

B1 CONSTRUCTION

	A	B	C	D	E	F	G	H	I	J
1	X	CONSTRUCT	VERB	REDUCED	PARTICIPLE					
2	1	theme	zero	no	no					
3	2	theme	zero	no	no					
4	3	theme	zero	no	no					
5	4	theme	zero	no	no					
6	5	theme	zero	no	no					
7	6	theme	zero	no	no					
8	7	theme	zero	no	no					
9	8	theme	zero	no	no					
10	9	theme	zero	no	no					
11	10	theme	zero	no	no					
12	11	theme	zero	no	no					
13	12	theme	zero	no	no					
14	13	theme	zero	no	no					
15	14	theme	zero	no	no					
16	15	theme	zero	no	no					
17	16	theme	zero	no	no					
18	17	theme	zero	no	no					
19	18	theme	zero	no	no					
20	19	theme	zero	no	no					
21	20	theme	zero	no	no					
22	21	theme	zero	no	no					
23	22	theme	zero	no	no					
24	23	theme	zero	no	no					
25	24	theme	zero	no	no					
26	25	theme	zero	no	no					
27	26	theme	zero	no	no					
28	27	theme	zero	no	no					
29	28	theme	zero	no	no					
30	29	theme	zero	no	no					
31	30	theme	zero	no	no					
32	31	theme	zero	no	no					
33	32	theme	zero	no	no					

```
##############################################################################
# logistic regression
##############################################################################


dat = read.csv("datLOAD.csv",T)
library(rms)
dat.dd = datadist(dat)
options(datadist="dat.dd")
dat.lrm = lrm(CONSTRUCTION~VERB+REDUCED+PARTICIPLE+VERB*PARTICIPLE, data=dat)
dat.lrm
#
#                       Model Likelihood     Discrimination    Rank Discrim.
#                          Ratio Test            Indexes          Indexes
#Obs           1920    LR chi2    1738.47    R2        0.796    C         0.964
# goal          871    d.f.             8    g         4.643    Dxy       0.928
# theme        1049    Pr(> chi2) <0.0001    gr      103.877    gamma     0.945
#max |deriv| 2e-08                          gp        0.459    tau-a     0.460
#                                           Brier     0.076
#
#                               Coef    S.E.   Wald Z Pr(>|Z|)
#Intercept                     -0.9465 0.2023 -4.68  <0.0001
#VERB=po                        6.7143 1.0220  6.57  <0.0001
#VERB=za                        1.0920 0.2451  4.45  <0.0001
#VERB=_zero                     2.3336 0.2446  9.54  <0.0001
#REDUCED=yes                   -0.8891 0.1748 -5.09  <0.0001
#PARTICIPLE=yes                -4.1862 1.0220 -4.10  <0.0001
#VERB=po * PARTICIPLE=yes       3.8953 1.5978  2.44  0.0148
#VERB=za * PARTICIPLE=yes       1.4087 1.0774  1.31  0.1910
#VERB=_zero * PARTICIPLE=yes -1.7717 1.4415 -1.23  0.2190

############################
##### table showing accuracy of lrm model inserted by LAURA
######################
probabilityTheme = plogis(predict(dat.lrm))
tab = table(dat$CONSTRUCTION=="theme", probabilityTheme >= 0.5)
tab
#                       FALSE TRUE
```

The role of the library

- The library holds knowledge and experience in:
  - Building and running databases
  - Metadata
  - Open Access publishing and infrastructure
  - Infrastructure and standards for exchanging database content
  - How to optimize dissemination and discovery of the content
- TROLLing contacted the library to this end

TROLLing and CLARINO

- CLARINO is an exciting project
- TROLLing will gain from connecting with CLARINO
  - And hopefully vice versa

Looking forward

- TROLLing will contribute to establishing best practices for the analysis, management, and sharing of data and statistical code
- If this can be accomplished, we will have much more to look forward to in terms of linguistic discoveries and theoretical insights