

Bootcamp 'Corpus linguistics and/or statistics with R'

Instructor
Stefan Th. Gries
University of California, Santa Barbara

Organizing contact
Stefanie Wulff
University of North Texas, Denton

After several years of successful bootcamps on corpus linguistics as well as statistics for linguists with R, this year offers participants the possibility to

- either take a **30 contact hours corpus linguistics bootcamp**, in which they learn how to write R scripts to extract and process data from differently -annotated corpora to generate different types of frequency lists, concordance displays, etc.;
- or take a **30 contact hours statistics for linguistics bootcamp**, in which they learn how to use R to perform many different monofactorial significance tests, visualize results, and perform regressions as well as hierarchical cluster analysis;
- or take **both** in a row to learn, first, how to get data out of corpora and then, second, how to evaluate them statistically.

I. Syllabi/contents

1. Bootcamp 'Corpus linguistics with R'

The corpus bootcamp is a 30-hours hands-on introduction to quantitative corpus linguistics for both graduate students and seasoned researchers. Using the open source software and programming language R, we will learn

- how to generate frequency lists and search for words and patterns;
- how to process corpora and perform corpus-linguistic searches in ways that typical corpus software does not support;
- how to write small functions for recurrent corpus-linguistic tasks.

Data to be dealt with include plain text corpora, corpora with SGML or XML annotation, ICE-GB files, and others. The participants will also get small functions and scripts they can use for their own corpus-linguistic tasks (concordancing, generating n -grams of words or characters, and others).

The content of this corpus linguistics bootcamp is based on Gries (2009e, <<http://tinyurl.com/QuantCorpLingWithR>>) but (i) structured differently to accommodate the workshop format of the bootcamp and (ii) provides functions and examples not discussed in it.

2. Bootcamp 'Statistics for linguistics with R'

The statistics bootcamp is a 30-hours hands-on introduction to statistical methods for both graduate students and seasoned researchers. Using the open source software and programming language R, we will

- briefly recap basic aspects of statistical evaluation as well as several descriptive statistics;
- discuss monofactorial statistical tests for frequencies, means, dispersions, correlations;

- explore different kinds of multifactorial and multivariate methods, in particular different kinds of regression approaches as well as hierarchical cluster analysis.

For all statistical methods to be explored, we will discuss how to test their assumptions and visualize their results with nice and annotated statistical graphs, and sometimes we will reanalyze published data from corpus-linguistic studies. The participants will also get small functions they can use for their own statistical applications. Also, there will be a small section on how to write small statistical/visualization functions yourself.

The content of this statistics bootcamp is based on Gries (2009d, <<http://tinyurl.com/StatForLingWithR>>), but goes beyond it in terms of the methods and datasets covered; in fact, the bootcamp will use materials currently being integrated into the second edition.

II. Schedule

1. Bootcamp 'Corpus linguistics with R'

Date	Activities
30 July 2012	9.00-10.00 arrival/admin 10.00-1.00 class 7.30 reception dinner
31 July 2012	9.00-12.15 class
1 August 2012	9.00-12.15 class
2 August 2012	free time
3 August 2012	9.00-12.15 class
4 August 2012	9.00-12.15 class
5 August 2012	departure

Class sessions of more than two hours include a 15-minute break. We are currently exploring the possibility of awarding UCSB credits for the participation in this bootcamp.

2. Bootcamp 'Statistics for linguistics with R'

Date	Activities
6 August 2012	9.00-10.00 arrival/admin 10.00-1.00 class 7.30 reception dinner
7 August 2012	9.00-12.15 class
8 August 2012	9.00-12.15 class
9 August 2012	free time
10 August 2012	9.00-12.15 class
11 August 2012	9.00-12.15 class
12 August 2012	departure

Class sessions of more than two hours include a 15-minute break. We are currently exploring the possibility of awarding UCSB credits for the participation in this bootcamp.

III. Accommodation

We reserved accommodation for up to 20 participants in ten two-bedroom apartments in Santa Ynez Apartments, a housing complex next to the UCSB campus. The apartments have

- two separate bedrooms;
- a shared living area and a shared kitchen area;
- 1.5 bathrooms;
- bed linens (sheets, a blanket, a pillow plus pillowcase);
- laundry facilities and wireless internet access.

Cf. <<http://www.housing.ucsb.edu/conferences/pdfs/summsheet-santaynez.pdf>>.

IV. Pricing, registration, and payment

Registration is on a first-pay-first-served basis, beginning immediately, with priority given to participants signing up for options 3a) or 3b), i.e., both bootcamps. If you would like to register, send an email to Stefanie (<Stefanie.Wulff@unt.edu>) with "Bootcamps 2012" in the subject header and state which of the six registrations options you would like to book. Provided that there are still open seats, Stefanie will send you a registration form with instructions for payment. The following six registration options are available:

1. Only the corpus bootcamp

a) Registration only: US\$300.-	b) Registration + accommodation: US\$650.-
This price includes 5 days of workshop instruction plus materials and a reception dinner. This option means you will have to organize your own accommodation.	This price includes 5 days of workshop instruction plus materials, a reception dinner, and accommodation in a bedroom in a Santa Ynez apartment from 29 July to 5 August (see III.).

2. Only the statistics bootcamp

a) Registration only: US\$300.-	b) Registration + accommodation: US\$650.-
This price includes 5 days of workshop instruction plus materials and a reception dinner. This option means you will have to organize your own accommodation.	This price includes 5 days of workshop instruction plus materials, a reception dinner, and accommodation in a bedroom in a Santa Ynez apartment from 5 August to 12 August (see III.).

3. Both bootcamps

a) Registration only: US\$600.-	b) Registration + accommodation: US\$1200.-
This price includes 10 days of workshop instruction plus materials and two reception dinners. This option means you will have to organize your own accommodation.	This price includes 10 days of workshop instruction plus materials, two reception dinners, and accommodation in a bedroom in a Santa Ynez apartment from 29 July to 12 August (see III.).

Note: Ideally, you **register (including payment) before 25 April 2012** because that is when we must commit financially to a minimum number of places and accommodation at UCSB; after that, registration may be more restricted. This also means that if you decide to cancel your registration, you will be refunded 50% of your payment (still much better than the 0% refund that conferences offer).

V. Recreation

Workshop participants can take advantage of the UCSB Recreation Center (<<http://www.recreation.ucsb.edu/>>) for a daily charge of approximately US\$12.-.

VI. The instructor

Stefan Th. Gries (STG) is Full Professor of Linguistics in the Department of Linguistics at the University of California, Santa Barbara (UCSB), Honorary Liebig-Professor of the Justus-Liebig-Universität Giessen, and he was a Visiting Professor at the 2007 and the 2011 LSA Linguistic Institutes at Stanford University and the University of Colorado at Boulder.

Methodologically, STG is a quantitative corpus linguist at the intersection of corpus linguistics, cognitive linguistics, and computational linguistics, who uses a variety of different statistical methods to investigate linguistic topics such as morpho-phonology (the formation of morphological blends), syntax (syntactic alternations), the syntax-lexis interface (collostructional analysis), and semantics (polysemy, antonymy, and near synonymy in English and Russian) and corpus-linguistic methodology (corpus homogeneity and comparisons, dispersion measures, *n*-gram identification and exploration, and other quantitative methods), as well as first and second language acquisition. Occasionally, he also uses experimental methods (acceptability judgments, sentence completion, priming, self-paced reading times, and sorting tasks).

Theoretically, he is a cognitively-oriented linguist (with an interest in Construction Grammar) in the wider sense of seeking explanations in terms of cognitive processes without being a cognitive linguist in the narrower sense of following any one particular cognitive-linguistic theory. The researchers who have influenced his work most are (in alphabetical order) R. Harald Baayen, Douglas Biber, Nick C. Ellis, Adele E. Goldberg, and Michael Tomasello.

STG has authored three books – one research monograph, an introduction to statistics with R for linguists, and a book on corpus linguistics with R. He has also co-edited four volumes – two on corpora in cognitive linguistics, one on corpus linguistics, and one on cognitive linguistics – and two more (on frequency effects in cognitive linguistics and psycholinguistics are contracted and scheduled to appear in 2012 (with Mouton de Gruyter). Since 1999, he has (co-)authored more than three dozen articles in the leading peer-reviewed journals of his fields

(*Cognitive Linguistics* and *International Journal of Corpus Linguistics*) as well as in many other peer-reviewed journals, plus another 50 or so articles in edited volumes, proceedings, etc. He is founding editor-in-chief of the international peer-reviewed journal *Corpus Linguistics and Linguistic Theory*, associate editor of *Cognitive Linguistics*, and performs editorial functions for the following international peer-reviewed journals: *Brazilian Journal of Applied Linguistics*, *CogniTextes*, *Constructions and Frames*, *Corpora*, *Journal of Advanced Linguistic Studies*, and *Language and Cognition*. Since the beginning of 2007, he has given more than 100 talks and workshops at national and international venues, more than half of them invited.

VII. Contacts

For further information, please contact

- Stefan Th. Gries (<stgries@linguistics.ucsb.edu>) for all questions having to do with the content of the bootcamps, or
- Stefanie Wulff (<Stefanie.Wulff@unt.edu>) for all questions having to do with organizational matters

by email; please use "Bootcamps 2012" as the subject header in all correspondence!

VIII. Links

Stefan Th. Gries (instructor)	http://tinyurl.com/stgries
Stefanie Wulff (organizer)	http://tinyurl.com/swulff
Bootcamp overview map:	http://g.co/maps/gp5fe
UC Santa Barbara	http://www.ucsb.edu/ http://www.aw.id.ucsb.edu/maps/
Department of Linguistics, UC Santa Barbara	http://www.linguistics.ucsb.edu/
Santa Barbara, CA	http://www.santabarbaraca.com/ http://www.santabarbaraca.gov/
Santa Barbara airport	http://www.flysba.com/
<i>Statistics for Linguistics with R</i>	http://tinyurl.com/StatForLingWithR http://groups.google.com/group/statforling-with-r/topics
<i>Quantitative Corpus Linguistics with R</i>	http://tinyurl.com/QuantCorpLingWithR http://groups.google.com/group/corpling-with-r/topics
R	http://www.r-project.org/ http://cran.r-project.org/index.html
RStudio	http://www.rstudio.org