Quantitative Methods in Cognitive Linguistics

Laura A. Janda
laura.janda@uit.no
Cognitive linguistics: Some basic facts

**Minimal Assumption:** language can be accounted for in terms of general cognitive strategies

- no autonomous language faculty
- no strict division between grammar and lexicon
- no a priori universals

**Usage-Based:** generalizations emerge from language data

- no strict division between langue and parole
- no underlying forms

**Meaning is Central:** holds for all language phenomena

- no semantically empty forms
- differences in behavior are motivated (but not specifically predicted) by differences in meaning
- metaphor and metonymy play a major role in grammar
What is Cognitive Linguistics?

- Explanation of linguistic phenomena via general cognitive mechanisms
- Meaning is the motive for language and is embodied in physical experience
- Radial categories based on prototypes with extensions via metaphor & metonymy
- Lexicon & grammar are a continuum, observe same patterns
- Empirical (statistical) analysis of authentic language data
• A survey of use of statistics in articles in *Cognitive Linguistics*, 1990-2014

• Concrete examples of how researchers have applied statistical models in linguistics
percent quantitative articles


Values: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%
The Quantitative Turn: 2008

Facilitated by theoretical and historical factors
- CL is usage-based, data-friendly – quantitative studies have always been part of CL
- Advent of digital corpora and statistical software
- Progress in computational linguistics

What this means for the future
- All linguists will need at least passive statistical literacy
- We need to develop best practices for use of statistics in linguistics
- Public archiving of data and code will contribute to advancement of the field
The Quantitative Turn: 2008

Two eras:

1990-2007: most articles did not present quantitative studies
   BUT: quantitative studies were always with us
2008-2014: most articles do present quantitative studies
   Seems to be leveling off at about 75%

Other indicators:

Quantitative Investigations in Theoretical Linguistics: six biannual conferences since 2002
   Journal: *Corpus Linguistics and Linguistic Theory* since 2005
   Textbooks of statistical methods for linguistics by cognitive linguists: Harald Baayen (2008), Stefan Gries (20013), and Natalia Levshina (2015)
What Brought About the Quantitative Turn?

THEORY
• A usage-based model of language is data-friendly

DATA
• Advent of electronic language resources

TOOLS
• Advent of analytical tools
THEORY: A usage-based model of language is data-friendly

• Cognitive linguistics has always been a “data-friendly” theory, with a focus on the relationship between observed form and meaning.

• We posit no fundamental distinction between “performance” and “competence”, and recognize all language units as arising from usage events.

• Usage events are observable, and therefore can be collected, measured, and analyzed scientifically.
DATA:
Advent of electronic language resources

• What is linguistic data?
  – constructed examples
  – individual intuition
  – corpus attestations
  – observation, experiments

• A usage-based theorist views language use as the data relevant for linguistic analysis, and this gives cognitive linguistics a natural advantage in applying quantitative methods.

• Language corpora
  – National corpora (Russian National Corpus)
  – Multimodal corpora (UCLA NewsScape Library)
  – Google Books Ngrams Corpus

Introspection does not yield data that can be subjected to statistical analysis

Observation does yield data that can be subjected to statistical analysis
Role of Introspection

Introspection is not a method for gathering data for statistical analysis

BUT

• Introspection is the source of inspiration for hypotheses, which are then tested via observation
• Introspection is indispensable in order to interpret the results and understand what they mean for both theory and facts of language
• The data do not speak for themselves; we need introspection in order to understand what they mean
• Introspection is necessary to ferret out suspicious results and alert us to problems in design and analysis
• Theoretical advances are typically born through introspection
TOOLS:
Advent of analytical tools

Statistical software

- The tool of choice for cognitive linguists is primarily “R” (R Development Core Team 2010), which is open-source, supports UTF-8 encoding for various languages, and has a programming package, “languageR”, specially developed by Harald Baayen for linguistic applications.

How to use it

Statistical Methods and How We are Using Them

These are the methods most common in Cognitive Linguistics:

- Chi-square
- Fisher Test
- T-test and ANOVA
- Correlation
- Regression
- CART
- Mixed Effects
- Cluster Analysis
- Correspondence Analysis

I will give some examples of how these methods have been applied in Cognitive Linguistics.

These are just some examples of how statistics are being used in cognitive linguistics. There is plenty of room for experimentation.
Chi-square test

When to use the chi-square test:

You have a matrix with two types of categories and a count for each cell in the matrix. You want to know whether there is a relationship between the two types of categories.

Example: Is there a relationship between the choice of с- (сглупить) vs. -ну (чихнуть) and verb classes in Russian?
Verb classes that prefer -ну

-ай

-auty

non-prod

1. conj

-*ђ

(звать >) Зевнул

(лизать >) Лизнула

(свистеть >) Свистнула
Verb classes that prefer *c*

- **-ова**
  - (малодушествовать >) Смалодушествовал

- **-и**
  - (грубить >) Сгрубил!

- **-*ѣй**
  - (робеть >) Сробела?
Chi-square:
Finding out whether there is a significant difference between distributions

Illustration: Is there a relationship between semelfactive markers and verb classes in Russian?

Result: chi-squared = 269.2249, df = 5, p-value < 2.2e-16
Cramer’s V = 0.83

CAVEATS: chi-square 1) assumes independence of observations; 2) requires at least 5 expected observations in each cell.
Semantic Profiles: “Empty” prefixes in Russian

Olga
Lyashevska

Big Questions:

What is the relationship between form and meaning? ➔ ...between prefixes and meanings of verbs?

Are there any “empty” forms? ➔ Are prefixes empty as claimed?

<table>
<thead>
<tr>
<th>Imperfective base</th>
<th>Prefixed perfective</th>
</tr>
</thead>
<tbody>
<tr>
<td>советовать ‘advise’</td>
<td>посоветовать ‘advise’</td>
</tr>
<tr>
<td>варить ‘cook’</td>
<td>сварить ‘cook’</td>
</tr>
<tr>
<td>писать ‘write’</td>
<td>написать ‘write’</td>
</tr>
<tr>
<td>твердеть ‘harden’</td>
<td>затвердеть ‘harden’</td>
</tr>
<tr>
<td>греметь ‘thunder’</td>
<td>прогреметь ‘thunder’</td>
</tr>
</tbody>
</table>
Semantic Profiles: “Empty” prefixes in Russian

Operationalization:

Semantic profiling: relationship between meanings (semantic tags) and forms

Distribution of Russian verb prefixes vs. semantic tags

Data:

382 verbs with “empty” prefixes from the Exploring Emptiness database (http://emptyprefixes.uit.no/index.php), semantic tags independently assigned in the Russian National Corpus (http://ruscorpora.ru/)

Statistics:

Chi-square, Cramer’s V effect size, Fisher Test
Distribution of prefixes/semantic classes

<table>
<thead>
<tr>
<th></th>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND &amp; SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td>po-</td>
<td>11</td>
<td>62</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>s-</td>
<td>23</td>
<td>11</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>na-</td>
<td>31</td>
<td>3</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>za-</td>
<td>47</td>
<td>22</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>pro-</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>51</td>
</tr>
</tbody>
</table>

Chi-squared = 248.0058, df = 12, p-value < 2.2e-16, Cramer's V = 0.465
chi-square = 248, df = 12, \( p = 2.2e-16 \); Cramer’s V effect-size = 0.46
More examples for chi-square:
The English Ditransitive

English Ditransitive
(1) a. John told a story to Mary. (prepositional dative construction)
   b. John told Mary a story. (ditransitive construction)

Alternating verbs take both constructions:
   tell, give, show, send, sell, bring, read, lend...
Non-alternating verbs take only prepositional dative construction:
   explain, whisper, transfer, return, entrust, deliver, present, repeat...

Question: How do children learn that the non-alternating verbs do not use the ditransitive construction, since there is no negative evidence?
Chi-square: Stefanowitsch 2011

Research question:
English ditransitive: does the ungrammatical ditransitive get preempted when the child gets as input the prepositional data in contexts that should prefer the ditransitive?

Data:
British Component, International Corpus of English (ICE-GB) sentences with prepositional dative construction, 50 sentences per verb

Factors:
verb class (alternating vs. non-alternating) vs. givenness (referential distance); syntactic weight (# words); animacy

Result: not significant; no support for preemption
Chi-square: Goldberg 2011

Research question:
Same as Stefanowitsch 2011, plus: are the alternative constructions really in competition?

Data:
Corpus of Contemporary American English
15000+ exx alternating verbs, 400+ exx non-alternating verbs

Factors:
verb class (alternating vs. non-alternating)
vs.
construction (prepositional dative vs. ditransitive)

Result: p<0.0001; 0.04 probability of prepositional dative for alternating verbs vs. 0.83 for non-alternating verbs; sufficient evidence for preemption
Research question:
Do bodily experience of paths vs. roads motivate metaphorical meanings?

Data:
Experiment + British National Corpus

Factors:
path vs. road
vs.
description of courses of action/ways of living vs. purposeful activity/political/
financial matters

Result: p<0.001; evidence that people’s understanding of their physical experiences with paths and roads also informs their metaphorical choices, making path more appropriate for descriptions of personal struggles, and road more appropriate for straightforward progress toward a goal.
Table 1. Proportion of responses to mental image questions in Study 1

<table>
<thead>
<tr>
<th>Question</th>
<th>Path</th>
<th>Road</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which is more likely to have obstacles along the way?</td>
<td>.58</td>
<td>.42</td>
</tr>
<tr>
<td>Which is more likely to be straight?</td>
<td>.16</td>
<td>.84</td>
</tr>
<tr>
<td>Which is more likely to go up and down?</td>
<td>.71</td>
<td>.29</td>
</tr>
<tr>
<td>Which is more likely to be wide?</td>
<td>.04</td>
<td>.96</td>
</tr>
<tr>
<td>Which is more likely to be paved?</td>
<td>.04</td>
<td>.96</td>
</tr>
<tr>
<td>Which is more likely to go through problematic terrain?</td>
<td>.87</td>
<td>.13</td>
</tr>
<tr>
<td>Which is more likely to take you to a specific destination?</td>
<td>.21</td>
<td>.79</td>
</tr>
<tr>
<td>Which is more likely to make you move fast?</td>
<td>.13</td>
<td>.87</td>
</tr>
<tr>
<td>Which is more likely to move you along in an aimless way?</td>
<td>.79</td>
<td>.21</td>
</tr>
<tr>
<td>Which is more likely for you to enjoy traveling?</td>
<td>.83</td>
<td>.17</td>
</tr>
<tr>
<td>Which is more likely for you to stop every now and then?</td>
<td>.83</td>
<td>.17</td>
</tr>
<tr>
<td>Which is more likely for you to be driving on?</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Which is more likely for you to be biking on?</td>
<td>.54</td>
<td>.46</td>
</tr>
<tr>
<td>Which is more likely for you to be moving along on foot?</td>
<td>1.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

Falck & Gibbs 2012: 
*path/road* vs. descriptions

Table 3. Results of corpus Study 2

<table>
<thead>
<tr>
<th>Target domain</th>
<th>Path</th>
<th>Road</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course of Action/Way of living</td>
<td>.58</td>
<td>.12</td>
</tr>
<tr>
<td>Purposeful activity</td>
<td>.18</td>
<td>.36</td>
</tr>
<tr>
<td>Development</td>
<td>.07</td>
<td>.00</td>
</tr>
<tr>
<td>Political/Financial</td>
<td>.06</td>
<td>.52</td>
</tr>
<tr>
<td>Computer/Mathematics</td>
<td>.08</td>
<td>.00</td>
</tr>
<tr>
<td>Other</td>
<td>.03</td>
<td>.00</td>
</tr>
</tbody>
</table>
Fisher test

When to use the Fisher test:

You have a matrix with two types of categories and a count for each cell in the matrix. You want to know whether a specific cell in the matrix deviates significantly from the overall distribution.

Example: We know there is a relationship between semantic class of the verb and choice of prefix in Russian, but precisely which combinations are significant?
chi-square = 248, df = 12, p = 2.2e-16; Cramer’s V effect-size = 0.4
**Fisher test:**
Finding out whether a value deviates significantly from the overall distribution

**Illustration:** There are 51 Natural Perfective verbs prefixed in *pro-* in the Russian National Corpus that have the semantic tag “sound & speech”. This exceeds the expected value, but is there a relationship between the prefix and the semantic class?

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a = value in the given cell</td>
<td>b = row total - value in the given cell</td>
<td>c = column total - value in the given cell</td>
<td>d = table total - value in the given cell - b - c</td>
</tr>
<tr>
<td>= 51</td>
<td>= 106 - 51</td>
<td>= 65 - 51</td>
<td>= 382 - 51 - 55 - 14</td>
</tr>
<tr>
<td></td>
<td>= 55</td>
<td>= 14</td>
<td>= 262</td>
</tr>
</tbody>
</table>

**NOTE:** You need to know the relationship between the observed and the expected value before you can do the Fisher test. If the observed value is greater than the expected value, you need to choose alternative="greater" to test the significance of finding a value this high or higher. If the observed value is less than the expected value, you need to choose alternative="less" to test the significance of finding a value this low or lower.

**CAVEAT:** Fisher test does not work well on large numbers and differences!!!
How can we use the Fisher Test to measure the relationship of each number to the rest of the matrix?

<table>
<thead>
<tr>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND&amp;SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Values</td>
<td></td>
<td></td>
</tr>
<tr>
<td>po-</td>
<td>11</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>s-</td>
<td>23</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>na-</td>
<td>31</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>za-</td>
<td>47</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>pro-</td>
<td>10</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Let’s look at pro- with sound & speech
First find the expected values!

<table>
<thead>
<tr>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND &amp; SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Values</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>po-</td>
<td>11</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>s-</td>
<td>23</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>na-</td>
<td>31</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>za-</td>
<td>47</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>pro-</td>
<td>10</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Expected value = 
(row sum x column sum) / total sum
In R:

\[ \text{ExpProSS} = \frac{(10 + 4 + 0 + 51) \times (37 + 9 + 8 + 1 + 51)}{382} \]

The observed value (51) is GREATER than the expected value (18).
How to do this in R

Let's take the same example of pro- with verbs of sound & speech ("pross"):

```r
> pross = matrix(c(51, 55, 14, 262), ncol=2, byrow=TRUE)
> pross
 [,1] [,2]
[1,]  51  55
[2,]  14 262

> fisher.test(pross, alternative="greater")

Fisher's Exact Test for Count Data
data:  pross
p-value < 2.2e-16
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 9.531013 Inf
sample estimates:
odds ratio
 17.16661
```

This is the probability that you would get 51 examples or more if the actual distribution in the population was random.
Now for a more interesting example:

<table>
<thead>
<tr>
<th></th>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND&amp;SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Values</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>po-</td>
<td>11</td>
<td>62</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>s-</td>
<td>25</td>
<td>11</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>na-</td>
<td>31</td>
<td>3</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>za-</td>
<td>47</td>
<td>22</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>pro-</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>51</td>
</tr>
</tbody>
</table>

Does 11 = 11 = 11 here?
χ² = 248, df = 12, p = 2.2e-16; Cramer’s V effect-size = 0.4

ypo- + impact, as in покрыть
ypo- + behav, as in полениться
с- + changest, as in сгнить
Find the expected values!

Expected value =
(row sum x column sum) / total sum
Expected value = (row sum x column sum) / total sum

<table>
<thead>
<tr>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND&amp;SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td>po-</td>
<td>39</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>s-</td>
<td>21</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>na-</td>
<td>19</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>za-</td>
<td>22</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>pro-</td>
<td>21</td>
<td>17</td>
<td>9</td>
</tr>
</tbody>
</table>
What is the difference between the observed values and the expected values?

<table>
<thead>
<tr>
<th></th>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND&amp;SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td>po-</td>
<td>-28</td>
<td>30</td>
<td>-5</td>
<td>3</td>
</tr>
<tr>
<td>s-</td>
<td>2</td>
<td>-7</td>
<td>14</td>
<td>-9</td>
</tr>
<tr>
<td>na-</td>
<td>12</td>
<td>-15</td>
<td>9</td>
<td>-8</td>
</tr>
<tr>
<td>za-</td>
<td>24</td>
<td>3</td>
<td>-9</td>
<td>-19</td>
</tr>
<tr>
<td>pro-</td>
<td>-11</td>
<td>-13</td>
<td>-9</td>
<td>33</td>
</tr>
</tbody>
</table>
Now prepare the four values you need for the Fisher Test:

<table>
<thead>
<tr>
<th>IMPACT</th>
<th>CHANGEST</th>
<th>BEHAV</th>
<th>SOUND&amp; SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Values</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>po-</td>
<td>11</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>s-</td>
<td>23</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>na-</td>
<td>31</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>za-</td>
<td>47</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>pro-</td>
<td>10</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
    a &= \text{value in the given cell} \\
    &= 11 \\
\end{align*}
\]

\[
\begin{align*}
    b &= \text{row total - value in the given cell} \\
\end{align*}
\]

\[
\begin{align*}
    c &= \text{column total - value in the given cell} \\
\end{align*}
\]

\[
\begin{align*}
    d &= \text{table total - value in the given cell - b - c} \\
\end{align*}
\]
Here are the input values for the whole table:

<table>
<thead>
<tr>
<th></th>
<th>Fisher Test input values</th>
</tr>
</thead>
<tbody>
<tr>
<td>po-</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>11, b = 111</td>
</tr>
<tr>
<td>c</td>
<td>110, d = 150</td>
</tr>
<tr>
<td>prob a &lt;= 11</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>62, b = 40</td>
</tr>
<tr>
<td>c</td>
<td>59, d = 221</td>
</tr>
<tr>
<td>prob a &gt;= 62</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>11, b = 41</td>
</tr>
<tr>
<td>c</td>
<td>110, d = 220</td>
</tr>
<tr>
<td>prob a &lt;= 11</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>37, b = 69</td>
</tr>
<tr>
<td>c</td>
<td>84, d = 192</td>
</tr>
<tr>
<td>prob a &gt;= 37</td>
<td></td>
</tr>
<tr>
<td>s-</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>23, b = 99</td>
</tr>
<tr>
<td>c</td>
<td>43, d = 217</td>
</tr>
<tr>
<td>prob a &gt;= 23</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>11, b = 91</td>
</tr>
<tr>
<td>c</td>
<td>55, d = 225</td>
</tr>
<tr>
<td>prob a &lt;= 11</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>23, b = 29</td>
</tr>
<tr>
<td>c</td>
<td>43, d = 287</td>
</tr>
<tr>
<td>prob a &gt;= 23</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>9, b = 97</td>
</tr>
<tr>
<td>c</td>
<td>57, d = 219</td>
</tr>
<tr>
<td>prob a &lt;= 9</td>
<td></td>
</tr>
<tr>
<td>na-</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>31, b = 91</td>
</tr>
<tr>
<td>c</td>
<td>56, d = 224</td>
</tr>
<tr>
<td>prob a &gt;= 31</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>17, b = 35</td>
</tr>
<tr>
<td>c</td>
<td>51, d = 225</td>
</tr>
<tr>
<td>prob a &lt;= 8</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>8, b = 98</td>
</tr>
<tr>
<td>c</td>
<td>51, d = 225</td>
</tr>
<tr>
<td>prob a &lt;= 8</td>
<td></td>
</tr>
<tr>
<td>za-</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>47, b = 75</td>
</tr>
<tr>
<td>c</td>
<td>49, d = 231</td>
</tr>
<tr>
<td>prob a &gt;= 47</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>22, b = 80</td>
</tr>
<tr>
<td>c</td>
<td>70, d = 260</td>
</tr>
<tr>
<td>prob a &gt;= 22</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1, b = 51</td>
</tr>
<tr>
<td>c</td>
<td>70, d = 206</td>
</tr>
<tr>
<td>prob a &lt;= 1</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1, b = 105</td>
</tr>
<tr>
<td>c</td>
<td>70, d = 206</td>
</tr>
<tr>
<td>prob a &lt;= 1</td>
<td></td>
</tr>
<tr>
<td>pro-</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>10, b = 112</td>
</tr>
<tr>
<td>c</td>
<td>55, d = 205</td>
</tr>
<tr>
<td>prob a &lt;= 10</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>4, b = 98</td>
</tr>
<tr>
<td>c</td>
<td>61, d = 219</td>
</tr>
<tr>
<td>prob a &lt;= 4</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>0, b = 52</td>
</tr>
<tr>
<td>c</td>
<td>65, d = 265</td>
</tr>
<tr>
<td>prob a &lt;= 0</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>51, b = 55</td>
</tr>
<tr>
<td>c</td>
<td>14, d = 262</td>
</tr>
<tr>
<td>prob a &gt;= 51</td>
<td></td>
</tr>
</tbody>
</table>
Here are the p-values from the Fisher Test

<table>
<thead>
<tr>
<th></th>
<th>IMPACT</th>
<th>CHANGETST</th>
<th>BEHAV</th>
<th>SOUND&amp; SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher Test p-value (probability that Observed - Expected could be greater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>po-</td>
<td>[-] 4e-12</td>
<td>+] 6.3e-13</td>
<td>[-] 0.052</td>
<td>+] 0.23</td>
</tr>
<tr>
<td>s-</td>
<td>[+ 0.33</td>
<td>[-] 0.03</td>
<td>+] 6.3e-07</td>
<td>[-] 0.003</td>
</tr>
<tr>
<td>na-</td>
<td>[+ 0.0003</td>
<td>-] 5.8e-06</td>
<td>[+ 0.0006</td>
<td>[-] 0.005</td>
</tr>
<tr>
<td>za-</td>
<td>[+ 3.7e-11</td>
<td>+] 0.22</td>
<td>[-] 0.0001</td>
<td>[-] 1.9e-10</td>
</tr>
<tr>
<td>pro-</td>
<td>[-] 0.0009</td>
<td>-] 6.3e-06</td>
<td>[-] 2.8e-05</td>
<td>[+ 3e-21</td>
</tr>
</tbody>
</table>

Strongest attractions:
- pro- / sound&speech
- po- / changest
- za- / impact
- s- / behav

Strongest repulsions:
- po- / impact
- na- / changest
- za- / sound&speech
- pro- / changest
- pro- / behav

Other repulsions:
- s- / changest
- s- / sound&speech
- pro- / impact
- na- / sound&speech
- za- / behav

Other attractions:
- na- / impact
- na- / behav

Neutral:
- s- / impact
- po- / sound&speech
- za- / changest
Chi-square = 248, df = 12, p = 2.2e-16; Cramer’s V effect-size = 0.4
Collostructional analysis...

• “starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction”; takes into account the overall frequency of both lexemes and constructions

• **Collexeme**: a word that is attracted to or repelled by a construction
  – This article looks at collostructions in the following constructions:
    • **NP waiting to happen**
    • *cause* **NP**
    • **NP think nothing of Vgerund**
    • Sagent *V* Opatient/agent *into Vgerund*
    • **ditransitive**
    • progressive
    • imperative
    • past tense

We will look only at **NP waiting to happen** and **ditransitive**
How the authors did the Fisher Test

• Authors use the following four data points:
  – frequency of the collexeme in the construction
  – frequency of all other lexemes in the construction
  – frequency of the collexeme in all other constructions
  – frequency of all other lexemes in all other constructions

Only boldfaced data comes directly from corpus; all other data supplied via subtraction from totals.

Italicized data depends on calculation of frequency of all constructions in BNC = total # of verb tags (not an exact measure, on the order of 10M)
**NP waiting to happen**

- Interpreting the data: The p-values tell you how likely it is that one could get this distribution in a sample of this size from a potentially infinite distribution in which there was NO attraction/repulsion
- *accident* p=2.12E-34, which is: 0.000000000000000000000000000000212 (33 0s after .)
- *event* p=6.92E-02, which is: 0.0692 (not significant)
Ditransitive

• An entirely schematic construction
• Prototype and extensions according to Goldberg:
  – prototype: actual transfer (give, pass, hand)
  – A. Satisfaction of conditions (guarantee, promise)
  – B. Enabling (permit, allow)
  – C. Negation (refuse, deny)
  – D. Future (leave, bequeath)
  – E. Intention (bake, make, build)
  – F. Communication as transfer (tell, teach, fax)
  – G. Perceiving as receiving (show)
  – H. Directed action as transfer (blow (a kiss))
  – Exceptions: cost, charge, envy, forgive
Ditransitive, cont’d.

- *give* is by far most attracted to the construction
- Strong polysemy of the construction – extended uses are strongly represented in top 15 collexemes:
  - A: *offer, owe, promise*; B: *allow*; C: *deny*; D: *grant*; E: *earn*; F: *tell, teach*; G: *show*; H: (none); Exception: *cost*
What Stefanowitsch and Gries conclude:

• Advantages of collostructional analysis
  – Improved description (lexicography, pedagogy)
  – Support for construction grammar
    • “If syntactic structures served as meaningless templates waiting for the insertion of lexical material, no significant associations between these templates and specific verbs would be expected”
  – Implications for psycholinguistics, acquisition
    • collostructional analysis identifies the most prototypical collostructions
So what’s wrong here?

• Fisher Test was designed for small numbers, all on approximately the same order of magnitude
• The “d” cell is not an exact measure (and note that the “c” value is dependent on “d”) and it is many orders of magnitude greater than the “a” and “b” values
• The p-value for a Fisher Test is not technically a measure
• For many collexemes, there is only one example
• Criticism from Baayen, Bybee, and Schmid & Küchenhoff
ANOVA = “analysis of variance”

When to use ANOVA:
You have two (or more) sets of numerical scores obtained under different conditions. You want to know whether there is a difference between the sets of scores, whether the different conditions make a difference.

Example: How do Russians perceive marginal factitive verbs (like осерьёзнить, увкуснить) in comparison with standard verbs and nonce verbs?
More about variance and ANOVA

Variance is a measure of the shape of a distribution in terms of deviations from the mean.

ANOVA divides the total variation among scores into two groups, the within-groups variation, where the variance is due to chance vs. the between-groups variation, where the variance is due to both chance and the treatment effect (if there is any).

The F ratio has the between-groups variance in the numerator and the within-groups variance in the denominator.

If F is 1 or less, the inherent variance is greater than or equal to the between-groups variance, meaning that there is no treatment effect.

If F is greater than 1, higher values show a greater treatment effect and ANOVA can yield p-values to indicate significance.

ANOVA can also handle multiple variables, for example priming vs. none and male vs. female and show whether each variable has an effect (called a main effect) and whether there is an interaction between the variables (for example if females respond even better to priming).
Experimental design by Anna Endresen

**STANDARD WORDS**

10 o + 10 y

**MARGINAL WORDS**

10 o + 10 y

**NONCE WORDS**

10 o + 10 y

Стандартные слова, которые могут быть сохранены в памяти, а не генерированы на лету.

Ненормальные слова, которые не могут быть сгенерированы и не существуют (потому что они не соответствуют фонетическим законам или не основаны на продуктивных морфологических паттернах).

Границные слова, которые генерируются некоторыми говорящими и могут быть поняты/приняты некоторыми говорящими.
Experiment: score-assignment test

The task: Evaluate the marked word using one of the statements.

Давно пора как-то оприличить наше общение более мягкими выражениями.
‘It’s high time we made our interaction respectable by using kinder statements.’

- 5 points - Это совершенно нормальное слово русского языка.
  ‘This is an absolutely normal Russian word’
- 4 points - Это слово нормальное, но его мало используют.
  ‘This word is normal, but it is rarely used’
- 3 points - Это слово звучит странно, но, может быть, его кто-то использует.
  ‘This word sounds strange, but someone might use it’
- 2 points - Это слово звучит странно, и его вряд ли кто-то использует.
  ‘This word sounds strange and it is unlikely that anyone uses it’
- 1 point - Этого слова в русском языке нет.
  ‘This word does not exist in the Russian language.’
ANOVA RESULTS:
F = 546, df = 2,
p-value < 2.2e-16

Marginal Verbs
MAX = 479
MEAN = 286.4
MIN = 169
stand dev = 67
variance = 4446

Standard Verbs
MAX = 605
MEAN = 595
MIN = 549
stand dev = 15
variance = 235

Nonce Verbs
MAX = 223
MEAN = 183.4
MIN = 150
stand dev = 19
variance = 360
ANOVA: Dąbrowska et al. 2009

Research question:
Do speakers perform as well on unprototypical examples of LDDs as on prototypical ones?
(LDD = long-distance dependency)
Prototypical LDD: What do you think the funny old man really hopes?
Unprototypical LDD: What does the funny old man really hope you think?

Data: Experiment

Factors:
construction (declarative vs. question)
prototypical vs. unprototypical
age

Result: Both construction (p = 0.016) and prototypicality (p = 0.021) were found to be main effects, but not age. Significant interaction between construction and age (p = 0.01). Support for usage-based approach, according to which children acquire lexically specific templates and make more abstract generalizations about constructions only later, and in some cases may continue to rely on templates even as adults.
The acquisition of long-distance dependencies

Dąbrowska et al. 2009

Table 5.  Mean number (standard deviation) of correctly repeated sentences (study 2, focused scoring)

<table>
<thead>
<tr>
<th>Condition</th>
<th>5-year-olds (SD)</th>
<th>6-year-olds (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototypical question</td>
<td>1.35 (1.46)</td>
<td>1.47 (1.22)</td>
</tr>
<tr>
<td>Unprototypical question</td>
<td>1.24 (1.25)</td>
<td>1.00 (0.94)</td>
</tr>
<tr>
<td>Prototypical declarative</td>
<td>0.71 (1.05)</td>
<td>1.53 (1.17)</td>
</tr>
<tr>
<td>Unprototypical declarative</td>
<td>0.47 (0.87)</td>
<td>1.00 (1.20)</td>
</tr>
</tbody>
</table>

Main effect of construction (F = 6.47, p = 0.016)
Main effect of prototypicality (F = 5.82, p = 0.021)
Interaction construction/age (F = 7.51, p = 0.010)
Correlation

When to use correlation:

You have one set of items, and each item has two numerical values associated with it, one for Factor A and one for Factor B. You want to know whether there is a relationship between the two numerical values, i.e., a relationship between Factor A and B.

Example: Is the use of the possessive suffix (NPx), which is being lost in North Saami, better preserved with words of high frequency?
Correlation:
Finding significant relationships among values

Correlation

\[ r = +1 \] for a perfect positive correlation
\[ r = 0 \] for no correlation
\[ r = -1 \] for a perfect negative correlation.

CAVEATS:

1) assumption of linear relationship
2) correlation does not imply causation
SCATTERPLOTS & CORRELATION

Correlation - indicates a relationship (connection) between two sets of data.

- Strong positive correlation
- Weak positive correlation
- Strong negative correlation
- Weak negative correlation
- Moderate negative correlation
- No correlation
Anscombe’s quartet: These four plots yield the SAME correlation values
An ongoing language change: NPx is being replaced by ReflN

Two examples from Elle Márjá Vars’ novel Kátjá

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

ReflN (analytic construction with)
(1b) Kátjá... ollii
    Kátjá.NOM reach.IND.PRET.3S REFL GEN
    ‘Kátjá... got to her room’

Is this language change affected by frequency?
Are high frequency words less vulnerable to this change?
North Saami:
No evidence that high frequency helps to retain NPx
$r = -0.14, p = 0.0001, 95\%$ confidence interval: $-0.2 \quad -0.07$
Research Question:
Are backgrounded constructions islands; are they hard to extract in LDDs?

Data: Experiment
“difference score” measures to what extent a clause is an island = difference in acceptability between extraction in questions (Who did Pat stammer that she liked) and declarative (Pat stammered that she liked Dominic)
“negation test” measures to what extent clause is assumed background = rating that She didn’t think that he left implies He didn’t leave.

Factors: difference score vs. negation test

Result: Mean negation test score was a highly significant negative predictor of mean difference score; $r = -.83$, $p = 0.001$
from 71 participants. A scatterplot of this correlation is shown in Figure 3.

This analysis revealed that the mean negation test score was a highly significant (negative) predictor of mean difference score ($r = -0.83$, $p = 0.001$), accounting for over two thirds of the observed variance ($R^2 = 0.69$).

The correlation of $-0.83$ is strikingly high, as perfect correlations ($\pm 1$) are almost non-existent when distinct measures are used. Separate measures of the same thing, e.g., mean length of utterance (MLU) at 28 months, have been found to correlate in the .75–.80 range (Bates and Goodman 1997).

5.4. Any role for subjacency?

The subjacency account clearly does not predict the pattern of results found in the present study. In particular, subjacency does not predict any distinctions based on the semantic class of the verbs involved without...

---

Figure 3. Correlation between difference scores (dispreference for question scores) and negation test scores
Regression:  
Finding significant relationships among values

Regression builds upon correlation (the regression line is a correlation line), so it inherits all the caveats of correlation.

Regression is useful when you have found (or suspect) a relationship between a dependent variable and an independent variable, but there are other variables that you need to take into account

Dependent variable = the one you are trying to predict
Independent variables = the ones that you are using to predict the dependent one
The Locative Alternation in Russian: Svetlana Sokolova

**Theme-object construction**

грузить сено на телегу

**Goal-object construction**

грузить телегу сеном

Variables: VERB (prefixes), (passive) PARTICIPLE, REDUCED

- **VERB:** unprefixed грузить or prefixed: на
gрузить, загрузить, по

- **PARTICIPLE:**
  - Theme-object: сено гружено на телегу
  - Goal-object: телега гружена сеном

- **REDUCED:**
  - Theme-object: грузить сено
  - Goal-object: грузить телегу
The Locative Alternation in Russian

RIVAL FORMS: the two constructions, theme-object vs. goal-object

DEPENDENT VARIABLE: CONSTRUCTION: theme-object vs. goal-object

INDEPENDENT VARIABLES:

VERB:
zero (for the unprefixed verb грузить) vs. на-
vs. за- vs. по-

PARTICIPLE:
yes vs. no

REDUCED:
yes vs. no

DATA: 1920 sentences from the Russian National Corpus
Optimal model: CONSTRUCTION~VERB+REDUCED +PARTICIPLE+VERB*PARTICIPLE

The model estimates how the log of the number of theme constructions divided by the log of the number of goal constructions depends on the predictors. The coefficient of the estimate (Coeff.) is POSITIVE if the combination of factors predicts more theme constructions, but NEGATIVE if they predict more goal constructions.
Optimal model: CONSTRUCTION~VERB+REDUCED+PARTICIPLE+VERB*PARTICIPLE

<table>
<thead>
<tr>
<th>Obs</th>
<th>LR chi2</th>
<th>R2</th>
<th>C</th>
<th>Dxy</th>
<th>gamma</th>
<th>Brier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920</td>
<td>1738.47</td>
<td>0.796</td>
<td>0.964</td>
<td>0.928</td>
<td>0.945</td>
<td>0.076</td>
</tr>
<tr>
<td>871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1049</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2e-08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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 |

The model estimates how the log of the number of theme constructions divided by the log of the number of goal constructions depends on the predictors. The coefficient of the estimate (Coeff.) is POSITIVE if the combination of factors predicts more theme constructions, but NEGATIVE if they predict more goal constructions.
CART: Tree & forest

- **Classification and regression tree** (CART) uses recursive partitioning to yield a classification tree showing the best sorting of observations separating the values for the dependent variable
  - optimal algorithm for predicting an outcome given the predictor values
- **Random forest** uses repeated bootstrap samples drawn with replacement from the dataset such that in each repetition some observations are sampled and serve as a training set and other observations are not sampled, so they can serve for validation of the model
  - predictor variables are also randomly removed from repetitions, making it possible to measure variable importance
Tree & forest: CART

Includes intercept of logistic regression model: VERB = na, PARTICIPLE = no
What a CART tree means

• A CART tree can literally be understood as an optimal algorithm for predicting an outcome given the predictor values.

• 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, since it is a generalization over a set of instances.

• For example, node 11 is a generalization over 169 examples in which finite (non-participial) unprefixed (zero) forms of Russian ‘load’ in full (not reduced) constructions show a strong tendency (over 80%) for theme use.
Validation and measurement of variable importance in CART

- A CART random forest analysis uses repeated bootstrap samples drawn with replacement from the dataset such that in each repetition some observations are sampled and serve as a training set and other observations are not sampled, so they can serve for validation of the model.
Tree & forest: variable importance
Regression: Diessel 2008

Research question:
Does the linear order of clauses reflect the order of the reported events such that adverbial clauses reporting prior events are more likely to precede the main clause, whereas adverbial clauses reporting posterior events are more likely to follow the main clause? Is a speaker is more likely to produce After I fed the cat, I washed the dishes than I washed the dishes after I fed the cat?

Data: ICE-GB

Factors:
dependent variable: position of adverbial clause (initial vs. final)
independent variables: conceptual order (iconicity), meaning (conditional, causal), length, and syntactic complexity

Result: All variables except syntactic complexity are significant. Meaning is significant only for the positioning of conditional once- and until-clauses, and length is significant only for once- and until-clauses.
only 22.3 percent of the simultaneously occurring -clauses are preposed. Leaving aside the one posterior -clause, a \( \chi^2 \)-analysis revealed a significant association between linear structure and conceptual order (\( \chi^2 = 14.25, \) \( \text{df} = 1, \) \( p < 0.001 \)), confirming the hypothesis that clause order is iconic.

Like -clauses, after- and before-clauses tend to occur at the end of a complex sentence. As can be seen in Table 3, there are 151 final and only 33 initial after- and before-clauses in the data. Of the initial subordinate clauses, 27 are introduced by after and only 6 are introduced by before. A \( \chi^2 \)-analysis revealed a significant association between clause order and iconicity.

Chi-squared = 14.25, \( \text{df} = 1, \) \( p < 0.001 \), but more factors need to be considered.

Figure 2. Clause order and iconicity
Diessel 2008

Regression analysis was used to predict the position of the adverbial clause (i.e., initial or final) from the following set of predictors: conceptual order (i.e., iconicity), meaning, length, and syntactic complexity. Figure 3 shows the research design.

Conceptual order and syntactic complexity were coded as dichotomous variables: adverbial clauses denoting a prior event were distinguished from adverbial clauses denoting a posterior or simultaneously occurring event, and simple adverbial clauses consisting of a single clause were distinguished from complex adverbial clauses containing another subordinate clause. Meaning was coded as a discrete variable with three levels: (i) purely temporal, (ii) temporal with an implicit conditional meaning, and (iii) temporal with an implicit causal or purposive meaning. Finally, length was coded as a continuous variable, measured by dividing the number of words in the adverbial clause by the total number of words in the complex sentence.

For all features, intercoder reliability was at least 95 percent.

2.2.2. Results.

Table 5 shows the raw frequencies of the categorical predictors, i.e., conceptual order, complexity, and meaning, and Figure 4 shows the histograms of the continuous predictor, relative length (i.e., the ratio of adverbial clause/complex sentence), for final and initial temporal clauses.

Figure 3. Research design
The regression coefficients indicate the direction of change induced by a particular predictor: positive values (which correspond to odds ratios larger than 1.0) indicate that the predictor variable increases the likelihood of the adverbial clause to precede the main clause; negative values (which correspond to odds ratios smaller than 1.0) indicate that the predictor variable decreases the likelihood of the adverbial clause to precede the main clause. The Wald $\chi^2$-values and the associated levels of significance indicate that the predictor variables (conceptual order, meaning, and length) are significant. The odds ratios show the change in odds for an adverbial clause to be placed in initial position. For instance, the odds ratio for conceptual order indicates that for adverbial clauses denoting a prior event the odds of preceding the main clause are 6.7 times larger than the odds for adverbial clauses denoting a posterior or simultaneous event. The two final columns show the lower and upper boundaries of the confidence intervals for the odds ratios (cf. Backhaus et al. 2005: 475–476).

Note that conceptual order and conditional meaning increase the likelihood of the adverbial clause to precede the main clause (compared to posterior/simultaneous temporal clauses with purely temporal meaning), whereas a causal/purposive meaning and an increase in length decrease the likelihood of the adverbial clause to precede the main clause (compared to purely temporal clauses that are shorter). Note also that conceptual order, i.e., the encoding of a prior event, is the strongest predictor for the initial occurrence of a temporal adverbial clause.

Since the positioning of temporal adverbial clauses varies with the subordinating conjunction (see above), I also computed regression models for individual types of temporal clauses. Specifically, I developed three separate logistic regression models for when-clauses, after- and before-clauses, and once- and until-clauses using the same stepwise procedure as in the model described above (Table 7 in the Appendix provides a summary of the frequency data). Interestingly, while conceptual order had a significant effect on the positioning of all temporal clauses (when: $\chi^2 = 12.149$, Table 6.

<table>
<thead>
<tr>
<th>Factor</th>
<th>reg. coef.</th>
<th>Wald $\chi^2$</th>
<th>df</th>
<th>p</th>
<th>odds ratio</th>
<th>lower CI</th>
<th>upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual order</td>
<td>1.902</td>
<td>73.69</td>
<td>1</td>
<td>0.001</td>
<td>6.70</td>
<td>4.34</td>
<td>10.35</td>
</tr>
<tr>
<td>Meaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. causal/purpose</td>
<td>–2.775</td>
<td>7.27</td>
<td>1</td>
<td>0.007</td>
<td>0.06</td>
<td>0.01</td>
<td>0.469</td>
</tr>
<tr>
<td>b. conditional</td>
<td>1.364</td>
<td>31.20</td>
<td>1</td>
<td>0.001</td>
<td>3.91</td>
<td>2.42</td>
<td>6.31</td>
</tr>
<tr>
<td>Length</td>
<td>–1.343</td>
<td>7.39</td>
<td>1</td>
<td>0.001</td>
<td>0.19</td>
<td>0.06</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Look at regression coefficient (first column): Positive values indicate that the predictor variable increases the likelihood of the adverbial clause to precede the main clause. Negative values indicate that the predictor variable decreases the likelihood of the adverbial clause to precede the main clause.
**Mixed effects**: Adding individual preferences into a regression model

**Mixed effects builds upon regression**: In an ordinary regression model, all effects are fixed effects. A mixed effects model combines fixed effects with random effects.

**Fixed effect**: an independent variable with a fixed set of possible values

**Random effect**: represent preferences of individuals sampled randomly from a potentially infinite population

Mixed effects models combine fixed effects and random effects in a single regression model by measuring the random effects and making adjustments so that the fixed effects can be detected.
When do we need mixed effects models?

Mixed effects models are used when individual preferences interfere with obtaining independent observations. Individuals with preferences need to be represented as random variables.

Some examples of random variables:

Subjects in an experiment will have different individual preferences, and different measures for baseline performance (e.g., reaction time)

Authors in a corpus will have different individual preferences for certain words, collocations, and grammatical constructions

Verbs in a language can have different individual behaviors with respect to ongoing changes and distribution of inflected forms
Research question:
In Dutch, English loanwords like *backpacker* co-exist with native equivalents like *rugzakker*. What factors contribute to the success/failure of loanwords?

Data: Dutch newspaper corpora

Factors:
dependent variable: success rate of English loanword
independent variables as fixed effects: length, lexical field, era of borrowing, luxury vs. necessary borrowing, concept frequency, data of measurement, register, region
independent variable as random effect: concept expressed

Result:
Two strongest main effects: a negative correlation between concept frequency and the success of an anglicism, and a significantly lower success rate for borrowings from the most recent era (after 1989) than from earlier eras. Interactions: concept frequency is a factor only when the anglicism is also the shortest lexicalization, and the difference between luxury and necessary borrowings is strongest in the 1945-1989 era.
Measuring variation in the success of anglicisms appears that speech economy clearly restricts concept frequency as a factor: if an anglicism is the shortest lexicalization, concept frequency does not have any impeding effect on the success of the anglicism.

For the interaction between era of borrowing and the distinction between luxury and necessary anglicisms, it is again important to remember that reverse estimate

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.890</td>
<td>1.171</td>
<td>5.032</td>
<td>0.000</td>
</tr>
<tr>
<td>log(concept frequency)</td>
<td>-0.830</td>
<td>0.149</td>
<td>-5.586</td>
<td>0.000</td>
</tr>
<tr>
<td>luxury anglicism</td>
<td>0.046</td>
<td>0.312</td>
<td>0.147</td>
<td>0.883</td>
</tr>
<tr>
<td>era borrowing-age group (helmert) 1</td>
<td>1.554</td>
<td>0.323</td>
<td>4.806</td>
<td>0.000</td>
</tr>
<tr>
<td>era borrowing-age group (helmert) 2</td>
<td>-0.216</td>
<td>0.216</td>
<td>-0.999</td>
<td>0.318</td>
</tr>
<tr>
<td>lexical field-sports and recreation</td>
<td>-0.652</td>
<td>0.727</td>
<td>-0.897</td>
<td>0.370</td>
</tr>
<tr>
<td>lexical field-legal moneymaking</td>
<td>-1.725</td>
<td>0.672</td>
<td>-2.566</td>
<td>0.010</td>
</tr>
<tr>
<td>lexical field-social life</td>
<td>-1.720</td>
<td>0.662</td>
<td>-2.598</td>
<td>0.009</td>
</tr>
<tr>
<td>lexical field-deviance</td>
<td>-1.695</td>
<td>0.658</td>
<td>-2.574</td>
<td>0.010</td>
</tr>
<tr>
<td>speech economy-anglicism not shortest</td>
<td>-0.097</td>
<td>1.508</td>
<td>-0.038</td>
<td>0.970</td>
</tr>
<tr>
<td>log(concept frequency): speech economy-anglicism not shortest</td>
<td>0.781</td>
<td>2.04</td>
<td>3.829</td>
<td>0.000</td>
</tr>
<tr>
<td>luxury anglicism: era borrowing-age group 1</td>
<td>-2.118</td>
<td>0.323</td>
<td>-6.551</td>
<td>0.000</td>
</tr>
<tr>
<td>luxury anglicism: era borrowing-age group 2</td>
<td>-0.209</td>
<td>0.221</td>
<td>-0.947</td>
<td>0.344</td>
</tr>
</tbody>
</table>

Table 8: Interaction model (mixed)

Fig. 1: Interaction concept frequency and speech economy

Zenner, Speelman, & Geeraerts 2012
Helmert coding was used for era of borrowing, which means that we first compare the words borrowed up to 1945 with words borrowed between 1945 and 1989, and that next, we take these two groups together and compare them to words borrowed after 1989. Within each age group, we set the behavior of luxury and necessary anglicisms side by side.

Figure 2 shows the first comparison. In the main model presented in the previous section, the difference in success between luxury and necessary loanwords was only borderline significant. Apparently, this is partly due to the interaction with era of borrowing: the difference in success in our present day corpus between luxury and necessary anglicisms is not significant for the oldest group of loanwords, but becomes highly significant for words borrowed between 1945 and 1989. Also, we see that luxury anglicisms follow the structural hypothesis: older loanwords are more successful, as they have had more time to establish themselves in the receptor language. In contrast, necessary anglicisms behave differently: older necessary anglicisms are less successful in our present day corpus than younger necessary anglicisms. Upon closer inspection, this pattern is highly intuitive. Luxury anglicisms are introduced as an alternative for an established lexicalization of a given concept, and hence need time to become a worthy competitor of this established lexeme. The older the luxury anglicism, the more time it has had to establish itself as alternative lexicalization, the more success it will have in our corpus. For necessary anglicisms, the situation is reversed. Given the restrictions to our dataset, all necessary anglicisms studied in this paper have at least one competitor. In these cases, the necessary anglicism is the first and most established lexicalization of the concept, which can gradually lose ground to the upcoming alternative(s). The older the necessary anglicism, the more time

Fig. 2: Interaction era of borrowing and luxury/necessary: age group 1
Cluster analysis:
Finding out which items are grouped together

Cluster analysis is useful when you want to measure the distances between items in a set, given that you have an array of datapoints connected to each item.

In hierarchical cluster analysis, squared Euclidean distances are used to calculate the distances between the arrays of data.

Other methods to achieve similar means include multidimensional scaling and correspondence analysis.
Cluster analysis:  
Janda & Solovyev 2009

Research question:  
Can we measure the distance among near-synonyms?

Data: Russian National Corpus and Biblioteka Moškova

Factors:  
Near-synonyms for ‘happiness’ and ‘sadness’  
(Preposition)+ case constructions

Result: Each noun has a unique constructional profile, and there are stark differences in the constructional profiles of words that are unrelated to each other. For the two sets of synonyms in this study, only six grammatical constructions are regularly attested. The study shows us which nouns behave very similarly as opposed to which are outliers in the sets. The clusters largely confirm the introspective analyses found in synonym dictionaries, giving them a concrete quantitative dimension, but also pinpointing how and why some synonyms are closer than others.
Chi-square = 730.35, df = 30, p < 0.0001, Cramer’s V = 0.305
‘Sadness’
Hierarchical Cluster
Correspondence analysis:
Another way to find out which items are grouped together

Correspondence analysis also measures the “distances” between row and column vectors. It constructs a multidimensional space and then represents the two most significant dimensions in a plot.

Example: Use grammatical profiles of Old Church Slavonic verbs as input and ask correspondence analysis to sort them.
Grammatical profiles of OCS verbs: 
*tvoriti* ‘make’ and *jęti* ‘take’

- To obtain the grammatical profile of a verb:
  - Count up attestations for each subparadigm
  - Calculate distribution across subparadigms in terms of percentages

<table>
<thead>
<tr>
<th></th>
<th>Aorist</th>
<th>Imperative</th>
<th>Imperfect</th>
<th>Infinitive</th>
<th>Past participle</th>
<th>Present</th>
<th>Present participle</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>tvoriti</em> ‘make’</td>
<td>0 0%</td>
<td>14 8%</td>
<td>12 7%</td>
<td>23 13%</td>
<td>0 0%</td>
<td>99 57%</td>
<td>26 15%</td>
</tr>
<tr>
<td><em>jęti</em> ‘take’</td>
<td>25 28%</td>
<td>7 8%</td>
<td>0 0%</td>
<td>10 11%</td>
<td>20 22%</td>
<td>28 31%</td>
<td>0 0%</td>
</tr>
</tbody>
</table>
Correspondence analysis of 129 OCS verbs: 
96% equivalent to Dostál’s sorting of verbs

Verbs with - Factor 1 values ≠ imperfective verbs

Verbs with + Factor 1 values ≠ perfective verbs

Factor 1 is the most important factor. It sorts all verbs into - vs. + values.

Factor 1 (39.7%)
Where Does the Quantitative Turn Lead Us? Opportunities and Dangers

Opportunities

• Discover structures in linguistic data
• Secure and maintain the status of linguistics as a science
• Establish best practices in quantitative approaches to theoretical questions
• joining forces with computational linguists to build better natural language processing and language technology applications
• Make a commitment to publicly archive data and statistical code: TROLLing
Take your linguistic data to the bank!

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 (Common Language Resources and Technology Infrastructure in the EU), 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.

Authors of scholarly works around the world are welcome to deposit their data in TROLLing, along with citations of their publications. Conversely, authors can reference their data by citing their TROLLing posts in their articles.

Visit us at
http://opendata.uit.no/
http://site.uit.no/trolling/

“In the age of Big Data, the creation of a general repository of datasets and statistical models for linguistic research is a welcome development. It will stimulate more research and new analyses.” -- Maria Polinsky, Director of the Polinsky Language Sciences Lab at Harvard University

“TROLLing will revolutionize research in linguistics and drive the discipline forward: making data publicly available significantly reduces the risk of bogus results, avoids duplication of efforts and facilitates large-scale analysis of meticulously annotated datasets.” -- Dagmar Divjak, Reader, Russian and Slavonic Studies, University of Sheffield

“TROLLing is crucial for the field of linguistics as it takes the next steps towards becoming more empirical. For the first time, it will be possible for researchers to deposit their primary linguistic data (the foundation for all research) in a central freely accessible on-line repository so that colleagues around the world have access to the same data. This invaluable resource will promote on-going academic exchange on an empirical basis.” -- Hans Boas, Professor, Department Germanic Studies and the Department of Linguistics, University of Texas at Austin

“TROLLing is exactly what our field needs - with the potential to become the most useful data resource in linguistics.” -- Marit Westergaard, Professor, Center for Advanced Study of Theoretical Linguistics, UiT The Arctic University of Norway

“I would like to recommend that scholars deposit their data at TROLLing. I strongly believe that sharing of data and methods for analysis can play a key role in the growth of cognitive linguistics. It will be beneficial for the community of linguists to have a single searchable repository rather than having data scattered about in many places.” -- Laura Janda, Professor, Center for Advanced Study of Theoretical Linguistics, UiT The Arctic University of Norway
Where Does the Quantitative Turn Lead Us? Opportunities and Dangers

Dangers

• Triviality and fractionalization of the field
• Fancy equipment and sophisticated software should not receive more attention than relevant linguistic principles
• “Cargo cult science”:
  – linguists perform empty rituals of calculations in hopes of conjuring up publishable results
  – shallow studies that do not advance the field, mere number-crunching without any real linguistic or theoretical goal
  – “arms race” to show off complex “black box” models that do not enhance understanding
Where Does the Quantitative Turn Lead Us? Opportunities and Dangers

More Dangers

• Substitution of “quantitative” for “empirical” and “scientific” in the minds of researchers:
  – marginalization of many of the traditional endeavors of linguists
  – erosion of the core of our field, linguistic description and theoretical interpretation, which are also the source for research hypotheses

• With Big Data, it is too easy to find trivial results
  – Language phenomena are never, ever, ever random (Kilgarriff 2005)
Where Does the Quantitative Turn Lead Us? Opportunities and Dangers

More Dangers

- **Fraud:**
  - fudged data or analyses, motivated by pressure to publish in prestigious journals
  - transparency (TROLLing) can help to avoid fraud

- **Invasion of privacy:**
  - Major corporations (Google, Amazon, Apple, Facebook), spyware operations and state governments, have access to massive quantities of human language data and are developing mining techniques
  - 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
Conclusion

- The quantitative turn is a hugely positive 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 application of statistical models to linguistic data
- We need to retain a humble attitude of respect for our venerable qualitative and theoretical traditions
- We need qualitative and theoretical insights now more than ever in order to make sense of all the data at our command because those insights are the wellspring for hypotheses and the yardstick for interpretation of results