31

The Quantitative Turn

Laura A. Janda

31.1 Introduction

The quantitative turn in cognitive linguistics is a force to reckon with. In this chapter, I track the history of our quantitative turn, which has been facilitated by a confluence of three factors: the usage-based nature of the cognitive linguistics framework, the advent of electronic archives of language data, and the development of statistical software. I give an overview of the types of statistical models cognitive linguists are turning to, illustrated by the kinds of research questions that are being asked and answered using quantitative tools. I also discuss the opportunities and dangers that we face now that we have taken our quantitative turn.

31.2 What Brought about the Quantitative Turn?

A survey of articles published in the journal *Cognitive Linguistics* (Janda 2013a) gives us a perspective on the quantitative turn in cognitive linguistics (see also Janda 2013b). Figure 31.1 presents the distribution of articles in the journal from its inaugural volume in 1990 through the most recent complete volume in 2015, according to whether or not they presented quantitative studies.[1]

Figure 31.1 reports percentages of quantitative articles for each year. A thick line marks 50 percent 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 2008–2015 – when most articles were quantitative. In 1990–2007, twelve out of eighteen volumes had 20–40 percent quantitative articles.

[1] This survey includes only articles proper, excluding review articles, book reviews, overviews, commentaries, replies, and squibs. For the purpose of this survey we define a 'quantitative article' as an article in which a researcher reports numbers for some kind of authentic language data.

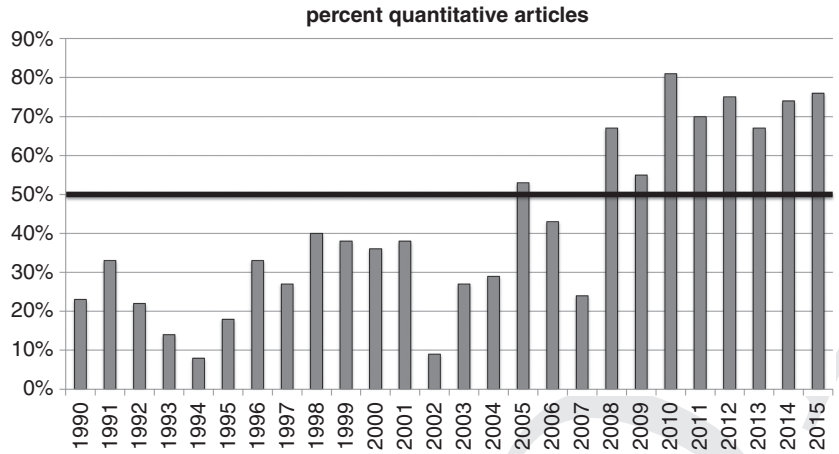

Figure 31.1 Percentage of articles presenting quantitative studies published in *Cognitive Linguistics* 1990–2015

The lowest points were 1994, with one out of twelve articles, and 2002, with one out of eleven articles. In 2005 the move was in the other direction, with ten out of nineteen articles. At present the publication of quantitative articles seems to be leveling off at the rate of about 75 percent.

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 of *Cognitive Lingustics*: 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 percent line, and it is unlikely that we will drop below that line again in the foreseeable future.

This survey indicates approximately when quantitative studies came to dominate our scholarly output. It also shows us that cognitive linguistics has always engaged in quantitative studies, yet there is no reason to expect quantitative studies to entirely eclipse non-quantitative studies either. I do not mean to imply that there is a dichotomy between quantitative versus non-quantitative studies. A variety of valuable types of studies require no quantitative analysis, such as descriptive linguistics, theoretical works, and overviews of the state of the art. Conversely, an ideal quantitative study relies on linguistic description, expands our theoretical framework, and thus contributes to the state of the art. Thus, in a sense, quantitative studies depend on and ideally integrate non-quantitative components, though the reverse is not necessarily true.

Although this survey is based on a single journal, *Cognitive Linguistics* is the signature journal of our field and it reflects the recent history of cognitive linguistics as a whole. Evidence from conferences and textbooks

devoted to quantitative studies points in the same direction. Since 2002 there have been six bi-annual meetings of Quantitative Investigations in Theoretical Linguistics, a conference series devoted to statistical analysis of language data predominantly from the point of view of cognitive linguistics. QITL has grown over the years from a workshop with only a dozen speakers to a three-day event. Three of the authors of the five textbooks on the use of statistical methods in linguistics that I cite in section 31.2.3 have close ties to cognitive linguistics: Harald Baayen, Stefan Gries, and Natalia Levshina.

How did we reach the quantitative turn? As is usually the case with historical developments, there was no single cause, but rather a combination of factors that pushed and pulled cognitive linguistics in this direction. Pushes have come from the theoretical framework of cognitive linguistics, which has proved to be fertile ground for developing research questions that rely on analysis of observed data. Pulls have come from the attraction of vast data resources and the access to sophisticated tools for their analysis.

31.2.1 A Usage-based Model of Language is Data-friendly

Cognitive linguistics is a usage-based model of language structure (Langacker 1987: 46, 2013: 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 2010a: 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.[2]

Even the question of what constitutes data in linguistics is controversial, and largely dependent upon the theory that one uses. Some researchers refer to constructed examples and individual intuitions as data, while others prefer to use corpus attestations or observations from acquisition or experiments. Introspection certainly plays an important role in linguistic analysis and indeed in the scientific method in general (cf. section 31.3.2), but reliance on introspection to the exclusion of observation undermines linguistics as a science, yielding claims that can be neither operationalized nor falsified (cf. section 31.4.2). 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

[2] For overviews of the use of corpus linguistics across various theoretical frameworks, see Joseph 2004 and Gries 2009b.

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 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 introspection about data (perhaps augmented by introspection solicited from a few colleagues) and the systematic elicitation of the intuitions of naïve informants under experimental conditions, which is a legitimate scientific method that normally involves quantitative analysis. The difference is that whereas the linguist's introspection does not necessarily yield reliable, replicable results, the elicitation of native speakers' intuitions can yield such results. Introspection on the part of linguists can present numerous problems in that there are disagreements between linguists (cf. Carden and Dieterich 1980, Cowart 1997, Anketa 1997); their intuitions about mental phenomena are often inaccurate (Gibbs 2006); and last but not least, linguists' intuitions may be biased by their theoretical commitments (Dąbrowska 2010). Even if we put aside the issue of whether a linguist can report viable intuitions about language data, it is a fact that a linguist is an individual speaker, and there is abundant evidence that different speakers of the same language have different intuitions about linguistic forms. Given the fact of inter-speaker variation, it is more reasonable to assume that there is not just one model, but instead many models of the grammar of a given language (Dąbrowska 2012, Barth and Kapatsinski 2014, Günter 2014). Every speaker, linguist or not, has to some extent a unique experience with the use of his or her native language, and a usage-based theoretical framework is well equipped to accommodate this fact.

31.2.2 Advent of Electronic Language Resources

Recent history has impacted the practice of linguistics through the development of language corpora and statistical software. Today we have access to balanced multipurpose 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. Many languages have national corpora, and open corpora are being built, providing free access not only to the linguistic forms and annotation in the interface, but also to the code itself, facilitating further exploration of data. A free resource that has attracted linguists is the Google Books Ngrams Corpus, which has a function that charts the frequency of words and phrases in a few of the world's largest languages. In addition to corpora of written language, spoken corpora are becoming available, and some resources are even

multimodal. For example, the UCLA NewsScape Library is an archive of billions of words in several languages, along with associated sound and images captured from television newscasts.

The attraction of all this data is predictably compelling, particularly for linguists who view usage events as linguistic data. It is no surprise that a large portion of the quantitative studies undertaken by cognitive linguists have involved the analysis of corpus data, either alone or in comparison with experimental results (see Gries this volume Ch. 36 for more details concerning corpus linguistics).

31.2.3 Advent of Analytical Tools

At approximately the same time that electronic corpora emerged, statistical software likewise became widely available. Thus linguists have at their disposal the means to explore the structure of complex data. 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.

A natural place to turn to for inspiration in the use of analytical tools is computational linguistics.[3] Computational linguistics has of course been around since the 1950s, and computational linguists have considerable expertise in digital exploration of language data. However, the goals of cognitive linguistics and computational linguists have traditionally differed significantly due to the theoretical focus of cognitive linguistics (though there is good potential for collaboration, cf. section 31.4.1). Therefore, in addition to drawing on the capacities of computational linguistics, we have looked for leadership to other disciplines that also deal with human behavior but took the quantitative turn earlier, in particular psychology (in addition to sociology and economics).

(We linguists are still in a formative period where we have not yet settled on a set of best practices for use of statistical methods. A pioneering work in bringing statistical methods to linguists was Butler's 1985 textbook. But ten years ago this textbook was out of print and there were very few alternatives. Since cognitive linguistics took its quantitative turn in 2008, several texts have been published such as Baayen (2008), Johnson (2008), Larson-Hall (2010), Gries (2013c), Levshina (2015). These books, together with scholarly works, are helping to establish norms for the application of statistical models to linguistic data and analysis. However, the field of statistics is itself in a state of considerable flux, particularly in the area of non-parametric models (especially relevant for us, since linguistic data is usually non-parametric; see section 31.3.1.2), adding an

[3] See, for example, the journal *Computational Cognitive Science* at www.computationalcognitivescience.com/.

extra challenge for cognitive linguists as relative late-comers to quantitative analysis.

31.3 What Does the Quantitative Turn Bring Us?

An introduction to statistical methods goes beyond the scope of this chapter and is better addressed by the textbooks cited above, so I will give only a bird's-eye view, sprinkled with illustrative examples of how cognitive linguists are applying such methods. The scope of this overview is restricted to tracking some trends and discussing the relationship between quantitative methods and introspection.

31.3.1 Quantitative Methods in Cognitive Linguistics

The goal of this section is to illustrate how quantitative methods are being used in cognitive linguistics and to identify some methods that are likely to stand the test of time. All statistical models are subject to assumptions and limitations concerning the nature of the data that need to be carefully observed and many models also facilitate the measurement of effect sizes which should be applied wherever possible, but since these issues are covered in textbooks, neither of them will be addressed in detail here.

31.3.1.1 Is A Different from B? Chi-square Test, Fisher Test, Binomial Test, T-test, ANOVA

The main idea of this set of tests is to find out whether there are significant differences between two (or more) measured phenomena. Just because two numbers are different does not mean that there is a statistically significant difference between them. This set of tests aims to discover whether there is sufficient reason to reject the 'null hypothesis.' The null hypothesis is the default position according to which there is no difference between the measured phenomena. If the null hypothesis is true, the observed difference can be accounted for by random fluctuations in samples taken from a larger population of observations in which there is no difference. If the null hypothesis is rejected, the observed difference is unlikely to be accounted for by such fluctuations.

Languages often give speakers choices, for example the choice between: A) the ditransitive (*read the children a story*), and B) the prepositional dative (*read a story to the children*) constructions in English. Corpus or experimental data might reveal a pattern such that there is more use of choice A in one environment (X) than in another environment (Y). But is the difference between the measurements of A and B a significant difference? In other words, is there reason to believe that there is a real difference between the frequency of A and B, or might the difference we observe be just a matter of chance (the null hypothesis)? A chi-square test can tell us the probability

that the observed difference is significant. Chi-square tests have been used, for example, to test differences between the two English constructions listed above (Stefanowitsch 2011, Goldberg 2011), the difference between physical and metaphorical understanding of English *path* versus *road* (Falck and Gibbs 2012), and the difference in the use of SVO constructions between a child and his mother (Theakston et al. 2012).

While a chi-square test can give an overall evaluation of whether there is something significant in a matrix of numbers, the Fisher test is useful when trying to find exactly which of those numbers deviates significantly from the overall distribution of the matrix. The Fisher test was brought to the attention of cognitive linguists by Stefanowitsch and Gries (2003, 2005) in collostructional analysis, where the point was to find out which words (such as *disaster, accident*) were more or less attracted to constructions (such as *an N waiting to happen*). This application of the Fisher test has since come under criticism (Bybee 2010: 97–101, Baayen 2011: 315, Schmid and Küchenhoff 2013, Küchenhoff and Schmid 2015),[4] primarily for the use of numbers on very different scales (especially when some of these numbers are estimated rather than actual numbers), and for the use of the p-value as a measure of collostruction strength. However, when used on actual (not estimated) numbers of low values (tens or hundreds rather than tens of millions), the Fisher test is a useful way to probe the relationships among values in a matrix.[5]

If you know the overall distribution of a phenomenon, a binomial test can tell you whether the frequency of that phenomenon in your sample is significantly different from that in the overall distribution. Gries (2011) compared the frequency of alliterations in the British component of the International Corpus of English (the ICE-GB, here taken to reflect the overall distribution of alliteration in English) with the frequency of alliteration in lexically specified idioms such as *bite the bullet* (as opposed to *spill the beans* with no alliteration). The binomial test showed that the frequency of alliteration in English idioms is indeed significantly higher than in English overall.

If two groups of items (e.g. two different semantic groups of lexemes – let's call them A and B) each get a set of scores (e.g. acceptability scores), those two sets of scores will probably overlap. If the means of scores of the two groups are different, how do we know whether there is a significant difference between group A and group B? In other words, how do we know whether the difference in means is likely to reflect a real difference, or just chance variation in a situation where A and B actually behave the same in a larger sample? A t-test can handle a simple comparison of two groups. ANOVA ('analysis of variance'), which is an extension

[4] See also Gries' responses to this criticism in Gries 2014b and Gries 2015a.

[5] A relevant example of the application of the Fisher test is presented here: http://emptyprefixes.uit.no/semantic_eng.htm.

of the t-test, compares the between-group variation in scores with the within-group variation in scores, making it possible to compare more than two groups or more than one variable across the groups. Dąbrowska, Rowland, and Theakston (2009) wanted to investigate the nature of long-distance dependencies such as *Who$_1$ did Mary hope that Tom would tell Bill that he should visit _____$_1$?* Dąbrowska, Rowland, and Theakston's hypothesis was that spontaneously produced long-distance dependencies 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. They conducted an experiment in which children and adults were asked to repeat long-distance dependencies that did versus did not follow the lexically specific templates. An ANOVA analysis showed that children rely on lexically specific templates as late as age six, and that even adults are more proficient with long-distance dependencies that match the 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.

31.3.1.2 What Factors are Associated with A? Correlation, Regression, Mixed Effects Regression, Classification and Regression Trees, Naïve Discriminative Learning

Suppose you want to find out what factors contribute to a given phenomenon, such as reaction time in a word-recognition task. The reaction time (A), termed the dependent variable in this example, may be related to various other phenomena such as frequency, length, and morphological complexity (B, C, D, etc.), known as independent variables. Correlation and regression are a family of models that can be used to explore such relationships.

Correlation refers to the degree of relationship between two variables, such that the stronger 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 thus it is possible to fit a line to a plot of data where one variable (frequency) is on the x-axis and the other variable (reaction time) is on the y-axis. If there is a correlation, given the frequency of a word it is possible to use the slope and intercept of the line to predict the reaction time, and conversely, given the reaction time associated with a word it is possible to predict its frequency.

Notice that the prediction goes both ways. A big caveat with correlation is that prediction is not the same as causation: an association between

frequency and reaction time does not necessarily mean that higher frequency causes shorter reaction times (or the converse). Even if you can use the value of B to predict the value of A with 100 percent accuracy, correlation tells you only that there is a relationship, not that B causes A. However, linguists are not immune to the temptation to assume causation when correlation is found (for a survey of correlation in relation to this problem, see Ladd, Roberts, and Dediu 2015). Another problem with interpreting correlation is that an apparent association between variables A and B might well be caused by other variables that have not been taken into account. The larger the dataset, the easier it is to find spurious relationships such as a positive correlation between linguistic diversity and traffic accidents (overlooking more telling factors such as population size and GDP; see Roberts and Winters 2013).

Correlation has been used in a wide variety of studies. For example, in a study of long-distance dependencies, Ambridge and Goldberg (2008) found a correlation between the backgrounding of a clause (measured by a negation test) and the difficulty of extracting a clause (measured by the difference between acceptability in questions versus declaratives), such that verbs like *know* and *realize* behaved very differently from verbs like *think* and *believe*. In a study of Polish prefixed verbs, Kraska-Szlenk and Żygis (2012) discovered a correlation between the reported morphological transparency of a prefixed verb and its acceptability rating by experiment participants.

A regression analysis allows you to consider the relationship between an independent variable (A) and a set of dependent variables (factors associated with A). Linear regression is based upon the same calculations as correlation, since the line of best fit in a correlation is the regression line, defined by the regression equation. 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 'residual error.' The standard error of estimate (which is an estimate of the standard deviation of the actual scores from the predicted scores) gives us a measure of how well the regression equation fits the data. Because regression is based upon the same calculations as correlation, it also inherits the same drawbacks, namely that by default it assumes a linear relationship (though this can be modified), it cannot tell us anything about causation, and any association that we find might actually be the result of other variables that we have not taken into account.

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 predictors). Ideally the independent variables should be independent not just of the dependent variable, but also of each other (thus 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. The goal of a logistic regression model is to predict the probability that a given value (e.g. initial versus final position) for the dependent variable will be chosen. If the dependent variable has an ordered set of more than two values (such as the values low, medium, and high acceptability), it is possible to use an ordinal regression model. The use of regression, and in particular logistic regression, has become fairly common in cognitive linguistics. For example, Diessel (2008) tested 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, the prediction is 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*. Diessel constructed a logistic regression model to explore the relationship between the position of the adverbial clause (initial versus final) as the dependent variable (the factor that is being predicted), and as independent variables conceptual order (iconicity), meaning, length, and syntactic complexity.

Mixed effects models are regression models that can take into account 'random effects,' which are the effects introduced by individual preferences. Mixed effects models are commonly used in experimental studies where random effects account for the behavior of individual stimuli and/or participants, and such models make it possible to arrive at generalizations that go beyond a specific sample of speakers or data. Random effects are relevant when we need to cope with what are called 'repeated measures,' such as in an experiment where multiple measurements are taken from each participant. In a word-recognition task where each participant responds to a set of words, some participants will be faster in general than others, so the baseline speed of each participant needs to be taken into account as a random effect. Random effects are opposed to fixed effects, which have a fixed set of values such as those for sex and age for experimental participants or tense, number, and person for verbs. For example, lexemes might act as random effects in a model, since they can have individual patterns of behavior. Janda, Nesset, and Baayen (2010) and Nesset and Janda (2010) applied 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: some verbs use more of the innovative forms while others tend to resist innovative forms. In a study of the relative success of anglicisms in Dutch, Zenner, Speelman, and Geeraerts (2012) treated the concept expressed as a random effect, along with a number of fixed effects: relative length of anglicisms versus Dutch equivalents, lexical field, era of borrowing, 'luxury borrowing' (when a Dutch equivalent exists) versus necessary borrowing (when there is no Dutch equivalent),

era of borrowing, concept frequency, date of measurement, register, and region.

Regression models rest on assumptions that are often violated by linguistic data. Linear regression is a parametric model, which means that it tests hypotheses about population parameters. In other words, this type of model assumes that data should follow the bell curve of what statisticians call a normal distribution. Corpus data is, however, usually highly skewed, thus rendering linear regression less appropriate. Logistic regression assumes that all of the combinations of the various levels of all variables should be represented in the dataset. However, linguistic data often involve systematic gaps where certain combinations of the relevant variables are necessarily absent. There are at present at least two alternatives to regression models that offer the advantage of being non-parametric tests that also do not require all levels of variables to be observed in the dataset: classification and regression trees and naïve discriminative learning.

The classification and regression tree model ('CART'; Strobl, Tutz, and Malley 2009) uses recursive partitioning to yield a tree showing the best sorting of observations separating the values for the dependent variable. Figure 31.2 shows an example of a CART tree from Baayen et al. 2013, showing the behavior of the Russian verb *gruzit'* 'load' with respect to two grammatical constructions: the 'goal' construction, as in *load the truck with hay*, versus the 'theme' construction, as in *load the hay onto the truck*.

The terminal nodes at the bottom of the tree show the number of examples in each node ('n=') and plot the distribution of theme versus goal uses for those examples. The top node of the tree (node 1) takes the entire dataset and makes the cleanest first division by finding the independent variable that is most effective at separating the goal uses from the theme uses, namely VERB: the 'load' verb prefixed in *na-*, *za-* or without prefix (the left branch) prefers goal use (represented by the light grey bars in the terminal nodes) more than when prefixed in *po-* (the right branch), where theme use (dark grey bars in terminal nodes) is strongly preferred. On the right side at node 13, the *po*-prefixed verb forms are further sorted into reduced constructions (yes), where a few goal uses are attested (light grey in node 15) versus full constructions (no), where only theme uses are attested (node 14). Most of the goal uses appear to the left, where we see that at node 2 the most important factor is whether the verb form is a participle (yes) or not (no): nearly all these examples are goal uses, though a few theme uses are found for the *za*-prefixed verb (dark grey in node 5).

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

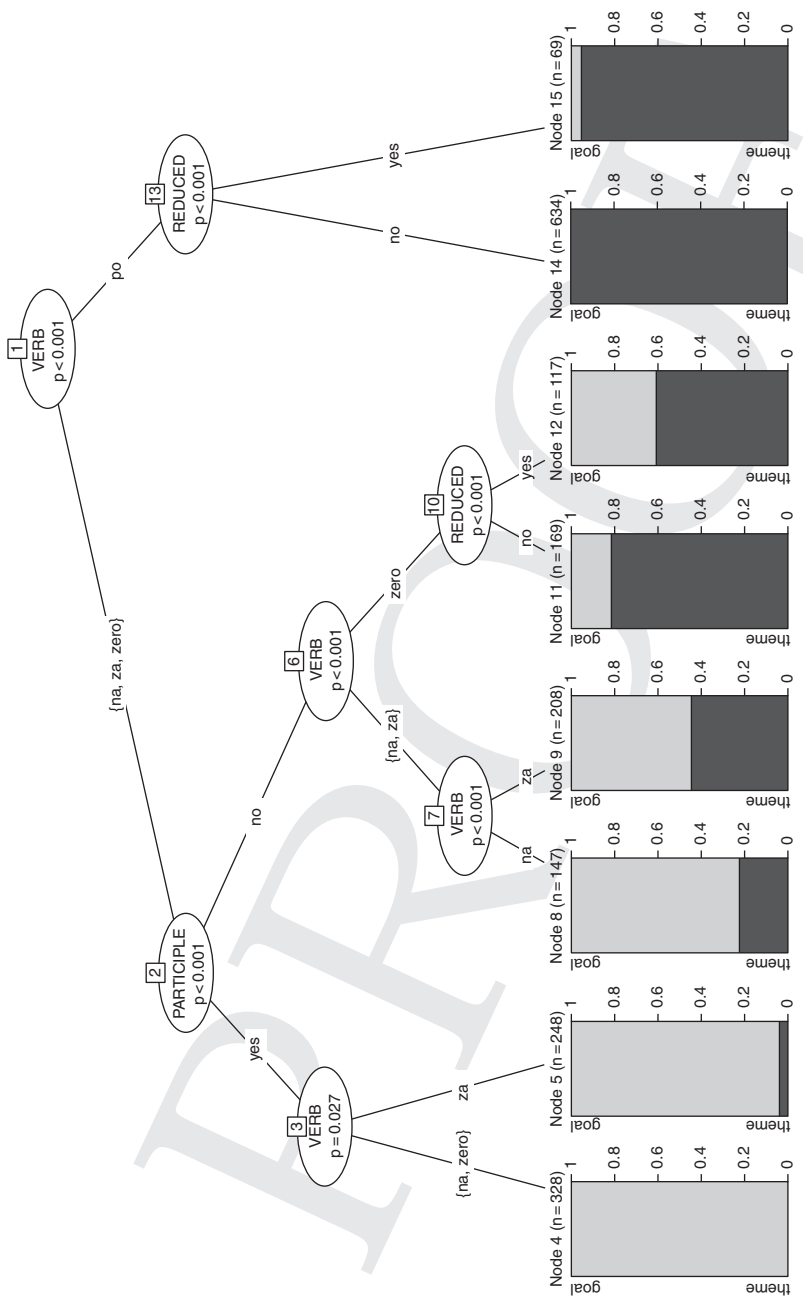

Figure 31.2 CART tree for Russian *gruzit'* 'load' from Baayen et al. 2013

over a set of instances. For example, in Figure 31.2, 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 percent) for theme use.

Naïve discriminative learning (Baayen 2011, Baayen et al. 2011) 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, modeling the usage-based experience of a speaker. Both CART and naïve discriminative learning offer means for measurement of the importance of variables and validation of results. 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 and for measurement of variable importance. Naïve discriminative learning partitions the data into ten subsamples, nine of which serve as the training set, reserving the tenth one to serve for validation. This process is repeated ten times so that each subsample is used for validation.

Baayen et al. (2013) test the performance of regression against classification tree and naive discriminative learning models across four datasets and find that the three models perform very similarly in terms of accuracy and measurement of the relative importance of variables.

31.3.1.3 What is the Structure of Relationships among a Group of Items? Cluster Analysis, Multidimensional Scaling, Correspondence Analysis

A given linguistic item, for example, a lexeme, might be measured in many different ways, yielding an array of data; and a group of lexemes could then each have an array. The linguist might want to ask: which of these items are more similar to others, how can these items be grouped? Cluster analysis, multidimensional scaling, and correspondence analysis take as input arrays of data associated with a set of items and use various mathematical techniques to arrange the items into a 'space' of two or more dimensions.

Janda and Solovyev (2009) approached the relationships within two sets of Russian synonyms, six words meaning 'sadness,' and five words meaning 'happiness,' by measuring the relative frequency distribution of the grammatical constructions for each word in a corpus. The output of a hierarchical cluster analysis shows us which nouns behave very similarly as opposed to which are outliers in the sets. These results largely confirm the introspective analyses found in synonym dictionaries, and point to asymmetries between metaphorical uses of grammatical constructions and concrete ones.

Multidimensional scaling has been used in various ways in cognitive linguistics; for example, to map out the functions of grammatical case in

Slavic languages (Clancy 2006) and to map the relations of aspect and expressions for spatial location (Croft and Poole 2008; see also Janda 2009).

Eckhoff and Janda (2014) used correspondence analysis to measure distances between verbs according to the frequency distributions of their grammatical forms, yielding a sorting that suggests that there was indeed a difference in behavior between perfective and imperfective verbs in Old Church Slavonic.

31.3.2 Role of Introspection

There should be a healthy balance between introspection and observation in any scientific inquiry. Introspection is the source of inspiration for hypotheses, which are then tested via observation. When it comes to analysis, 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. The critical eye of introspection is necessary to ferret out suspicious results and alert us to problems in design and analysis. Whereas theory should of course be informed by data, theoretical advances are typically born through introspection.

Introspection is irreplaceable in the descriptive documentation of language. In fieldwork, a linguist interacts with speakers and posits the structure of a grammar 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 language at all.

31.4 Where Does the Quantitative Turn Lead Us?

Like any journey, taking the quantitative turn both opens up new opportunities and exposes us to new perils. It is worth taking stock of the pros and cons of this situation.

31.4.1 Opportunities

The most obvious advantage to taking the quantitative turn is of course the opportunities we gain to discover structures in linguistic data that would otherwise escape our notice. In addition, we can bolster the scientific prestige of our field and foster greater accountability and collaboration.

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. 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.

One important step we can take as a community is to make a commitment to publicly archive both our data and the statistical code used to analyze it. This will help to move the field forward by providing standards and examples that can be followed. In so doing, we can create an ethical standard for sharing data, stimuli, and code in a manner explicit enough so that other researchers can access the data and re-run our experiments and statistical models. 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 advance more efficiently if we pool our efforts in a collective learning experience. In many cases, funding agencies require researchers to share their data, adding further motivation for public archiving of data. Ultimately, the most important reason for making data publicly accessible stems from the basic principles of the scientific method, namely that scientific findings should be falsifiable and replicable. Researchers should be held accountable for their findings and only findings that can be replicated can be considered valid. One good option for linguists is the Tromsø Repository of Language and Linguistics ('TROLLing' at opendata.uit.no), a professionally managed, free, and open international archive of linguistic data and statistical code built on the Dataverse platform from Harvard University.

As cognitive linguists become more familiar with quantitative methods, the opportunity for joining forces with computational linguists also increases. We can bring to the table valuable descriptive analyses and theoretical perspectives that can enrich collaboration in the building of better natural language processing and language technology applications.

31.4.2 Dangers

There are at least two types of dangers lurking just beyond the quantitative turn. One involves over-reliance on quantitative methods, and the other involves various kinds of misuse or neglect of data. In the face of these dangers we can lose sight of the bigger picture of our theoretical principles and values.

If taken too far, quantitative research runs the risk of triviality and fractionalization of the field. It is very easy for researchers to be seduced by fancy equipment and sophisticated software to the point that these

receive more attention than relevant linguistic principles. The most harmless negative outcome of this situation are shallow studies that do little or nothing to advance the field because they involve number-crunching without any real linguistic or theoretical goal. The potential outcome is a cognitive linguistic version of 'cargo cult science'[6] in which linguists perform empty rituals of calculations in hopes of conjuring up publishable results.

More problematic is the substitution of 'quantitative' for 'empirical' and 'scientific' in the minds of researchers. The use of quantitative methods in a study does not make it better or necessarily any more empirical or scientific than language documentation or qualitative analysis. Confusion of these concepts could result in the marginalization of many of the traditional endeavors of linguists that could then be disadvantaged in the selection of works presented at conferences and in publications. We thus risk erosion of the core of our field, linguistic description and theoretical interpretation, which are also the source for research hypotheses. As Langacker stated in 2015, "linguistic investigation is a highly complex and multifaceted enterprise requiring many kinds of methods and expertise"[7] and these various kinds of expertise should ideally be mutually supportive.

In the age of big data, it becomes far too easy to find results simply because as the number of observations increases toward infinity (or just millions and billions), and statistical tests are able to find effects that are infinitesimally small and therefore meaningless. To some extent this can be corrected for by the use of effect sizes as a check on results. However, Kilgarriff (2005) argues that since languages do not behave in a random fashion, the use of statistics to test null hypotheses is perhaps misguided to begin with. There will always be some patterns in linguistic data. The linguist's job is to bring enough insight to the enterprise to know what is worth looking for and to distinguish between results that have a real impact on the advancement of our science and those that do not.

Focus on big data analysis also threatens to marginalize languages themselves. Only a tiny fraction of the world's languages have the resources to support large corpora, experimental studies, and comprehensive language technology coverage. The quantitative turn has the potential to exacerbate the existing imbalance between the few languages that many linguists study and the majority of languages that are largely ignored.

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

[6] This term is used by Feynman (1992) to compare inept scientists to 'cargo cult' south sea islanders, who, after experiencing airlifts during WWII, constructed mock runways manned by mock air traffic controllers, in hopes that this would cause more airplanes to land and bring them cargo.

[7] Quoted from Langacker's presentation at the 'Theory and Method' panel at the International Cognitive Linguistics Conference (2015a).

will be most accessible to readers. Sometimes the structure of the data dictates a more complex model, but very complex models carry with them the disadvantage 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 understand and communicate their results.

Wherever numbers are involved, there is a temptation to misrepresent them. 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 (cf. section 31.4.1) can help us to protect our field from this kind of dishonor. 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.

Major corporations such as Google, Amazon, Apple, and Facebook, along with hacking and spyware operations and state governments, have access to massive quantities of human language data. The lure of developing mining techniques via language analysis is part of what Kelly (2010) terms the 'technium,' the collective of archives and devices that constitute an organism-like system with a powerful momentum. This technology is advancing rapidly, and like it or not, we as linguists are contributing to it by improving our understanding of languages. This development is unstoppable; 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.

31.5 Conclusion

Since about 2008, cognitive linguistics has shifted its focus, and is now dominated by quantitative studies. On balance, the quantitative turn is a hugely positive step forward since it puts powerful new tools into the hands of cognitive linguists. Time always brings changes, and changes always bring challenges, but in this case the pros clearly outweigh the cons. Our field can gain in terms of scientific prestige and precision and collaboration. We can show leadership in best practices and the norming of application of statistical models to linguistic data. At the same time, I hope we can retain a humble attitude of respect for our venerable qualitative and theoretical traditions, which we should continue to nurture. If anything, 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.