Methodological Issues

Laura A. Janda
Overview

1. Introduction: Some History
2. What is Language?
3. What is Linguistic Data?
4. The Quantitative Turn: Dangers and Opportunities
5. Some Popular Quantitative Methods
6. Conclusions
1. Introduction: Some History

A Quantitative Turn is definitely underway
   – in Cognitive Linguistics
   – in linguistics in general

A view from a single journal:
   *Cognitive Linguistics* 1990-2014
Quantitative studies were always with us. Over 50% since 2008, leveling off at 75-80%?
Other Indicators of a Quantitative Turn in Cognitive Linguistics

Conferences: QITL

- Since 2002 Quantitative Investigations in Theoretical Linguistics, conference series devoted to statistical analysis of language data from the point of view of cognitive linguistics.

Recent textbooks on statistical methods in cognitive linguistics:

What has motivated the Quantitative Turn?

Historical confluence of three factors:
- advent of electronic archives of language data and means to share them
- development of free open-access statistical software
- usage-based nature of the cognitive linguistics framework
Electronic archives of language data

• Modern **corpora** became widespread only a little over a decade ago
• Today we have access to balanced multi-purpose corpora for many languages, often containing **hundreds of millions of words**, some with linguistic annotation
• Many languages have **national corpora**, and **open corpora** are being built, providing free access not only to linguistic forms and, but also to the code, facilitating further exploration of data
• For a few languages, “very large” corpora of **billions of words** are being created, usually by means of webcrawling + deduplication
• Google Books Ngrams Corpus has a function that charts the frequency of words and phrases in a few of the world’s largest languages
• **Spoken** corpora are becoming available
• **Multimodal** corpora: UCLA NewsScape Library is an archive of billions of words in several languages, along with associated sound and images captured from television newscasts
Electronic archives of language data

- Modern corpora became widespread only a little over a decade ago.
- Today we have access to balanced multi-purpose corpora for many languages, often containing hundreds of millions of words, some with linguistic annotation.
- Many languages have national corpora, and open corpora are being built, providing free access not only to linguistic forms but also to the code, facilitating further exploration of data.
- For a few languages, “very large” corpora of billions of words are being created, usually by means of webcrawling and deduplication.
- Google Books Ngrams Corpus has a function that charts the frequency of words and phrases in a few of the world’s largest languages.
- Spoken corpora are becoming available.
- Multimodal corpora: UCLA NewsScape Library is an archive of billions of words in several languages, along with associated sound and images captured from television newscasts.

We now have access to LOTS of language data.
Development of free open-access statistical software: R

- Find R at: https://cran.r-project.org/
- “R is … a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.”
- R contains functionality for a large number of statistical procedures. There is also a large set of functions which provide a flexible graphical environment for creating various kinds of data presentations.
- Initiated by Ross Ihaka and Robert Gentleman at University of Auckland, New Zealand in 1993. Many individuals contribute to R by sending code and bug reports.
- Since mid-1997 there has been a core group (the “R Core Team”) who modify the R source code archive.
Why use R?

Advantages of R

• FREE, open-source
• very powerful
• supports UTF-8 encoding
• it is the standard in linguistics
• huge network of users, you can google most problems
• there is a special group for linguists using R: https://groups.google.com/forum/#!forum/statforling-with-r

• Baayen has created an R package for linguists: languageR (see https://cran.r-project.org/web/packages/languageR/index.html)

Disadvantages of R

• help function is not very helpful
• the learning curve is steep
• sometimes the software will get updated and your code won’t work anymore (not likely for simpler tests, but this does happen with regression and more advanced models)
Alternatives to R (none of these are open-source)

- Microsoft XL
  - $ (comes with Microsoft Office)
  - limited capacity
- Stata
  - $$ ($295-$595)
- SPSS
  - $$$ ($2250-$6750)
- SAS
  - $$$$$ (about $10000)
2. What is Language?

Generativism
- Universal Grammar
- Hard-wired
- Autonomous
- Rule-based
- Grammar vs. Lexicon
- Competence vs. Performance

Cognitive Linguistics
- Emergent structure
- Uses general cognitive mechanisms
- Motivated by meaning
- Usage-based
## 2. What is Language?

<table>
<thead>
<tr>
<th>Generativism</th>
<th>Cognitive Linguistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Universal Grammar</td>
<td>• Emergent structure</td>
</tr>
<tr>
<td>• Hard-wired</td>
<td>• Uses general cognitive mechanisms</td>
</tr>
<tr>
<td>• Autonomous</td>
<td>• Motivated by meaning</td>
</tr>
<tr>
<td>• Rule-based</td>
<td>• Usage-based</td>
</tr>
<tr>
<td>• Grammar vs. Lexicon</td>
<td></td>
</tr>
<tr>
<td>• <strong>Competence vs. Performance</strong></td>
<td></td>
</tr>
</tbody>
</table>


3. What is Linguistic Data?

• Even the question of what constitutes data in linguistics is controversial.
• Some researchers refer to constructed examples and individual intuitions as data, while others prefer to use corpus attestations or observations from acquisition or experiments.
• WHY?
• The definition of what counts as data depends on theory.
Competence vs. Performance

Competence
• Langue
• i-language
• available via introspection

Performance
• Parole
• e-language
• available via observation
Competence vs. Performance

**Competence**
- Langue
- i-language
- available via introspection

**Performance**
- Parole
- e-language
- available via observation

From the Generativist perspective, only Competence is language.
Chomsky rejects performance (corpus) data

Noam Chomsky (1957) Syntactic Structures:

- p. 15: “...it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances ... a grammar mirrors the behavior of the speaker, who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences.”

Chomsky changed the direction of linguistics away from empiricism and towards rationalism. He invalidated the corpus as a source of evidence in linguistic enquiry. Chomsky suggested that the corpus could never be a useful tool for the linguist, as the linguist must seek to model language competence rather than performance.
Chomsky in the 1960s…

• “We thus make a fundamental distinction between competence (the speaker-hearer’s knowledge of his language) and performance (the actual use of language in concrete situations). [...] A record of natural speech will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on. The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system of rules that have been mastered by the speaker-hearer and that he puts to use in actual performance” (Chomsky 1965).

Chomsky’s orthodoxy (Hill 1962)
  Chomsky: The verb ‘perform’ cannot be used with mass word objects: one can ‘perform a task’ but one cannot ‘perform labor’
  Hatcher: How do you know, if you don’t use a corpus and have not studied the verb ‘perform’?
  Chomsky: How do I know? Because I am a native speaker of the English language
• But, from BNC corpus, it is possible to ‘perform magic’!
Chomsky more recently


• "Corpus linguistics doesn’t mean anything. It’s like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they’re going to do is take videotapes of things happening in the world and they’ll collect huge videotapes of everything that’s happening and from that maybe they’ll come up with some generalizations or insights."
question: "Think of the occurrence of 'Can you...' or, 'Could you... .' rather than 'Are you able to... .' in polite requests in given communicative situations (a domain studied extensively by speech act theorists). Such chunks of linguistic expressions can be traced by the researcher via the application of corpus linguistic methods. It is from a corpus that one can identify their frequency and trace shifts in their meaning and use. Would you attribute significance to such data in your approach to linguistic analysis and description?"

answer: "People who work seriously in this particular area do not rely on corpus linguistics. They may begin by looking at facts about frequency and shifts in frequency and so on, but if they want to move on to some understanding of what’s happening they will very quickly, and in fact do, shift to the experimental framework. Where you design situations, you enquire into how people will act in those situations. You design them within a framework of theoretical inquiry which has already suggested that these are likely to be important questions and I want the answers to them. But that’s not corpus linguistics.”
**Competence vs. Performance**

<table>
<thead>
<tr>
<th>Competence</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Langue</td>
<td>• Parole</td>
</tr>
<tr>
<td>• i-language</td>
<td>• e-language</td>
</tr>
<tr>
<td>• available via</td>
<td>• available via</td>
</tr>
<tr>
<td>introspection</td>
<td>observation</td>
</tr>
</tbody>
</table>

No distinction in Cognitive Linguistics

**Competence and Performance** are both language
Introspection & Observation

Introspection
• Source of inspiration for hypotheses and theoretical advances
• Indispensable to interpret results and understand what they mean for theory and facts of language
• Necessary to ferret out suspicious results and alert us to problems in design and analysis

Observation
• Used to test hypotheses
• Essential to the scientific method

When intuitions become observations: systematic elicitation of intuitions of naïve informants under experimental conditions and insights gained during fieldwork
What happens when you rely only on introspection?

Scientific researchers naturally want to find an orderly logical system in which crisp distinctions yield absolute predictions.

- But there is no a priori reason to make this assumption, and usage data typically do not support it.

Continuousness and variation are “messy”

- But that doesn’t mean we should pretend they aren’t there.

What about “performance error”?

- It is usually marked by speaker repair.
- All naturalistic data contains some noise.
- Statistics can sort trends from noise.
What happens when you rely only on introspection?

Scientific researchers naturally want to find an orderly logical system in which crisp distinctions yield absolute predictions. But there is no a priori reason to make this assumption, and usage data typically do not support it. Continuousness and variation are "messy." But that doesn't mean we should pretend they aren't there.

What about "performance error"? It is usually marked by speaker repair. All naturalistic data contains some noise. Statistics can sort trends from noise.

Reliance on introspection to the exclusion of observation undermines linguistics as a science, yielding claims that can be neither operationalized nor falsified.
What happens when you rely only on introspection?

Scientific researchers naturally want to find an orderly logical system in which crisp distinctions yield absolute predictions. But there is no a priori reason to make this assumption, and usage data typically do not support it. Continuousness and variation are “messy.” But that doesn’t mean we should pretend they aren’t there.

What about “performance error”? It is usually marked by speaker repair. All naturalistic data contains some noise. Statistics can sort trends from noise.

Reliance on observation to the exclusion of introspection is also bad...
4. The Quantitative Turn: Dangers and Opportunities

Dangers

- “Cargo cult science”
- Erosion of the core of the field of linguistics
- “Language is never, ever, ever random”
- Exacerbation of discrimination against small languages
- Arms race of statistical models
- Temptation to fudge data
- Empowerment of hacking and spyware
WWII aftermath: Cargo cult
“Cargo cult science”

• Feynman (1992) compared inept scientists to “cargo cult” south sea islanders, who, after experiencing airlifts during WWII, constructed mock runways manned by mock air traffic controllers, in hopes that this would cause more airplanes to land and bring them cargo

• Linguists perform empty rituals of calculations in hopes of conjuring up publishable results

• Very easy for researchers to be seduced by fancy equipment and sophisticated software

• Number-crunching without any real linguistic or theoretical goal

• Triviality and fractionalization of the field
Erosion of the core of the field of linguistics

- Core of our field are linguistic description and theoretical interpretation, which are also the source for research hypotheses.

- Substitution of “quantitative” for “empirical” and “scientific” in the minds of researchers is problematic.

- Use of quantitative methods in a study does not make it better or necessarily any more empirical or scientific than language documentation or qualitative analysis.

- Confusion of these concepts could result in marginalization of traditional endeavors of linguists in conferences and publications.

- “Linguistic investigation is a highly complex and multifaceted enterprise requiring many kinds of methods and expertise” (Langacker 2015).
“Language is never, ever, ever random”

- Kilgarriff (2005) argues that since languages do not behave in a random fashion, the use of statistics to test null hypotheses is perhaps misguided to begin with.

- It becomes far too easy to find results simply because as the number of observations increases toward infinity (or just millions and billions), statistical tests are able to find effects that are infinitesimally small and therefore meaningless.
Exacerbation of discrimination against small languages

- Big data analysis also threatens to marginalize minority languages
- Only a tiny fraction of the world’s languages have the resources to support large corpora, experimental studies, and comprehensive language technology coverage
- Exacerbate the existing imbalance between the few languages that many linguists study and the majority of languages that are largely ignored
Speaker count vs. language count:
Data from Ethnologue (1999)

6059 languages with 1 or more speakers ... 80 languages with $10^7$ or more speakers
Arms race of statistical models

Complex models are well understood only by the statisticians who developed them.

“Black box” methods will not enhance the ability of linguists to understand and communicate their results.

It is usually the case that the simplest model that is appropriate to the data is the best one to use, since the results will be most accessible to readers.
Temptation to fudge data

• Pressures to publish present an incentive to falsify results in hopes of impressing reviewers at a prestigious journal

• Most academic fields in which researchers report statistical findings have experienced scandals involving fudged data or analyses
Diederik Alexander Stapel, former professor of social psychology, fired in 2011 for manipulating data in at least 55 publications...

"I have failed as a scientist."
Empowerment of hacking and spyware

- Major corporations such as Google, Amazon, Apple, and Facebook, along with hacking and spyware operations and state governments, have access to massive quantities of human language data.
- Technology for mining techniques is advancing rapidly, and linguists contribute to it by improving our understanding of languages.
- The “technium”: collective of archives and devices that constitute an organism-like system with a powerful momentum (Kelly 2010).
- Our only defense is to keep as much of it as possible in the public domain rather than behind clandestine corporate, state, and criminal firewalls.
Opportunities

• Discover structures in linguistic data that would otherwise escape our notice
• Secure and maintain the status of linguistics as a science
• Foster greater accountability and collaboration
• Establish best practices in quantitative approaches to theoretical questions
A PLACE FOR US:

TROLLing
The Tromsø Repository of Language and Linguistics
TROLLing

- is an international archive of linguistic data and statistical code
- is built on the Dataverse platform from Harvard University and complies with DataCite, the international standard for storing and citing research data
- is compliant with CLARIN, the EU research infrastructure for language-based resources
- assigns a permanent URL to each post
- uses metadata that ensures visibility and retrieval through international services
- is professionally managed by the University Library of Tromsø and an international steering committee
Find TROLLing at opendata.uit.no

<table>
<thead>
<tr>
<th>TROLLing</th>
<th>Search Studies</th>
<th>Studies: 18</th>
<th>Downloads: 177</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication data for: Ditransitive constructions in Russian and Ukrainian&lt;br&gt;by Mykhaylyk, Roksolana; Redina, Yuliia; Andersen, Merete&lt;br&gt;Description: Publication abstract: This paper presents new experimental data on the acquisition of structures involving ditransitive verbs in two East Slavic languages: Russian and Ukrainian. The results of an elicited production task with adults and 3... Continue »</td>
<td>hdl:10037.1/10005</td>
<td>39 downloads</td>
<td>Last Released: Jun 18, 2014</td>
</tr>
<tr>
<td>Replication data for: Old Church Slavonic bytъ Part One: Grammatical Profiling Analysis&lt;br&gt;by Eichhoff, Horne M.; Janda, Laura A.; Neeset, Tore&lt;br&gt;Description: There is controversy over whether bytъ 'be' in Old Church Slavonic functioned as an imperative verb with an unusually large number of inflected forms or as an aspectual pair of verbs, reflecting its suppletive origin from two stems (ex-... Continue »</td>
<td>hdl:10037.1/10008</td>
<td>6 downloads</td>
<td>Last Released: May 21, 2014</td>
</tr>
<tr>
<td>Spanish deverbal adjectives (1): hvo&lt;br&gt;by Fabregas, Antonio&lt;br&gt;Description: Attested adjectives in -hvo in contemporary Spanish, with information about their bases, their semantic interpretation and their morphological properties.</td>
<td>hdl:10037.1/10018</td>
<td>18 downloads</td>
<td>Last Released: Jun 16, 2014</td>
</tr>
<tr>
<td>Metonymy in Word-Formation: Russian, Czech, and Norwegian&lt;br&gt;by Janda, Laura A.&lt;br&gt;Description: Publication abstract: A foundational goal of cognitive linguistics is to explain linguistic phenomena in terms of general cognitive strategies rather than postulating an autonomous language module (Langacker 1987: 12-13). Metonymy is ident... Continue »</td>
<td>hdl:10037.1/10029</td>
<td>3 downloads</td>
<td>Last Released: Jun 13, 2014</td>
</tr>
<tr>
<td>Why Russian Prefixes Aren’t Empty&lt;br&gt;by Janda, Laura A.&lt;br&gt;Description: Publication abstract: We claim that Russian verbal prefixes always express meaning, even when they are used to form the perfective partners of aspectual pairs. The prefixes in verbs like написать/napisat’ ‘write’ and совершить/sveshit’ ‘do... Continue »</td>
<td>hdl:10037.1/10022</td>
<td>10 downloads</td>
<td>Last Released: Jun 16, 2014</td>
</tr>
<tr>
<td>Russian nu-drop verbs</td>
<td>hdl:10037.1/10024</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Some Popular Quantitative Methods

There are many statistical models
Three that are particularly useful for linguists:
• Chi-square & Cramer’s V
• Correspondence Analysis
• Classification & Regression Trees (CART)
The Chi-square test:
Influence on choice of linguistic form

• Languages often give speakers choices, for example the choice between: A) the ditransitive (*read the children a story*), and B) the prepositional dative (*read a story to the children*) constructions in English

• Corpus or experimental data might reveal a pattern such that there is more use of A in one environment (X) than in another environment (Y)

• But is the difference between the measurements of A and B a significant difference?

• In other words, is there reason to believe that there is a real difference between the frequency of A and B, or might the difference we observe be just a matter of chance (the null hypothesis)?
Russian semelfactive morphemes

• Semelfactive verbs are verbs that mean “do X once”
• In Russian, there are two ways to form such verbs:

  Suffix –nu  
  Example:  
  čixat’ ‘sneeze’  
  >>  
  čixnut’ ‘sneeze once’

  Prefix s-  
  Example:  
  glupit’ ‘act stupid’  
  >>  
  sglupit’ ‘act stupid once’
Some verb classes prefer *-nu*

- *-aj*

non-prod

1. conj

- *-ě*

Zevnul

Liznula

Svistnula
Some verb classes prefer $s$-

-ova

Smalodušestvoval

-i

Sgrubil!

-*ěj

Srobela?
The chi-square test compares the observed and the expected values

The observed values:

<table>
<thead>
<tr>
<th>aj</th>
<th>non_prod</th>
<th>e</th>
<th>ova</th>
<th>i</th>
<th>ej</th>
</tr>
</thead>
<tbody>
<tr>
<td>nu</td>
<td>185</td>
<td>57</td>
<td>20</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>s</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36</td>
</tr>
</tbody>
</table>

The expected values:

<table>
<thead>
<tr>
<th>aj</th>
<th>non_prod</th>
<th>e</th>
<th>ova</th>
<th>i</th>
<th>ej</th>
</tr>
</thead>
<tbody>
<tr>
<td>nu</td>
<td>141.05398</td>
<td>43.22622</td>
<td>15.92545</td>
<td>26.542416</td>
<td>40.95116</td>
</tr>
<tr>
<td>s</td>
<td>44.94602</td>
<td>13.77378</td>
<td>5.07455</td>
<td>8.457584</td>
<td>13.04884</td>
</tr>
</tbody>
</table>
The p-value

• The p-value for a chi-square test tells you the probability that there is no difference between the observed values and the expected values for your contingency table.

• Another way to say this: The p-value tells you the probability that you could get this much or more difference between the observed and expected values if the observed values were drawn from a population that had the distribution of the expected values.

• For the Russian semelfactives, this means that the p-value will tell me the probability that I would get such an extreme distribution if there was no association between –nu vs. s- and the verb classes; i.e. if the choice of –nu vs. s- was just random for every verb regardless of verb class.
Now let’s evaluate our Russian semelfactive data

Pearson's Chi-squared test

X-squared = 269.22, df = 5, p-value < 2.2e-16

Cramer’s V is the square root of the chi-square value divided by the sample size multiplied by the smallest degree of freedom (for rows or for columns)

Cramer’s V < 0.1 means effect is too small to be worth reporting
0.3 = medium effect size
0.5 = large effect size

For this data, Cramer's V : 0.832
Correspondence Analysis: Measuring the distances between forms

How learnable is Russian Aspect?

• All forms of all verbs obligatorily express **perfective** vs. **imperfective** aspect

  • **Perfective aspect**: unique, complete events with crisp boundaries
    – *Pisatel’ na-pisal/na-pišet roman* ‘The writer has written/will write a novel’

  • **Imperfective aspect**: ongoing or repeated events without crisp boundaries
    – *Pisatel’ pisol/pišet roman* ‘The writer was writing/is writing a novel’

• Morphological marking is not entirely reliable:
  – **bare verb**: usually **imperfective** (*pisat’* ‘write’), some biaspectual
    (*ženit’sja* ‘marry’), a few perfective (*dat’* ‘give’)
  – **prefix + verb**: usually **perfective** (*pere-pisat’* ‘rewrite’), some imperfective
    (*pre-obladat’* ‘prevail’, *pere-xodit’* ‘walk across’)
  – **prefix + verb + suffix**: **imperfective** (*pere-pis-yva-t’* ‘rewrite’)

CLEAR
Correspondence Analysis: Measuring the distances between forms

How learnable is Russian Aspect?

• All forms of all verbs obligatorily express perfective vs. imperfective aspect
  - Perfective aspect: unique, complete events with crisp boundaries
    – Pisatel’ na-pisal/na-pišet roman ‘The writer has written/will write a novel’
  - Imperfective aspect: ongoing or repeated events without crisp boundaries
    – Pisatel’ pisal/pišet roman ‘The writer was writing/is writing a novel’

• Morphological marking is not entirely reliable:
  – bare verb: usually imperfective (ženit’sja ‘marry’), a few perfective
  – prefix + verb: usually perfective (pre-obladat’ ‘prevail’, pere-xodit’ ‘walk across’)
  – prefix + verb + suffix: imperfective

If morphological marking is not reliable, how do native speakers figure out which verbs are perfective and which are imperfective?
Correspondence Analysis of Journalistic Data

**Input:** 185 vectors (1 for each verb) of frequencies for verb forms

Each vector tells how many forms were found for each verbal category:
- indicative non-past
- indicative past
- indicative future
- imperative
- infinitive
- non-past gerund
- past gerund
- non-past participle
- past participle

**Rows are verbs, columns are verbal categories**

**Process:**

Matrices of distances are calculated for rows and columns and represented in a multidimensional space defined by factors that are mathematical constructs. Factor 1 is the mathematical dimension that accounts for the largest amount of variance in the data, followed by Factor 2, etc.

Plot of the first two (most significant) Factors, with Factor 1 as x-axis and Factor 2 as the y-axis

You can think of **Factor 1** as the **strongest parameter** that splits the data into **two groups** (negative vs. positive values on the x-axis)
On the Following Slide...

- Results of correspondence analysis for Journalistic data
- Perfective verbs represented as “p”
- Imperfective verbs represented as “i”
- Remember that the program was not told the aspect of the verbs
- All it was told was the frequency distributions of grammatical forms
- All it was asked to do was to construct the strongest mathematical Factor that separates the data along a continuum from negative to positive (x-axis)
Factor 1 (39.1 %)

- Factor 2 (19.7 %)

- Perfective

- Imperfective

- Factor 1 looks like aspect
Factor 1 correctly predicts aspect 91.5%  
(negative = perfective vs. positive = imperfective)

Of the 185 verbs:
– 87 perfectives
  • 84 negative values, 3 positive values, so 96.6% correct
    – 3 deviations are: obojit’ ‘make do without’, smoč’ ‘manage’, prijit’ ‘be necessary’
– 96 imperfectives
  • 83 positive values, 13 negative values, so 86.5% correct
– 2 biaspectuals with low negative values
  • obeščat’ ‘promise’, ispol’zovat’ ‘use’
Correspondence Analysis of Russian Verbs

Summary

• When we look at the distribution of verb forms, aspect (or a close approximation) emerges as the most important factor distinguishing verbs.

• It is possible to sort most high-frequency verbs as perfective vs. imperfective based only on the distribution of their forms.
Classification and Regression Trees (CART): Factors and their power to influence form

• Carolin Strobl, head of research unit on Psychological Methods, Evaluation and Statistics at the University of Zurich
• 2009 landmark publication on CART and Random Forests together with Gerhard Tutz and James Malley
Drawbacks of logistic regression

- Assumes normal distribution (parametric assumption), which is usually not the case with corpus data
- Assumes that all combinations of all variables are represented in the dataset, but linguistic data often involves paradigmatic gaps
More drawbacks of logistic regression

• Building the optimal model can be a laborious task
• Results are a table and coefficients that can be hard to interpret
Alternative to regression: CART and random forests

• CART:
  – is non-parametric (does not assume a normal distribution)
  – does not assume all combinations of variable levels are represented
  – does not require tedious building and fine-tuning – just put in all variables
  – provides opportunities for validation
  – yield results that can be easier to interpret (especially trees)
An ongoing language change in North Saami: NPx is being replaced by ReflN

Two examples from Elle Márrjá Vars’ novel Kátjá

NPx (possessive suffix, HIGH morphological complexity):
(1a) Kátjá... ollii latnjasis
Kátjá.NOM reach.IND.PRET.3S room.ILL.SG.PX.3S
‘Kátjá... got to her room’

ReflN (analytic construction with reflexive genitive pronoun):
(1b) Kátjá... ollii iezas latnji
Kátjá.NOM reach.IND.PRET.3S REFL.GEN.3S room.ILL.SG
‘Kátjá... got to her room’
The S-curve: longitudinal data from literary texts, showing only anaphoric and endophoric use.
North Saami CART analysis: variables and levels

**Dependent variable:** PossCon (possessive construction)
- NPx (noun with possessive suffix), ReflN (reflexive genitive pronoun)

**Independent variables:**
- Generation (of author)
  - Old, Mid, Young
- PMClass (semantic class of possessum)
  - Abstraction, Body, Event, Human, Kin, Place, Property, Other
- PMCcase (case of possessum)
  - Acc, Com, Ess, Gen, Ill, Loc, Nom
- PRCase (case of possessor)
  - Acc, Gen, Ill, Loc, Nom, Verb
- Geography
  - East, West
CART and Cognitive Linguistics

A CART tree can literally be understood as an optimal algorithm for predicting an outcome given the predictor values, and Kapatsinski (2013: 127) suggests that from the perspective of a usage-based model, each path of partitions along a classification tree expresses a schema, in the Langackerian sense (Langacker 2013: 23), since it is a generalization over a set of instances.
6. Conclusions

• Since about 2008, cognitive linguistics has shifted its focus, and is now dominated by quantitative studies.
• The quantitative turn is a step forward since it puts powerful new tools into the hands of cognitive linguists.
• Our field can gain in terms of scientific prestige and precision and collaboration.
• We can show leadership in best practices and the norming of application of statistical models to linguistic data.
• We should retain an attitude of respect for our venerable qualitative and theoretical traditions.
• If anything, we need qualitative and theoretical insights now more than ever in order to make sense of all the data.
References