Laura A. Janda (Ed.)
Cognitive Linguistics: The Quantitative Turn

Cognitive Linguistics: The Quantitative Turn

The Essential Reader

edited by

Laura A. Janda

De Gruyter Mouton

ISBN 978-3-11-033388-6
e-ISBN 978-3-11-033525-5

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Cover image: Caroline Sale/Flickr/Getty Images
Printing: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Table of contents

The papers are reprinted with permission. They appear in their original form except for the following changes: adjusted pagination, removal of DOI. The article by Zenner, Speelman and Geeraerts features the latest De Gruyter typography.

Publication sources

Anatol Stefanowitsch
2011 Constructional preemption by contextual mismatch: A corpus-linguistic investigation. *Cognitive Linguistics* 22(1): 107–129.

Adele E. Goldberg
2011 Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics 2*2(1): 131–153.

Marlene Johansson Falck and Raymond W. Gibbs, Jr.
2012 Embodied motivations for metaphorical meanings. *Cognitive Linguistics* 23(2): 251–272.

Anna L. Theakston, Robert Maslen, Elena V. M. Lieven and Michael Tomasello
2012 The acquisition of the active transitive construction in English: A detailed case study. *Cognitive Linguistics* 23(1): 91–128.

Beate Hampe
2011 Discovering constructions by means of collostruction analysis: The English Denominative Construction. *Cognitive Linguistics* 22(2): 211–245.

Stefan Th. Gries
2011 Phonological similarity in multi-word units. *Cognitive Linguistics* 22(3): 491–510.

Ewa Dąbrowska, Caroline Rowland and Anna Theakston
2009 The acquisition of questions with long-distance dependencies. *Cognitive Linguistics* (20)3: 571–597.

Holger Diessel
2008 Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19(3): 465–490.

Eline Zenner, Dirk Speelman and Dirk Geeraerts
2012 Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics* 23(4): 749–792.

Laura A. Janda and Valery D. Solovyev
2009 What constructional profiles reveal about synonymy: A case study of Russian words for SADNESS and HAPPINESS *Cognitive Linguistics* 20(2): 367–393.

Quantitative methods in *Cognitive Linguistics*: An introduction*

Laura A. Janda

1. Introduction

Both the field of cognitive linguistics as a whole and the journal *Cognitive Linguistics* have taken a quantitative turn in recent years. The majority of conference presentations, articles, and books in our field now involve some kind of quantitative analysis of language data, and results are often measured using statistical methods. This does not mean that other types of contributions (theoretical, introspective) are in any way less welcome in cognitive linguistics, but the quantitative turn in our field is now a fact to be reckoned with.

This book presents some of the people and the statistical methods that have played a leading role in defining the current state of the art in cognitive linguistics, focusing specifically on researchers and methods that have appeared prominently in our journal in the past five years. The ten articles gathered here showcase recent achievements of the following individuals (plus coauthors) who have made quantitative contributions repeatedly in the pages of *Cognitive Linguistics*: Ewa Dąbrowska, Holger Diessel, Dirk Geeraerts, Raymond W. Gibbs, Adele E. Goldberg, Stefan Th. Gries, Beate Hampe, Laura A. Janda, Elena V. M. Lieven, Caroline Rowland, Anatol Stefanowitsch, Anna L. Theakston, and Michael Tomasello. Collectively these researchers have done much to shape contemporary practice in statistical analysis in cognitive linguistics, addressing issues at all levels of language, including phonology, morphology, syntax, semantics, acquisition, sociolinguistics, etc. Other significant leaders in quantitative analysis in our field include Ben Ambridge, Antti Arppe, Harald Baayen, Jeremy Boyd, Steven Clancy, William Croft, Dagmar Divjak, Dylan

* I would like to thank: the CLEAR (Cognitive Linguistics: Empirical Approaches to Russian) group (Anna Endresen, Julia Kuznetsova, Anastasia Makarova, Tore Nesset, and Svetlana Sokolova), Ewa Dąbrowska, Ludmila Janda, and Francis Tyers for their comments on this article; and the University of Tromsø and the Norwegian Research Council for their support of this research.

Glynn, Martin Hilpert, Willem B. Hollmann, Irraide Ibarretxe, Vsevolod Kapatsinski, Maarten Lemmens, John Newman, Sally Rice, Dominiek Sandra, Hans-Jörg Schmid, Doris Schönefeld, Dan Slobin, Dirk Speelman, Javier Valenzuela, and Stefanie Wulff.

The methods represent those that have proven useful and versatile in linguistic analysis: chi-square, Fisher test, binomial test, ANOVA, correlation, regression, and cluster analysis. Each of these methods, with their advantages and limitations, will be discussed in turn and illustrated by highlights from the articles in this collection. Additional methods that are gaining popularity and may become part of standard use are also presented in that section, and suggestions are made for best practices in the management and sharing of data and statistical code.

Based on a study of articles published in *Cognitive Linguistics*, the time period 2008–2012 emerges as a noticeably different era in our history. As described in section 2, the year 2008 marks the quantitative turn for our journal, and the past five years have been substantially different from the two decades that preceded them. It seems unlikely now that we will ever turn back, so this is an appropriate time to take stock of the situation, how it came about, and what it means for our future.

2. How we got here, where we are now, what challenges lie ahead

There are many reasons why cognitive linguists have become increasingly attracted to quantitative methods. A combination of theoretical and historical factors has facilitated the quantitative turn.

Unlike most other modern theories of linguistics, cognitive linguistics is a usage-based model of language structure (Langacker 1987: 46; 2008: 220). In other words, 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 (Glynn 2010: 5–6). In this sense, cognitive linguistics has always been a "data-friendly" theory, with a focus on the relationship between observed form and meaning. Linguistic theories that aim instead to uncover an idealized linguistic competence have less of a relationship to the observation of usage, though there are of course notable exceptions. For overviews of the use of corpus linguistics across various theoretical frameworks, see Gries 2009 and Joseph 2004.

Even the question of what constitutes data in linguistics is controversial, and largely dependent upon the theory that one uses. Many researchers in formal theories refer to constructed examples and individual intuitions as data, while others prefer to use corpus attestations or observations from acquisition

or experiments. While introspection does play an important role in linguistic analysis, reliance on introspection to the exclusion of observation undermines linguistics as a science, yielding claims that can be neither operationalized nor falsified. It may seem attractive to assume that language is a tightly ordered 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. Instead we find complex relationships among factors that motivate various trends in the behavior of linguistic forms. A usage-based theorist views language use as the data relevant for linguistic analysis, and this gives cognitive linguistics a natural advantage over other theories in applying quantitative methods, an advantage that we have been steadily realizing and improving upon over the past quarter century.

It is crucial to distinguish between the linguist's own intuitions about data (or intuitions solicited from a few colleagues) and judgment experiments involving the systematic study of the intuitions of naive informants under experimental conditions (which is a legitimate scientific method that normally involves quantitative analysis). There is a difference between these two uses of introspection in that the former does not yield reliable, replicable results, whereas the latter can. The linguist's intuitions present numerous problems in that there are disagreements between linguists (cf. Carden and Dietrich 1980, Cowart 1997); intuitions about mental phenomena are often inaccurate (Gibbs 2006); and last but not least, linguist's intuitions may be biased by their theoretical commitments (Dąbrowska 2010).

Computational linguists have made remarkable progress in developing technological applications for language in recent years. In terms of digital manipulation of language data, on the whole they have more experience than we typically find among cognitive linguists. The goals of computational linguists and cognitive linguists of course differ, but this opens up considerable opportunity for collaboration. We bring to the table a strong focus on foundational theoretical issues. Joining forces with computational linguists can help us to realize the potential that digital resources provide for investigating linguistically interesting questions. And hopefully computational linguists will inspire us to put our research results to work in developing language technology.

Recent history has impacted the practice of linguistics through the development of language corpora and statistical software. Today we have access to balanced multi-purpose corpora for many languages, often containing hundreds of millions of words, some even with linguistic annotation. Modern corpora of this kind became widespread only a little over a decade ago, but have already become the first resource many linguists turn to when investigating a phenomenon. At approximately the same time, statistical software likewise became widely available, in particular "R", which is open-source and supports

UTF-8 encoding for various languages. Thus we now have access to both vast quantities of data and the means to explore its structure.

Cognitive linguists are on the leading edge in terms of implementing data analysis in the context of a theoretical framework and we may well have a historic opportunity now to show leadership not only within cognitive linguistics, but in the entire field of linguistics. We can establish best practices in quantitative approaches to theoretical questions. Best practices should include acknowledgement of the most valuable kinds of statistical methods and significance measures, as well as public archiving and sharing of data and statistical code. This will help to move the field forward by providing standards and examples that can be followed. It is also a means of reducing the risk of fraud. Most academic fields in which researchers report statistical findings have experienced scandals involving fudged data or analyses, and current pressures to publish present an incentive to falsify results in hopes of impressing reviewers at a prestigious journal. Data sharing and best practices (see section 2.2) can help us to protect our field from this kind of dishonor.

2.1. *The quantitative turn in the pages of* Cognitive Linguistics

In this book I use the journal *Cognitive Linguistics* as a microcosm for the entire field, and here I present the quantitative turn as it has unfolded on our pages. Of course it would in principle be possible to undertake a comprehensive investigation, including other journals such as *Corpus Linguistics and Linguistic Theory*, and books such as Glynn and Fischer 2010, Gries and Stefanowitsch 2007, Schmid and Handl 2010, and Stefanowitsch and Gries 2007. However I justify this choice on the grounds that the journal gives us the most consistent longitudinal perspective available on this development.

I have surveyed all of the articles published in the journal *Cognitive Linguistics* from its inaugural volume in 1990 through the most recent completed volume in 2012. The numbers here represent the findings of this survey as an overview of the situation rather than a scientifically exact account. If we exclude review articles, book reviews, overviews, commentaries, replies, squibs, CLiPs (surveys of recent publications), and introductions to special issues, we find a total of 331 articles published in the journal in that interval. If we define a "quantitative article" as an article in which a researcher reports numbers for some kind of authentic language data, then we find 141 quantitative articles in that period, and they are distributed as shown in Figure 1.

In order to put all the data on the same scale, Figure 1 reports percentages of quantitative articles for each year. A thick line marks 50% to make this visualization clearer. On the basis of this distribution we can divide the history of *Cognitive Linguistics* into two eras, 1990–2007 – when most articles were not quantitative, and 2008–2012 – when most articles were quantitative.

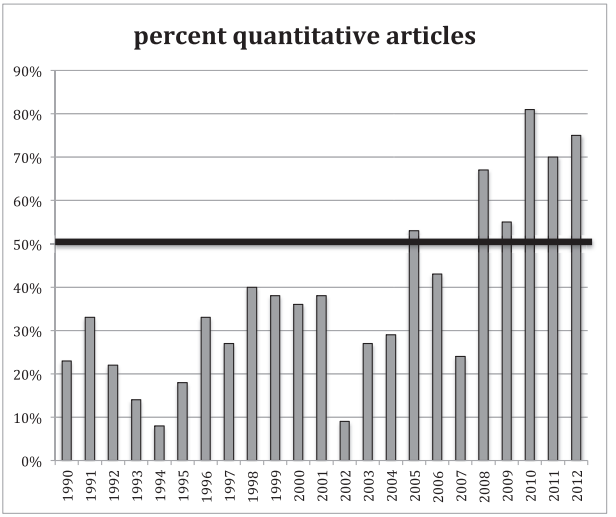

Figure 1. *Percent quantitative articles in* Cognitive Linguistics 1990–2012

In 1990–2007, twelve out of eighteen volumes had 20–40% quantitative articles. The lowest points were 1994, with one out of twelve articles, and 2002, with one out of eleven articles. 2005 reached in the other direction, with ten out of nineteen articles.

It is important to note that quantitative articles have always been with us; no year has ever been without quantitative studies. Three quantitative articles appeared already in the very first volume: Goossens 1990 (with a database of metaphorical and metonymic expressions), Delbecque 1990 (citing numbers of attestations in French and Spanish corpora), and Gibbs 1990 (presenting experimental results). However 2008 is the year in which we definitively crossed the 50% line, and it is unlikely that we will drop below that line again in the foreseeable future. Over half (75 out of 141 = 53%) of all quantitative articles published in *Cognitive Linguistics* have appeared in 2008–2012.

The majority of quantitative articles in our journal report corpus data (34%) or experimental data (48%) or a combination of the two (6%), and acquisition data (which can involve both corpus and experimental data) is also steadily represented (12%). 54 articles (38%) reported only raw and/or percent frequencies in the absence of any statistical test. The most popular statistical measure is by far the chi-square test (40 articles), but an accompanying effect size (Cramer's V) is reported only in 3 articles. The remaining measures that appear more than once are given here in descending order of frequency with the number of relevant articles (note also that some articles report several kinds of tests): ANOVA (26), t-test (13), correlation (11), regression (of various types,

also including both fixed and mixed effects models; 8), clustering (5), Fisher test (4), binomial test (2). Visualization of data was spotty in the first decade of the journal, with only four graphs appearing before 2000 (in Hirschberg and Ward 1991, Sanders et al. 1993, Sandra and Rice 1995, and Hudson 1997). Between 2000–2007 the number of graphs ranges from zero (in 2002 and 2004) to five (in 2005), but becomes frequent in 2008–2012 when half or more of the quantitative articles appear with graphs.

We can thus securely identify 2008–2012 as a distinct period in the history of *Cognitive Linguistics*. During this period quantitative analysis emerges as common practice, dominating the pages of our journal. The selection of articles, authors, and statistical models represented in this anthology are motivated by these observations. The purpose of this book is to explicitly acknowledge the norms that we are implicitly forging as a community. In the next subsection we consider what this means for our future.

2.2. *The road beyond the quantitative turn*

Now that we have started off down a path dominated by quantitative methods, it is worth asking ourselves where we are headed. We have much to look forward to, but some words of caution are also in order.

It is essential for the legitimacy of our field to secure and maintain the status of linguistics as a science. In applying quantitative measures we are developing linguistics as a discipline, following psychology and sociology in bringing the scientific method best known from the natural sciences to the fore. However, we face two challenges, one involving the relationship between introspection and observation and the other involving the archiving and sharing of data and code.

Although I maintain that exclusive reliance on introspection can be problematic, especially in the presence of unfounded assumptions, it is important to remember that there always has been and always should be a place for introspection in linguistics. Our journal has always published both quantitative and non-quantitative articles, and there is no reason to expect that this should cease to be the case even after the quantitative turn. In other words, it is not the case that we are dealing with an S-curve in which a phenomenon was initially absent, there was an innovation, and then the innovation will necessarily reach 100% (cf. Blythe and Croft 2012). While it is not infallible as a method, introspection has a place in our field. There should be a healthy balance between introspection and observation in any scientific inquiry. Introspection is a source of inspiration for hypotheses, which are then tested via observation. When it comes to analysis, we need introspection again in order to interpret the results and understand what they mean for both theory and facts of language.

Introspection is irreplaceable in the descriptive documentation of language. In fieldwork a linguist interacts with speakers and posits the structure of a gram-

mar based on a combination of observations and insights. The foundational role of descriptive work and reference grammars is not to be underestimated, for without this background we would have no basis for stating any hypotheses about languages at all. Linguists who pursue quantitative methods should never forget that they stand with one foot on the shoulders of descriptivists. Although it is not strictly within the mission of *Cognitive Linguistics* to publish purely descriptive work, contributions that present a previously unknown language phenomenon as attested by authentic data (whether quantitative or not) are welcome on our pages.

The other foot of quantitative linguists should be on the shoulders of theorists. Whereas theory should of course be informed by data, theoretical advances owe much to introspection and are often presented without recourse to new findings or in the context of summaries over multiple studies. It would be foolish to banish theoretical polemics from our journal and our field. Reducing our theoretical perspective would hinder our ability to pose linguistically interesting questions, both in quantitative and non-quantitative studies.

Both theoretical and descriptive components have long been common in the training of linguists, but now we should ask how much statistics should be added to our graduate programs and our professional expectations. The answer depends in part upon the goals of programs and individuals, however we have reached a point at which all programs should offer some quantitative component, and all linguists should have at least some passive statistical literacy. Relevant handbooks are available (King et al. 2010, Johnson 2008, Baayen 2008, Gries 2013, Cohen et al. 2003), and this book gives illustrations of how several statistical methods can be successfully applied to pertinent linguistic questions.

One important step we should take as a community is to make a commitment to publicly archive both our data and the statistical code used to analyze it. The goal should be to create an ethical standard for sharing data and code in a manner explicit enough so that other researchers can access the data and re-run the models. This can be done by creating designated websites for public access using standard and preferably open-source software. For example, Janda et al. 2013 presents a series of studies using chi-square, Fisher test, and logistic regression. Any visitor to this site http://emptyprefixes.uit.no/book.htm can find all the relevant data in csv (comma-separated-values) files and open-source annotated R scripts. The website gives instructions on how to access R, run the scripts, and interpret the results, and explains how the datasets are organized and what the values stand for. The annotations in the scripts describe every step needed to set up the model for analysis. Similarly Baayen et al. forthcoming presents a series of case studies comparing the results yielded by logistic regression, classification and regression trees and random forests, and naive discriminative learning, and all of the data and code are already housed at this site: http://ansatte.uit.no/laura.janda/RF/RF.html. De Gruyter Mouton has the facil-

ity to archive supplementary materials associated with the works it publishes, and this can include data, code, graphics, and sound files. To date, no author in *Cognitive Linguistics* has yet made use of this opportunity, perhaps because it is not widely known. I strongly encourage linguists to publicly archive data and code, for it has important implications for the advancement of the field and for its integrity.

Publicly archived linguistic data and statistical code have great pedagogical value for the community of linguists. As anyone who has attempted quantitative analysis of linguistic data knows, one of the biggest challenges is to match an appropriate statistical model to a given dataset. Access to examples of datasets and corresponding models will help us all over the hurdle of choosing the right models for our data. We can help each other and bring our whole field forward much more efficiently if we pool our experience. I think it is quite misguided to be overprotective of one's data and code. This does not need to be a race with winners and losers; it can instead be a collective learning experience. A shared pool of data and code will also have a normative effect on the use of statistics in linguistics, further clarifying the trends that I try to identify in this book.

While transparency does not guarantee integrity, it does make some kinds of fraud easier to detect, and it always improves the quality and depth of scholarly communication. It has long been the case in natural sciences, medicine, and psychology that authors are routinely requested to submit their data along with their manuscripts when seeking publication in a journal. I expect similar requests to become more common in connection with submissions to *Cognitive Linguistics* in the future. In many cases funding agencies also require researchers to share their data with any colleagues who ask for it (this is particularly common in medicine), and it is not unthinkable that such conditions could be placed upon grant funding for linguistics as well. For the researcher, both public archiving and submission of data can be accomplished via the same task, preparing annotations for datasets and code that facilitate the work of peer reviewers and colleagues.

Lastly I would like to make an appeal for elegance in analysis. We should not engage in an arms race to find out who can show off the most complex statistical models. 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. Sometimes the structure of the data dictates a more complex model, but some models carry with them the problem that they are well understood only by the statisticians who developed them. Overuse of "black box" methods will not enhance the ability of linguists to communicate with each other. Recall from section 2.1 that over one-third (38%) of the quantitative studies published in *Cognitive Linguistics* did not use any statistical test at all: the goals of the authors were achieved by reporting frequencies and ratios that are easy for everyone to interpret. I refer the reader also to Kuznetsova 2013 for several exam-

ples of how to find linguistic insights in quantitative studies without invoking heavy statistical machinery.

3. Methods

A research question must of course come first, along with some kind of hypothesis. Next the researcher can consider what kind of data can be collected in order to address the question. The design of a study inevitably involves some compromise between accessing an ideal dataset and the limitations of what is realistically obtainable. Already in the design, decisions must be made about what to collect, how to code it, etc. and these decisions will impact the choice of the statistical model. The choice of a model is very much dependent upon the structure and type of data involved. Ideally the researcher will be familiar with some possible statistical models and take this into consideration when designing a study.

This section presents the articles in this anthology organized according to the statistical models they use. First some information is given about each model and then the relevant articles are discussed, with focus on the theoretical linguistic issue that the author has posed, the type of data examined, and reasons why the given model is appropriate. The purpose of this discussion is not to serve as a textbook on applying statistical models, but rather to illustrate how the models are being used and provide sufficient orientation for readers who want to gain confidence in reading and understanding such articles.

3.1. Chi-square: Finding out whether there is a significant difference between distributions

Stefanowitsch 2011, Goldberg 2011, Falck and Gibbs 2012, Theakston et al. 2012

The chi-square test is very common and popular, so it is worth giving some detail about how it works, what it means, and what kinds of data it is appropriate for. This test is usually appropriate when you have a matrix of data and you want to explore the relationship between two variables. One factor is assigned to the rows and another to the columns. The matrix must have at least two rows and two columns, each column and row represents a given value for a variable, and each cell in the matrix has a number of observations. The chi-square test evaluates the distribution of observations in relation to what would be expected in a random distribution given the totals for the rows and the columns. If the distribution is very uneven, and this unevenness cannot be attributed to chance, then there is probably a relationship between the two variables. The chi-square test gives a p-value (probability value) that tells you the likelihood that you

could get a distribution that is as uneven as the one observed (or even more extreme) if your observations are a sample from a (potentially infinite) population of data points in which there is no relationship between the factors and no difference in distribution. A very low number indicates a low likelihood that you could get this distribution by chance, and this is a measure of statistical significance. Usually the largest p-value that is acknowledged as significant is 0.05 (often signaled by one asterisk *), while more significant values are $p < 0.01$ (**) and $p < 0.001$ (***).

Here is a concrete example to illustrate how the chi-square test can be used. Dickey and Janda (2009) wanted to challenge the traditional definition of allomorphy, suggesting that allomorphy should be recognized as a gradient rather than all-or-nothing phenomenon because there are cases where the distribution of morpheme variants fails the classical criterion of complementary distribution, but displays a strong relationship akin to allomorphy. To this end, Dickey and Janda presented the distribution of Russian verbs derived with two semelfactive markers, the suffix *-nu* and the prefix *s-*, across the morphological classes of verbs. This distribution supports their argument that *-nu* and *s-* behave much like allomorphs. Here is the raw data:

Table 1. *Distribution of semelfactive markers across Russian verb classes from Dickey and Janda 2009*

		verb classes					
		-aj	non-prod	-*ě	-ova	-i	-*ěj
semelfactive	-nu	185	57	20	17	16	0
markers	s-	1	0	1	18	38	36

The two variables are the semelfactive markers and the verb classes. There are 185 verbs in the *-aj* class with the *-nu* marker, 57 verbs in the non-productive class with the *-nu* marker, etc. The chi-square test returns these values for this distribution: chi-squared = 269.2249, df = 5, p-value < 2.2e-16. 2.2e-16 is a very low number (0.00000000000000022), in fact it is the lowest p-value that R reports for the chi-square test, so it tells us there is almost no chance that we could have taken a sample with this distribution (or one even more extreme) from a hypothetically infinite population of verbs in which there is no relationship between the two variables. In other words, this result is statistically significant (***).

In addition to the chi-square test, Dickey and Janda report the effect size (Cramer's V), which measures the chi-square value against the total number of observations. Cramer's V ranges from 0 to 1, and it is generally acknowledged that 0.1 is the minimum threshold for a reportable though small effect size, 0.3 is the threshold for a moderate effect size, and 0.5 is the threshold for

a large effect size. The Cramer's V in this study is 0.83, indicating a large effect. While effect sizes are not yet commonplace in linguistic studies, I strongly encourage all researchers to measure effect sizes when reporting p-values, especially when the number of observations is large (thousands or more). In a large dataset the chi-square test will find even infinitesimal differences in distribution to be statistically significant. For instance, Janda and Lyashevskaya 2011 is a study of the distribution of verb forms across aspect and aspectual markers for nearly 6 million observations from the Russian National Corpus. The p-values for all distributions were found to be significant, but only the p-value for the aspectual difference (perfective vs. imperfective) was confirmed by a robust Cramer's V effect size of 0.399, whereas effect sizes for differences in aspectual markers (prefixes vs. suffixes) were 0.076 and 0.037, an order of magnitude too small to be considered reportable. Thus a measure of effect size can be used to distinguish between effects that are worth our attention and ones that are not.

Some words of caution are in order with regard to the use of the chi-square test. Note that the input for this test must always be raw frequencies, not percentages. The chi-square test has a lower limit on the quantity of data needed: no cell in a matrix should have an expected value of five or less. While there are some lower values in Table 1, the expected values (based on the row and column totals) for all cells are greater than 5. If there is a large matrix and/or very uneven distribution of data, this will result in a paucity of data for chi-square, which gives error ("unreliable") messages in R. The chi-square test is also founded upon an assumption of independence of observations. In other words, no two observations should be related to each other, for example by having the same source. For corpus data this usually means that one should not have more than one example from any given author in order to avoid biasing the data according to individual preferences of authors, unless one is sampling within a population of utterances or using the author/utterer as one of the variables; see the discussion of Theakston et al. 2012 below. Note also that mixed effects regression models are designed to deal with such factors; see section 3.6.

Stefanowitsch 2011

The linguistic issue addressed is: How do children learn that a given syntactic structure, such as the English ditransitive, is ungrammatical for some verbs in the absence of negative evidence? Does the ungrammatical ditransitive get preempted when the child gets as input the prepositional dative in contexts that should otherwise prefer the ditransitive (see Pinker 1984)? Stefanowitsch uses corpus data (from the British Component of the International Corpus of English = ICE-GB) to address this issue, and analyzes this data by means of chi-square tests. The first variable in all tests is verb class, which can be either

alternating (appearing in both the ditransitive and the prepositional dative constructions, like *read* and *tell*) or non-alternating (appearing only in the prepositional dative constructions, like *explain* and *mention*). The second variable was selected from a set of factors relevant to the information structure of these verbs. There were three such variables coded with reference to both the recipient and the theme: givenness (referential distance), syntactic weight (number of orthographic words), and animacy. Stefanowitsch extracts 50 sentences each for alternating and non-alternating verbs; all examples are of the prepositional dative construction. In nearly all tests of the first variable in relation to one selected from the second set of variables, the chi-square test yields a p-value too high to suggest statistical significance. Further tests show that the differences between verbs belonging to the same class are often greater than other differences. Stefanowitsch concludes that preemption is an unlikely explanation since corpus data do not support the relevant inferences.

Goldberg 2011

Goldberg addresses the same question as Stefanowitsch, namely whether preemption gives sufficient evidence for learners of English to understand that some verbs can only take the prepositional dative construction, as opposed to other verbs that can appear in both the prepositional dative construction and the ditransitive construction. For Goldberg the most important issue is whether the alternative constructions are actually in competition, and for this reason her data reflects use of both constructions, not just the prepositional dative. Goldberg argues that Stefanowitsch's sample of data (100 sentences, all of the prepositional dative construction) is too small and too restricted, and that the hypothesis is also too narrow. Goldberg takes a different sample from a corpus (Corpus of Contemporary American English = COCA), with over 15,000 examples of alternating verbs and over 400 examples of non-alternating verbs (the latter are of overall lower frequency), representing both the prepositional dative construction and the ditransitive construction with a pronominal recipient and a full NP for the theme. Goldberg shows that the probability of using the prepositional dative (ratio of prepositional dative/ditransitive uses) is very low (0.04 on average) for alternating verbs, but very high (0.83 on average) for non-alternating verbs. Goldberg compares the overall distribution of the two constructions across the two classes of verbs using the chi-square test. The first variable is the same as we see for Stefanowitsch: the class of verb as alternating vs. non-alternating. The other variable is the construction as prepositional dative vs. ditransitive. The p-value reported for this chi-square test is $p < 0.0001$, indicating a very significant result. Goldberg thus argues that the different distributions are indeed sufficient to give learners evidence for preemption. Several additional arguments are also adduced, such as frequency, experimental

data (reported in other studies on use of adjectives) and a variety of other alternative hypotheses involving more complex sets of competing constructions and lexemes.

Falck and Gibbs 2012

Falck and Gibbs present a combination of experimental and corpus data addressing the question of how bodily experiences motivate metaphorical meanings. Their study focuses on differences between the use of the English words *path* and *road* both in reference to physical experience and to metaphorical understanding of other kinds of experience. Twenty-four undergraduates at UC Santa Cruz participated in an experiment by answering fourteen questions about their experiences of paths vs. roads. This questionnaire showed that the subjects expected paths to be more likely to involve problematic terrain and aimless pedestrian movement, whereas roads were judged more likely to be wide, paved and straight and traveled by vehicles. A chi-square test was performed for each question with one variable being the choice of *path* vs. *road*, and the other relating to each given question (e.g. more likely to have obstacles vs. not). The result for one question was significant at the $p < 0.05$ level, the result for one other question was significant at the $p < 0.01$ level, the results for ten questions were significant at the $p < 0.001$ level, and the results for two questions (involving presence of obstacles and which would be more used for biking) were not significant. These experimental results were compared to dictionary entries and to corpus examples. 1000 examples each for *path* and *road* were extracted from the *British National Corpus* (BNC) and the Pragglejaz Metaphor Identification Procedure was used to identify and classify all metaphorical uses in the sample. At an abstract level all of the metaphors were similar in that they used travel as the source domain, and had various life experiences as the target domain. However, at a more fine-grained level, the distribution of metaphorical uses was very different for the two words. While *path* was often used to describe courses of action and ways of living, *road* (with overall far fewer metaphorical uses) was more likely to be associated with purposeful activity and political or financial matters. A second set of chi-square tests, with the same first variable, but different second variables involving choice of metaphorical types, showed these results to be significant at the $p < 0.001$ level. Falck and Gibbs take this as 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.

Theakston et al. 2012

A twelve-month sample (from age 2;0 to 3;0) of acquisition data representing both the output of a child (Thomas) and the input of his mother was analyzed to track the use of SVO transitive constructions. The question motivating this research is whether children have preliminary biases favoring learning the expression of prototypical transitive events or they instead gradually build up competence based on previous use of the same verbs in SV and VO constructions. Chi-square tests are used in this study to show that there are significant differences across several types of distributions. For example, it is shown that Thomas's use of SVO constructions are different from his mother's use. When the first variable is Thomas vs. his mother (input) and the second variable is the form of the subject or object (pronoun/omitted, noun, or proper noun), the difference is significant at $p < 0.001$ at 2;6. Overall Thomas shows a propensity for expressing subjects as proper nouns and objects as pronouns (*it*), contrary to the input pattern of using pronouns for subjects and noun phrases for objects, which conforms to preferred argument structure. During the second half of the study phase (2;7 to 3;0) the proportional use of SVO (vs. SV vs. VO) is significantly different from month to month for most of the sample, with $p < 0.01$. However, even though these changes bring Thomas closer to the adult model, even at 3;0 his proportional use of SVO is significantly different from that of his mother, with $p < 0.001$. Thomas also shows more use of "Old" verbs (attested before 2;7) than "New" verbs (attested at or after 2;7) in the SVO construction ($p = 0.006$ at 2;9 and $p = 0.017$ at 2;11). Theakston et al. take this as evidence that children do not come to the acquisition task equipped with preliminary biases, but instead acquire the SVO construction via a complex process that involves different stages of development for different verbs (those acquired early vs. those acquired late), gradual abstraction of patterns, and integration of various semantic types.

3.2. Fisher test: Finding out whether a value deviates significantly from the overall distribution
Hampe 2011

The Fisher test is useful to evaluate the relationships among variables when data is very unevenly distributed and/or sparse. Like the chi-square test, the Fisher test takes into account the overall distribution of values in a matrix, and yields p-values. The difference is that a Fisher test can be applied to each cell, where it can tell us the probability that each value could deviate even more from the expected value, given the overall distribution. If the expected value is less than the observed value, we calculate a right-sided p-value, which indicates the probability that we would get this many items or more in the cell given the

overall distribution of items. If the expected value is greater than the observed value, we calculate a left-sided p-value, which indicates the probability that we would get this many items or fewer in the cell given the overall distribution of items. In order to compute the Fisher test probability, four values are needed. These values relate the value in the cell to the sum for the row, the sum for the column, and the sum for the entire table.

This website http://emptyprefixes.uit.no/semantic_eng.htm gives a link to a Fisher Test calculator and shows how the Fisher test is applied to data relating the use of Russian verbal prefixes to the semantic tags assigned for verbs in the Russian National Corpus (Janda et al. 2013). For example, 51 verbs are found with the prefix *pro-* and the semantic tag "sound & speech", there are a total of 65 verbs prefixed by *pro-* (the column total), there are a total of 106 verbs with the "sound & speech" tag, and there are a total of 382 verbs in the study. Table 2 shows the values used for computing the Fisher test probability for *pro-*/"sound & speech":

Table 2. *An example of values used as input for a Fisher test (boldfaced)*

a = (value in the given cell) = **51**	b = (row total) – (value in the given cell) = 106 − 51 = **55**
c = (column total) – (value in the given cell) = 65 − 51 = **14**	d = (table total) – (value in the given cell) = 382 − 51 = **331**

Based on this array of values we can apply the Fisher test and we calculate a right-sided p-value of 5.7e-25 (an extremely low number, with twenty-four zeroes after the decimal point followed by the digits 57). This value indicates a strong relationship between the prefix *pro-* and the semantic tag "sound & speech" since there is an extremely small chance that we could get 51 or more verbs in that cell if we took another sample of the same size from a potentially infinite population of verbs in which there was no relationship between the prefix and the semantic class.

Hampe (2011) turns her attention to the family of complex transitive argument structures. She observes that whereas both generativists and cognitivists have paid considerable attention to both the caused-motion construction with a prepositional phrase (*John pushed Sally into the hole*) and the resultative construction with a predicate adjective (*John hammered the metal flat*), there has been less focus on a similar construction with a predicate noun phrase that Hampe calls the "denominative construction" (*Schoolmates called John a hero*). Hampe argues that the denominative construction deserves a place among complex transitive constructions and seeks support in corpus data from

the ICE-GB. Following Stefanowitsch and Gries (2003, 2005), Hampe uses the Fisher test in collostruction analysis to measure the attraction of lexemes to constructions. She reports the p-values log-transformed on base 10, so that the number corresponds to the number of decimal places in the p-value (0.001 = 3, for example). Thus higher log-transformed numbers reflect lower p-values and stronger attractions, and Hampe arranges lists of verbs that appear in the relevant constructions according to their attraction to each construction. This results in distinctive lists that are very different from each other, supporting Hampe's claim that the denominative construction should be recognized as a construction in its own right. Hampe also finds that the denominative construction is attracted to the active voice, whereas the resultative construction is attracted to the passive voice.

3.3. Exact Binomial test: Finding out whether the distribution in a sample is significantly different from the distribution of a population Gries 2011

Like the chi-square test and the Fisher test, the exact binomial test gives a p-value that reflects the chance that you could get a given distribution in a sample. The difference is that this test is appropriate when you have values for only two alternatives, provided that you also know the relative frequency of the two alternatives in the total population. In other words, if you know that there are ten white balls and ten red balls in an urn, you can calculate the chance of drawing three red balls when four total balls are drawn (and replaced each time) as p = 0.3125, or nearly a one in three chance (this example adapted from Gries 2001: 497–498). The exact binomial test is handy when you know the overall frequency of two alternatives in a corpus and want to know whether your sample differs significantly from what one would expect given the overall distributions in the corpus. For example, one could use the exact binomial test to compare the frequency of a given lexeme in a certain context with its overall frequency in the corpus to see whether there is an association between the context and the word.

Gries (2011) investigates the hypothesis that phonological similarity as realized in alliteration contributes to the cohesiveness of idiomatic expressions. Is the alliteration we see in phrases like *bite the bullet* and *turn the tables* just a random fact or does alliteration play a significant role in the formation of idioms? Gries undertakes two studies to find evidence in support of his hypothesis. The first study involves 211 high-frequency fully lexically specified idioms with a verb and a direct object. These idioms include 35 alliterations like the two cited above, but many others without any alliteration, like *spill the beans*. Gries makes several computations of baseline frequencies involving all allowable initial phonemes in English and their occurrence in the ICE-GB corpus and uses the binomial test to show that the frequency of alliteration in lexically-specified

idioms is significantly above chance, with all p-values <0.001. Gries' second study is of the partially lexically specified *way*-construction as in *wend one's way*, where the direct object *way* is specified, but the verb can vary (since it can be replaced by *make*, *find*, and many other verbs). The question here is whether the verbs that fill the unspecified slot also have a tendency to alliterate with *way*. Again Gries undertakes a series of calculations to determine relevant baseline measures in the ICE-GB corpus and uses the exact binomial test to show that the alliteration in the *way*-construction is highly significant, again with p-values < 0.001.

3.4. T-test and ANOVA: Finding out whether group means are significantly different from each other
Dąbrowska et al. 2009

In order to understand ANOVA, it is helpful to start by tackling the t-test on which ANOVA is based. The t-test is useful for determining whether distributions of scores, for example from psycholinguistic experiments, are indeed different from each other. Let's say that we do an experiment collecting word-recognition reaction times from two groups of subjects, one that is exposed to a priming treatment that should speed up their reactions (the test group), and one that is not (the control group). The mean scores of the two groups are different, but the distributions overlap since some of the subjects in the test group have reaction times that are slower than some of the subjects in the control group. Do the scores of the test group and the control group represent two different distributions, or are they really samples from a single distribution (in which case the difference in means is merely due to chance)? The t-test can answer this question by giving us a p-value.

The t-test can only handle a simple comparison of two groups. ANOVA takes the t-test to a further dimension by making it possible compare more than two groups or more than one variable across the groups. ANOVA stands for "analysis of variance", and to understand ANOVA, one must first come to terms with variance. Variance is a measure of the shape of a distribution in terms of deviations from the mean. Since the sum of the deviations from the mean in any distribution is necessarily zero (half of the deviations will be positive and half will be negative), variance is measured by summing the squared deviations (all of which are rendered positive) and dividing them by the number of scores in the distribution. The square root of the variance gives us the standard deviation of the distribution. What ANOVA does is to divide 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, so if the F value is 1 or less, the inherent variance is greater than or equal to the between-groups variance, meaning that there is no treatment effect. But 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).

Generative linguists account for long-distance dependencies (LDDs) such as *What*$_1$ *do you think* _____$_1$ *is in the box?* and *Who*$_1$ *did Mary hope that Tom would tell Bill that he should visit* _____$_1$? in terms of abstract syntactic representations and iterate-able WH movement operations. If speakers really have such representations, they should perform equally well on simple, ordinary examples as on ones that are complex and deeply embedded. However, in a study of the BNC spoken corpus Dąbrowska discovered that 67% of LDD questions follow the lexically specific templates *WH do you think S-GAP?* or *WH did you say S-GAP?*, where S-GAP is a subordinate clause with a missing constituent, and the majority of the remaining attestations are minimal variations on these patterns. In other words, spontaneously produced LDD questions are highly stereotypical and might best be accounted for by means of these two lexically specific templates than by abstract schemas. Dąbrowska et al. (2009) tested this hypothesis in experiments on both children and adults. The results of an initial experiment with children were ambiguous since they could have been influenced by different frequencies of words. The design of the experiment was adjusted and both children and adults were asked to repeat four examples each of four types of questions using all the same lexemes (here only one example of each is given):

Prototypical LDD question: *What do you think the funny old man really hopes?*
Prototypical declarative: *I think the funny old man will really hope so.*
Unprototypical LDD question: *What does the funny old man really hope you think?*
Unprototypical declarative: *The funny old man really hopes I will think so*.

The children were stratified according to age: about half of them were five-year-olds and half of them were six-year-olds. For the children the results were analyzed using a $2 \times 2 \times 2$ ANOVA with the first variable as construction (declarative, question), the second variable as prototypicality (prototypical, unprototypical), and the third variable as age (5-year-olds, 6-year-olds). Both construction ($p = 0.016$) and prototypicality ($p = 0.021$) were found to be main effects, but not age. However, there was a significant interaction between construction and age ($p = 0.01$); five-year-olds performed better on questions than declaratives,

but six-year-olds were equally good on both constructions. For adults a 2×2 ANOVA was used with the variables construction and prototypicality. Neither of the variables was significant as a main effect, but there was a significant interaction between construction and prototypicality ($p = 0.021$), suggesting that even adults make use of lexically specific templates for LDD questions, but not for declaratives. Overall, the results reported by Dąbrowska et al. indicate that children rely on lexically specific templates for both LDD questions and declaratives as late as age 6, and that even adults are more proficient with LDD questions that match these templates. These results support the 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.

3.5. Correlation and Regression: Finding significant relationships among values
Diessel 2008

Correlation refers to the degree of relationship between two variables, such that the greater the correlation, the better we are able to predict the value of one variable given the value of the other. Let's say, for example, that we want to explore the relationship between the corpus frequency of a word and reaction time in a word-recognition experiment. A likely outcome would be that there is a correlation, such that the higher the frequency of a word, the shorter the reaction time, and this relationship can be quantified as a coefficient. If this correlation exists, given the frequency of a word one would be able to use the coefficient to predict the reaction time, and conversely given the reaction time associated with a word one would be able to predict its frequency. There are two main ways to calculate correlation, also known as r, using Pearson's coefficient (which is appropriate for ordinary numerical scores) and Spearman's coefficient (which is appropriate for rank-ordered scores), and the two are very similar. Both involve calculations based on the deviations of individual data points from the mean and both yield measures that range from $r = +1$ (perfect positive correlation) to $r = 0$ (no correlation) to $r = -1$ (perfect negative correlation). In our example with frequency and reaction time we would expect to find a negative correlation since a higher value for frequency should give a lower value for reaction time. If the relationship is weak the value will be closer to zero, but if the relationship is strong it will be closer to -1. The value of the coefficient is an indication of how closely the data points come to approximating a straight line of best fit: if the data points follow a straight line the coefficient will be close to $+1$ or -1, but if the data points are scattered at random the coefficient will be close to zero.

Two caveats are important when using correlation. The first caveat is that the correlation coefficients assume that the relationship in question is linear, when

in fact there are infinitely many other possible kinds of relationships (with various curves and clumps of data points) and indeed even for any given *r* value there is an infinite number of distributions of data points that it might describe. While correlation is handy for data that is perhaps a bit scattered but otherwise reasonably well behaved, in more complex cases the correlation coefficient might hide more structure than reveals. In some cases various transformations of the data can correct for the problem of non-linearity.

The second caveat is that the presence of a correlation does not mean that there is any causal relationship involved. There might be a causal relationship, but it cannot be inferred from a correlation. So while it might be the case that high frequency causes low reaction times, this is not proved by a correlation. The correlation would be just as likely (or unlikely) to prove the opposite: that low reaction times cause high frequency. For a perspective from another domain, it has long been known that there is a strong positive correlation between the wealth of a country and its cancer rate, but it would be very strange to assert that money gives people cancer. This correlation is probably due to other variables that are related to both wealth and cancer, such as for example that people in wealthy countries live longer and thus have more opportunity to eventually get cancer, and that they also have more access to doctors who can diagnose cancer, etc. Similar hidden variables can also lurk in linguistic data.

While correlation is not used as a measure in the articles in this anthology, it is worth understanding for two reasons: one reason is that correlation is well-represented in recent articles in *Cognitive Linguistics* (see Ambridge and Goldberg 2008, Ambridge and Rowland 2009, Chandler 2010, Ghesquière and Van de Velde 2011, Akita 2012, and Kraska-Szlenk and Żygis 2012) and the other reason is that the line of best fit described by correlation is the basis for regression models.

The line of best fit is called the regression line, and the equation that locates that line is called the regression equation. Like the correlation coefficient, the regression equation can predict the value of one variable given the value of the other variable, but this regression equation fits the data exactly only when the correlation is perfect ($+1$ or -1). Because the correlation is generally not perfect, there is a difference between the predicted values and the actual values, and this difference is referred to as the "error". The standard error of estimate (which is a kind of standard deviation of the actual scores from the predicted scores) gives us a measure of how well the regression equation fits our data.

Because regression is based upon the same calculations as correlation, it also inherits the same drawbacks, namely that it assumes a linear relationship (which may or may not be true), and that it cannot tell us anything about causation. Regression models come in a variety of types and all involve the prediction of a dependent variable based upon one or more independent variables (also called predictor values). Ideally the independent variables should be independent not

just of the dependent variable, but also of each other (avoiding what is called collinearity). In logistic regression (named after the logistic function used to divide all values into a categorical choice between two levels) the dependent variable has only two values, and this is particularly useful for linguistic phenomena that involve a choice between two forms. For example, the locative alternation involves a choice between two constructions, the theme-object construction as in *load the boxes onto the cart*, and the goal-object construction as in *load the cart with boxes*. This website http://emptyprefixes.uit.no/constructional_eng.htm presents the data and R script for a logistic regression analysis of the locative alternation in Russian where the dependent variable is the construction (theme-object vs. goal-object) and the independent variables are the prefix on the verb, the status of the construction as full (with both theme and goal overt) vs. reduced, and the use of an active construction vs. a passive one (with a participle). (Note that multinomial extensions of logistic regression are also possible, allowing more than two choices.)

A regression analysis allows you to consider the relationship between an independent variable and a dependent variable, while making it possible to take into account the effects of additional independent variables. A regression model specifies the change in the group means when going from one variable level to another. The goal of a logistic regression model is to predict the probability that a given value (X, or alternatively, Y) for the dependent variable will be used. This is achieved by means of the logarithm of the odds ratio of X and Y. The odds ratio is the quotient of the number of observations supporting X and the number of observations supporting Y. This ratio is negative when the count for Y is greater than the count for X. It is zero when the counts are equal. It is positive when the counts for X exceed the counts for Y.

Like the chi-square test, the binomial test, and ANOVA, regression will also give you p-values. Usually there will be an overall p-value to indicate the significance of the data sample (the likelihood that we would find a sample with this strong a deviation from a random pattern or even stronger if there were no pattern at all in a potentially infinite population of examples), as well as p-values indicating the significance of each of the variables in the model. A series of other measures come with a regression model, among them r in a new guise as r^2 (often written as R^2), which indicates the amount of the variance that is accounted for by the model and its variables. Like r, the maximum limit for this measure is 1, and higher numbers indicate a better model. Another common measure is C, the index of concordance, which should have a value of 0.8 or higher if a model is performing well. Measures of the performance of the model are important because it is usually necessary to undertake some trial-and-error in fitting a model to the data, and each model has to be evaluated in order to arrive at the optimal one, while avoiding overfitting (see section 3.8). Usually this is done by first putting all of the variables (and interactions) into

the regression formula and then gradually trimming away variables that are not found to be significant, and chi-square, ANOVA, or AIC (Akaike Information Criterion) can be used to compare models and see whether subsequent ones are significantly better than previous ones.

Diessel (2008) sets out to test the hypothesis that there is an iconic relationship between the position of a temporal adverbial clause (which can come before or after the main clause) and the order of the event reported in the adverbial clause as prior, simultaneous, or posterior to the event in the main clause. In other words, Diessel's question is: Is there a tendency for the linear order of clauses to 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? In terms of examples, the prediction would be that 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* and more likely to produce *I fed the cat before I washed the dishes* than *Before I washed the dishes, I fed the cat* (since feeding the cat is conceptually prior in all these cases). Diessel conducts two studies based upon corpus data from the ICE-GB, with samples of clauses beginning with *when*, *after*, *before*, *once*, and *until*. A chi-square test shows that there is a relationship between conceptual order and the linear order of clauses, with $p < 0.001$. However, there are certainly many examples of sentences that violate the iconic order and there are many differences among the sampled clauses that cannot be accounted for by iconicity, so it seems necessary to include more variables in the study. These additional variables include: 1) the meaning of the clause (which may account for the distributional differences between *once*-clauses, which are frequently conditional and *after*-clauses, which are frequently causal), 2) the length of the clause (since long clauses tend to occur sentence-finally), and 3) the syntactic complexity of the clause (since complex clauses tend to occur sentence-finally). Thus Diessel's logistic regression model has the position of the adverbial clause (initial vs. final) as the dependent variable, and has as independent variables conceptual order (iconicity), meaning, length, and syntactic complexity. Whereas syntactic complexity did not turn out to be significant and was removed from the model, all of the other variables were indeed significant. Quite a bit of detail is revealed by the regression model, for example that meaning is significant only for the positioning of conditional *once*- and *until*-clauses, and that length is significant only for *once*- and *until*-clauses. The analysis supports Diessel's hypothesis concerning iconicity and gives us much information about other factors that are involved in the order of clauses as well.

3.6. Mixed effects: Adding individual preferences into a regression model Zenner et al. 2012

The variation found in data can have many sources. Hopefully the variables that you are testing are a major source of differences in the data, showing that the variables you have identified are indeed relevant. These independent variables are sometimes referred to as fixed effects since they have a fixed set of values. In Diessel's logistic regression model described above, all of the independent variables are fixed effects: syntactic complexity was coded with two values (simple, complex), meaning was coded with three values (purely temporal, temporal with implicit conditional meaning, temporal with causal or purposive meaning), and length was a continuous variable measured by dividing the number of words in the adverbial clause by the total number of words in the complex sentence (theoretically ranging from 0 to 1).

However, individual preferences or tendencies can also come into play, and since these are keyed to individuals sampled randomly from a potentially infinite population, they are called random effects. Recall our example of the correlation between corpus frequency and reaction time. If we ran this experiment, we would likely discover that each individual subject has a personal range of reaction times, since some people are just naturally faster than others. This is a well-known problem, and in fact in many psychological studies it turns out that the random effects of personal preferences are actually more pronounced than the effect that the researcher is trying to measure. Imagine, for example that the average baseline difference in reaction times between participant A and participant B in the experiment is 100 milliseconds, but the frequency effect is only 50 milliseconds. If you don't know and cannot account for the individual differences, the frequency effect will be overwhelmed by the random effects of the participants.

Mixed effects models can combine both 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. In addition to use in psycholinguistic experiments, mixed effects models can be useful in various ways in corpus research too. For example, if a corpus has multiple data points from a set of authors, each author can serve as a random effect in order to take into account the fact that different authors will have different preferences for use of various linguistic forms. The source of random effects need not necessarily be human beings. For example, lexemes might also act as random effects in a model, since they can have individual patterns of behavior. For example, Nesset et al. (2010) and Nesset and Janda (2010) apply a mixed effects model to a historical change underway in Russian verbs; in this model the individual verbs are a random effect since each verb has its own tendencies in relation to the ongoing change. Note also that Baayen et al. forthcoming includes a mixed effects model for

an experiment in which subjects (as a random effect) chose between Russian prefix allomorphs *o-* vs. *ob-* and all the data and R code associated with this model are available at this site: http://ansatte.uit.no/laura.janda/RF/RF.html.

Zenner et al. (2012) bring a quantitative perspective to a sociolinguistic study of anglicisms in Dutch. Several possible factors in the success of loanwords have been suggested by previous research, but very little empirical work has been undertaken, and no prior studies use a multivariate approach. Corpus data (from two newspaper corpora), along with a host of other measures are collected in relation to 149 lexemes with human reference such as *manager*. An onomasiological profile shows the relative distribution of the English loanword and its Dutch equivalents (if any). For example, English *backpacker* is attested 425 times in the corpus, while its Dutch equivalents *rugzakker*, *rugzaktoerist* are attested 941 times, and thus the success rate of *backpacker* is $425/(425 + 941) = 31\%$, which serves as the dependent variable. Zenner et al. investigates the variables that have been proposed as factors in the penetration of English loanwords, namely: 1) relative length of the anglicism vs. Dutch equivalent; 2) lexical field (media & IT; sports & recreation; etc.); 3) era of borrowing (up to 1945, 1945–1989, after 1989); 4) luxury vs. necessary borrowing (where necessary borrowing occurs when there is no Dutch equivalent); 5) concept frequency (how often the concept was named by either a Dutch or an English word, for example, the concept frequency for BACKPACKER cited above is $425+941 = 1366$, however these figures were log transformed in order to reduce the effects of extreme numbers, so in this case $\log(1366) = 7.23$); 6) date of measurement (a diachronic corpus factor); 7) register (popular vs. quality newspapers); and 8) region (Belgian Dutch vs. Netherlandic Dutch). In addition to all of these fixed effects, because several measuring points were used for each concept and those data points would therefore not be independent observations, the concept expressed was taken as a random variable. In other words, the mixed effects model took into account any individual preferences associated with the concepts themselves. The model found both main effects and interactions. The regional, register, and diachronic variables were not found to be significant. The two strongest main effects, both with $p = 0.000$, were 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 the earlier eras. Both of these findings make sense because highly frequent concepts are likely to have well entrenched Dutch expressions that would be resistant to borrowing and loanwords from the most recent era have had less time to become established as successful. The interactions in the model give more nuance to the study, for example showing that concept frequency is a factor only when the anglicism is also the shortest lexicalization, and that the difference between luxury and necessary borrowings is strongest in the 1945–1989 era.

3.7. Cluster analysis: Finding out which items are grouped together
Janda and Solovyev 2009

All of the models discussed so far have involved testing whether a value is significant or not. In other words, the question we have asked has always been, given the value X that we obtain in this data, what is the probability that X reflects a meaningful property in a potentially infinite population, rather than being merely a chance artifact of the sample? Cluster analysis asks a different kind of question, namely: Given a set of items, which of them are grouped closest together and which are farthest apart? Another way to state this question is: What is the distance between the items in the set? If each item in a set has an array of values associated with it, it is possible to use mathematical means such as squared Euclidean distances to calculate the distances between the arrays of values. A cluster analysis does just this, yielding a proximity table that shows the distances between each pair of items in the set from which a graph of the clusters can be derived.

Janda and Solovyev (2009) approach the relationships within two sets of Russian synonyms, six words meaning 'sadness', and five words meaning 'happiness', by introducing the constructional profile method. The constructional profile of a word is the relative frequency distribution of the grammatical constructions that a word appears in, as measured in a corpus. The assumption is that the constructional profile is a possible measure of a word's meaning, since there should be a relationship between the meaning of a word and its behavior. Although a Russian noun can appear in seventy constructions involving prepositions and case endings, for most nouns fewer than ten such constructions occur regularly. 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, and these are the basis for the constructional profiles of these words. Within the set of 'sadness' synonyms, for example, there were significant differences in the constructional profiles (a chi-square test gives $p < 0.001$ and a Cramer's V effect size of 0.305), but this does not tell us which of the synonyms are closer to each other and which are further apart. The constructional profile for each noun, with the frequency found in each construction, is the array of values that serves as input for the cluster analysis. The output 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. There appear to be asymmetries between metaphorical uses of grammatical constructions and concrete ones. For example, metaphorically sadness can function as a pit and while the constructions for falling into and being in

sadness are quite common, the construction for getting out again is exceedingly rare; by contrast, nouns denoting physical pits appear robustly in all three constructions.

3.8. Other alternatives: tree & forest, naive discriminative learning, multidimensional scaling, correspondence analysis

In addition to the models described here and illustrated in the articles in this anthology, there are many other statistical models that might be applied to linguistic data. Here we review a few additional models that the reader is likely to encounter. These can be divided into two groups: 1) alternatives to regression models, and 2) alternatives to cluster models.

Alternatives to regression

In addition to the weaknesses that follow from correlation cited above (assumption of linearity and lack of causal implication), regression rests on two assumptions that are often violated by linguistic data. One is that because regression is a parametric model, it assumes that data should follow the bell curve of what statisticians call a normal distribution. Corpus data is however usually highly skewed, thus rendering regression less appropriate. The other assumption is that all of the combinations of the various levels of all variables should be represented in the dataset. However, linguistic data often involves paradigmatic gaps where certain combinations of the relevant variables are necessarily absent. For example, in evaluating the distribution of certain suffixes in Russian, both the factors of form (with levels finite, gerund, participle) and prefixation (prefixed, unprefixed) are relevant, but it is categorically impossible to find examples of unprefixed gerunds (see Baayen et al. forthcoming).

There are two alternatives that can be used for similar data that avoid both the parametric assumption and the assumption concerning combinations of values: 1) classification and regression trees in combination with random forests (here called "tree & forest"; Strobl et al. 2009), and 2) naive discriminative learning (Baayen 2011, Baayen et al. 2011). The tree & forest model uses recursive partitioning to yield a classification tree, showing the best sorting of observations separating the values for the dependent variable. It can literally be understood as an optimal algorithm for predicting an outcome given the predictor values, and Kapatsinski (2013) suggests that from the perspective of a usage-based model, each path of partitions along a classification tree expresses a schema (see also Kyröläinen 2013 for an application of tree & forest modeling in cognitive linguistics). Naive discriminative learning is a quantitative model for how choices can be made between rival linguistic forms, making use of a system of weights that are estimated using equilibrium equations.