INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES
SECOND EDITION
EDITED BY
JAMES D. WRIGHT

From Gries, S.T., 2015. Quantitative Linguistics. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 19. Oxford: Elsevier. pp. 725–732.
ISBN: 9780080970868

Quantitative Linguistics

Stefan Th Gries, University of California, Santa Barbara, CA, USA

Abstract

This article surveys a selected variety of statistical methods that are currently used in experimental and observational studies in linguistics. It covers goodness-of-fit tests, monofactorial and multifactorial hypothesis testing methods, and hypothesis-generating techniques. In addition, for the two major sections of significance testing and exploratory methods, the article also discusses a wide range of statistical desiderata, i.e., perspectives and methods whose more widespread recognition or adoption would benefit linguistics as a discipline.

Introduction

Over the last 20 or so years, linguistics has taken a decidedly quantitative turn. While subdisciplines such as phonetics, sociolinguistics, and experimental psycholinguistics have employed statistical methods for a long time, work in most other central subdisciplines – morphology, syntax, semantics, to name but a few – has only done so since about the 1990s. This shift has no doubt been facilitated by a variety of developments.

For example, the predominance of generative linguistics, with its stance that experimental or observational evidence for one's judgments is not really required, has waned and functional, cognitive, and exemplar-usage-based theories have become more widely adopted; since these theories have relied much on empirical data, an increase in the use of statistical tools was a natural by-product of these theoretical developments.

In addition to any major theoretical shifts in linguistics, there are also growing overall tendencies to (1) study linguistic phenomena from quantitative perspectives including, but not limited to, probability theory, information theory, etc. and (2) cross-disciplinary boundaries and, thus, get involved with neighboring disciplines in which statistical modeling has been much more widespread such as psychology, sociology, and communication.

As yet another example, technological progresses – computers with faster processors, more RAM, and larger hard drives as well as the invention of the WWW – have made it much easier to compile and process large(r) corpora and other kinds of 'big data.' Large(r) corpora in turn yield higher frequency data, which cannot be studied by mere eyeballing; so this, too, has led to a burgeoning use of statistics in all areas of linguistics in which corpora are studied.

As a result of these converging trends, quantitative methods in linguistics are now established and constitute a vibrant methodological domain. However, given the ever-evolving nature of the field of statistics and the recency of the more widespread adoption of quantitative methods in linguistics, the latter are also still in a field of flux as best practices are still being developed/established. This article provides an overview of how statistical tools are used in linguistics – given the vastness and diversity of linguistics as well as all the ways in which statistical methods can be used; however, this can only be a highly selective bird's-eye overview. The section 'Hypothesis-Testing Methods' covers quantitative methods that are hypothesis testing in nature, i.e., that return, among other things, *p*-values from significance tests. The section 'Hypothesis-Generating Methods', then, deals with hypothesis-generating approaches, i.e., methods that aim at detecting or hypothesizing structures in data without necessarily testing these for significance. Each of these sections will first provide an overview of what may be considered the state of the art before turning to a variety of desiderata, i.e., methods or developments. The section 'Concluding Remarks' then concludes with a few general developments that linguistics as a discipline would benefit from.

Hypothesis-Testing Methods

Statistical tools that belong to the domain of null-hypothesis significance testing are the most widespread in linguistics. Within this set of tools, one needs to distinguish between *goodness-of-fit tests* and *tests for independence/differences*. The former are concerned with testing whether a characteristics of a particular data set – a mean, a standard deviation, the overall distribution – is different from that of some other data set (e.g., one from a previous study) or a known distribution (e.g., the bell-shaped normal distribution). The latter can be divided up into *monofactorial* and *multifactorial* tests: both involve one *dependent variable* (or response or effect), but the monofactorial designs contain only one *independent variable* (or predictor or cause) whereas multifactorial designs contain more than one independent variable. While not a standard statistical terminology, it has occasionally been didactically useful to distinguish two kinds of multifactoriality; Adopting this distinction, multifactorial$_1$ refers to designs in which multiple independent variables are involved but without interactions, whereas multifactorial$_2$ then refers to designs in which multiple independent variables are involved such that they may interact with each other.

The following sections will survey monofactorial and multifactorial approaches. However, given the fact that no linguistic phenomenon is truly monofactorial in nature and that,

International Encyclopedia of the Social & Behavioral Sciences, Second Edition, 2015, 725–732

correspondingly, the field has been moving in the direction of multifactorial testing, the emphasis will be on this latter type of approach.

Goodness-of-Fit Tests in Common Use

Nearly all statistical tests can be said to involve one of the following five statistics:

- Distributions
- Frequencies
- Averages (such as means or medians)
- Dispersions (such as standard deviations or interquartile ranges)
- Correlations (such as Pearson's r or Kendall's τ).

Thus, for each of these statistics, goodness-of-fit tests are conceivable. The probably most frequent goodness-of-fit test involves testing distributions, more specifically whether data are normally distributed, i.e., whether a histogram or density plot would result in a symmetric bell-shaped form. This scenario is frequent because many statistical tests require that the data tested exhibit this shape lest their significance tests produce unreliable results. Figure 1 exemplifies this scenario: a researcher may have collected the data represented by the histogram and the dashed density curve and now needs to determine whether the data are sufficiently similar to the heavy normal curve to assert that his data are normally distributed.

However, goodness-of-fit tests are available for the other four statistics, too. An example for frequencies involves testing whether the frequencies with which foreign language learners choose three modal verbs in a gap-filling experiment differ significantly from native speakers' choices in some prior study. An example for averages involves testing whether foreign language learners' average acceptability judgment for n sentences is significantly different from native speakers' average acceptability judgment in previously studied data. And an example for correlations involves testing whether the correlation between the age of a child and the child's mean length of utterance in morphemes in corpus data is the same as that found for other children from a comparable corpus.

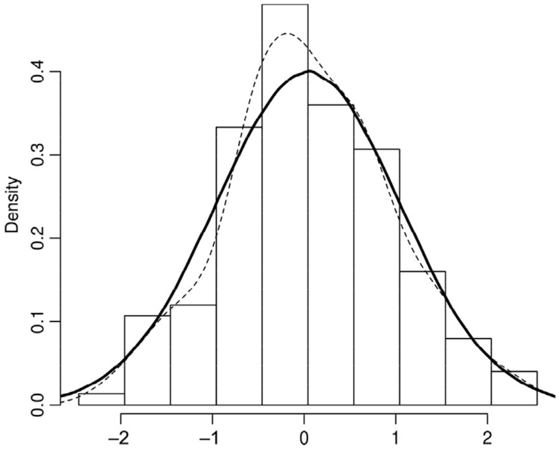

Figure 1 Histogram and density curve (dashed line) of fictitious data compared to a normal distribution with the same mean and standard deviation (heavy line).

Tests for Independence/Differences in Common Use

The much more frequent kind of statistical scenario involves both dependent and independent variables, and the statistic of the dependent variable is typically one of the five listed above.

Monofactorial Designs

In monofactorial hypothesis tests, one is usually interested in whether the value of the statistic in question of the dependent variable is dependent on the value of the independent variable; put differently, the question is, 'do the values of the independent variable make a difference for those of the dependent variable?' An example for a test for independence/differences of distributions would be the question of whether the similarities of the source words of blends (e.g., *breakfast* and *lunch* for *brunch*) are differently distributed from the similarities of the source words of complex clippings (e.g., *system* and *administrator* for *sysadmin*). If one was *not* interested in whether the average source word similarities are different, but just in the overall distributions of source word similarities – are the more larger values for blends? – then one could run a Kolmogorov–Smirnov test. If one *was* interested in the difference of average source word similarities of blends and complex clippings, then one could perhaps run a t-test for independent samples. If the source word similarities are not normally distributed and have different variances, one could perhaps run a U-test.

An example for a test for independence/differences of dispersions would be the question of whether two groups of subjects – say, native speakers and foreign language learners – exhibit differently variable sentence lengths (because, perhaps, the native speakers' higher competence allows them to exploit longer sentence lengths more for stylistic/expressive means); such a question could, if certain conditions are met, be tested with an F-test for variance homogeneity.

Finally, testing frequencies for independence/differences are often applied to frequency tables such as Table 1. If 100 subjects were asked "at what time does your shop close?" and an additional 100 subjects were asked "what time does your shop close?" one noted the frequencies with which one received the answers "at five o'clock" and "five o'clock," then one could test the resulting Table 1 with a chi-squared test, which would here return a significant result indicating that subjects prefer to use *at* when they were asked with *at*.

The above tests all have in common that the data evaluated are all independent; for example, no subject or no word provides more than one data point. In cases where this is not

Table 1 Fictitious data from a priming experiment

	Answer with at	*Answer without at*	*Totals*
Question with *at*	70	30	100
Question without *at*	45	55	100
Totals	115	85	200

the case, as in when subjects take a test before and after a treatment, alternatives to many of the above tests are available.

Multifactorial Designs

As mentioned above, statistical designs in many subdisciplines of linguistics now do more justice to the fact that linguistic choices are bound to always be affected by multiple causes. As a result, multifactorial approaches are now more common than ever. Most such studies involve different types of regression modeling, which can be considered state of the art at this point:

- *linear modeling*, a type of approach where the predictors can be categorical or numeric and where the dependent variable is numeric. In traditional references, such models may be treated under the headings of ANOVA (analysis of variance) or ANCOVA (analysis of covariance);
- *generalized linear modeling*, a type of approach where the predictors can be categorical or numeric and where the dependent variable can be of different forms: binary, ordinal, categorical/multinomial, or frequencies, which give rise to binary logistic regression, ordinal logistic regression, multinomial/polytomous regression, and Poisson regression respectively. (Strictly speaking, multinomial regression is not a type of generalized linear model, but for simplicity's sake and because it is closely related to a sequence of binary logistic regressions, it is included here in this bullet point.) This process still involves linear modeling after the application of a so-called link function to the dependent variable (which makes sure that the model generates only sensible predictions, e.g., no negative values are predicted for frequencies).

Regression analyses are usually conceptualized as shown in (1) and (2) using an example where a constructional choice by a speaker for an *of*- or an *s*-genitive (will the speaker say *the nut of the squirrel* or *the squirrel's nut*?) may be modeled as a function of the animacy of the possessor (*squirrel*), the animacy of the possessed (*nut*) and the difference of the lengths of the possessor and the possessed (say, in characters). The two regression equations begin with the dependent variable (Genitive), followed by a tilde meaning 'as a function of,' followed by all included predictors connected with pluses.

(1) Genitive (*of* vs *s*) ~ PossessorAnim + PossessedAnim + LenDiff
(2) Genitive (*of* vs *s*) ~ PossessorAnim + PossessedAnim + LenDiff + PossessorAnim:PossessedAnim + PossessorAnim:LenDiff + PossessedAnim:LenDiff + PossessorAnim:PossessedAnim:LengthDiff

The equation in (1) is multifactorial$_1$: multiple independent variables are studied at the same time. The equation in (2) is multifactorial$_2$: multiple independent variables and their interactions (indicated by the colons) are studied at the same time, which means that the regression in (2), but not the one in (1), can determine whether PossessorAnim has the same effect on the choice of genitive regardless of the level of PossessedAnim, whether PossessorAnim has the same effect on the choice of genitive regardless of what the length differences are, etc. The second type of modeling is statistically much more challenging, but it is often also the more revealing since many studies have shown that independent variables can interact in quite complicated ways. Often, researchers begin with a larger model such as that in (2) and then use various statistical operationalizations of Occam's razor (such as significance tests or information criteria) to trim the larger model down in a process called model selection to a so-called minimal adequate model; the model that, with some simplification, contains all, and only all, predictors that are significantly correlated with the dependent variable.

These regressions are immensely powerful and versatile tools and are now applied to many experimental and observational data sets (but see below). For instance, dependent variables that are reaction times or acceptability judgments would be analyzed with linear regression modeling; binary-dependent variables such as above in (1) and (2) and many other cases of syntactic alternations (such as the dative alternation, particle placement, preposition stranding, analytic vs synthetic comparisons, *will* vs *going to*, etc.) have been studied with binary logistic regressions; and dependent variables with more than two levels have been studied with multinomial regressions in, for instance, the semantic analysis of which of several near synonyms is the most likely choice given a particular context and why.

The above two kinds of models assume that all data points are independent of each other. However, especially in multifactorial settings, this is hardly ever the case: speakers provide more than one response in an experiment, verbs are reacted to by more than one subject, an author provides more than one constructional choice in a corpus file, etc. At the same time, these three variables – Speaker, Verb, and Author – are typically also different from variables such as TimeOfTest (before vs after), SexOfSpeaker (male vs female), or Animacy (animate vs inanimate): The levels of the latter variables (often called *fixed effects*) cover all the possibilities one would expect to exist in the population whereas the levels of the former variables (often called *random effects*) cover only the (ideally random) sample of speakers, verbs, or authors that was included in the study although a researcher would probably want to generalize beyond the set of speakers, verbs, or authors tested.

Dealing with this kind of scenario requires adjustments to the kind of (generalized) linear modeling where independence of data points is required. For quite some time, experimental psycholinguistic studies were heavily influenced by Clark (1973) and Forster and Dickinson (1976), two of the most widely cited early discussion of this problem in psycholinguistics. The current state-of-the-art approach in linguistics, however, involves what is called (generalized linear) *mixed-effects modeling*, or *multilevel modeling*. These models address the dependence by taking into account the variability that is associated with random effects. Consider a scenario in which one dependent variable is modeled as a function of one independent variable on the basis of 10 data points from three speakers each. **Figure 2** represents a regular linear model of some data in the left panel (the numbers 1, 2, and 3 represent the speakers' data points) and it is obvious that a linear model cannot account well for the data. In the right panel, a linear mixed-effects model is computed on the same data where every speaker gets his own intercept but they all share the same slope. In that panel, the dashed line reflects the overall trend and the three straight lines are the separate regression lines for each

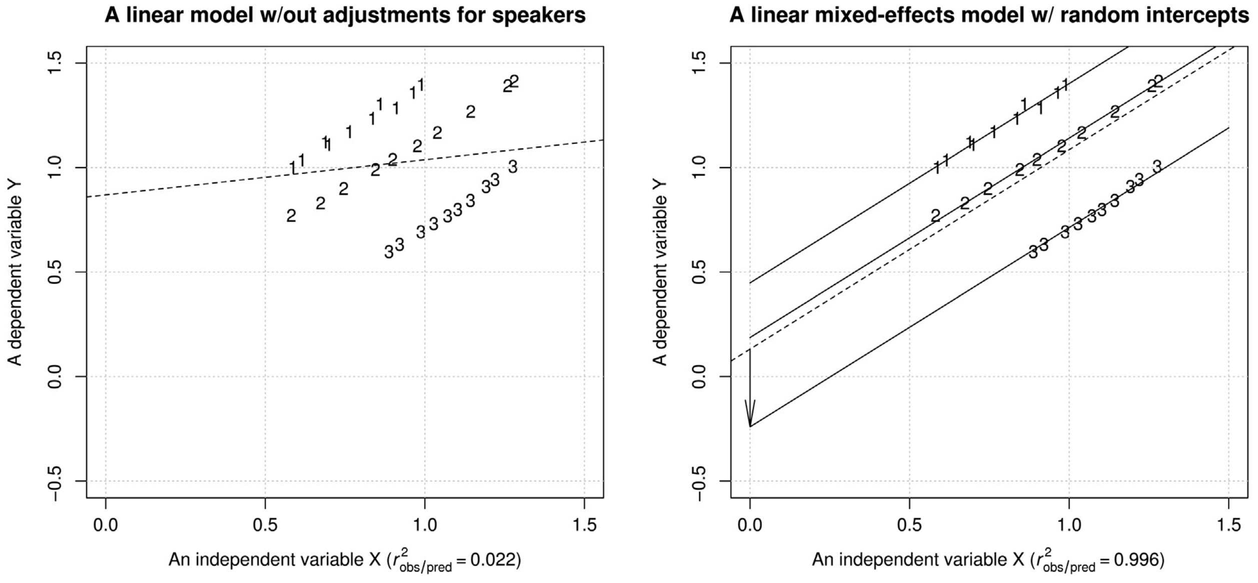

Figure 2 A linear model and a linear mixed-effects model (random intercepts) analysis of a fictitious data set.

speaker; the difference between the overall trend and the regression line for speaker 3 is represented by the arrow. Again, it is obvious that this model, which takes the relation between data points of a speaker into consideration, accounts very well for the data.

More complex versions of such mixed-effects model are available, e.g., models where speakers do not just get their own intercepts, but also their own slopes, and these adjustments can be included for crossed and nested variables that have made this a very powerful and popular approach in a very wide variety of experimental and observational studies. While some details are currently still being worked out – how to compute *p*-values for predictors in some models, how to do model selection with mixed effects, how many different random effects to include, etc. – mixed-effects modeling is bound to become one of the most important tools of the trade.

While regression models are probably the most widespread tool for multifactorial statistical analysis, another increasingly popular method is that of *classification and regression trees* or related, but more advanced, methods such as *random forests*. Such trees are based on the idea of successively splitting the data in a binary fashion such that each split groups the data based on one independent variable into two groups that predict the behavior of the dependent variable best. The output of such an analysis is typically a prediction/classification accuracy and a decision tree-like structure and the analyst interprets the tree in a top-down fashion to see which variable splits explain the dependent variable. This method is sometimes used an exploratory precursor to regression modeling, but can also be illuminating in its own right.

Desiderata

The field of linguistics has evolved enormously with regard to the number and sophistication of hypothesis-testing techniques that are now used regularly. However, given the recency of this development, many of the techniques mentioned above need to become more fleshed out in how they are applied to linguistic data and more entrenched among a larger number of practitioners. In addition to these developments, there is also a need for linguists to recognize the multitude of techniques that are already regularly applied in fields that struggle with data that pose the exact same distributional challenges that linguists face every day. Typically, psychology is mentioned as one of the closest neighboring fields whose statistical tools linguists might inspire – however, the field of ecology is maybe an even better contender (cf Zuur et al., 2007, 2009 for examples). This section will briefly outline a few methods that more linguists ought to pay more attention to, given how they can help tackle data and questions that are sometimes hard to address with the currently established tools.

One very basic but still quite useful improvement would involve linguists realizing more the power of regression approaches. For example, if a researcher has done a corpus study and represented his data in a 2 × 2 frequency table and then wanted to compare these results to another 2 × 2 frequency table from someone else's previous study, then he might resort to a heterogeneity chi-squared test to see if both data sets may represent the same population. However, not only is this somewhat cumbersome, but the heterogeneity chi-squared test also does not generalize to $r > 2 \times c > 2$ frequency tables. Instead, one natural solution may be to just analyze the data with a regression and test whether the two variables tested for in the chi-squared test also interact with a third variable that indicates which study the data are from. If there is no such significant three-way interaction, then the two data sets can be argued to represent the same population, otherwise, they cannot. Thus, while it may feel like an exaggeration to run a binary logistic regression where a chi-squared test would be entirely appropriate, adopting the regression perspective allows one to pursue a wider range

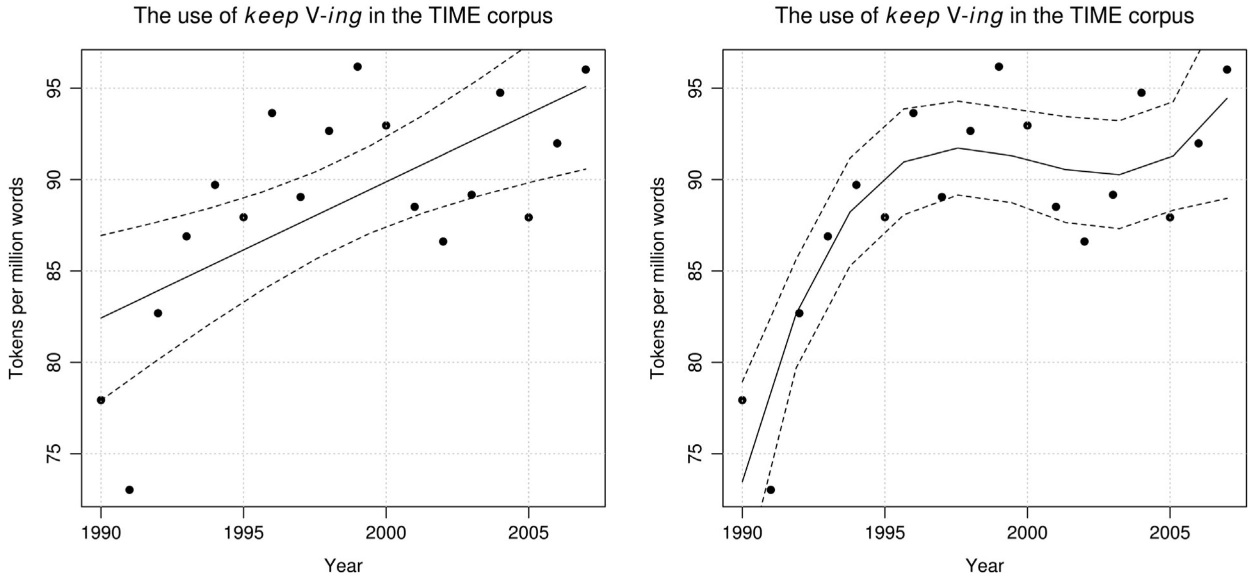

Figure 3 A linear regression model without polynomial predictors (left panel) and a linear regression model with a predictor as a third-degree polynomial (right panel).

of questions with a single unified modeling approach. (See below for another similar example.)

While conceptualizing some data sets in terms of regressions is already useful, there are of course also extensions and additional methods. Some of these extensions/methods are very closely related to regression approaches. One kind of particularly useful extension allows researchers to handle curvature in data or nonlinear effects. One example would be to include polynomial terms in regression analyses, another more general one is *generalized additive modeling*. Consider Figure 3 for data on the frequency of the *keep* V-*ing* construction in the TIME corpus (from Hilpert, 2011). While the left panel shows that it is possible to fit a simple linear regression to these data (and even obtain a significant result), the right panel shows that a polynomial to the third degree provides a much better fit with the data and suggests that, after 1995, the growth trend seems to be leveling off.

Another technique linguists may consider using much more is *general linear hypothesis test*. This is a test that can be used as follow-up analysis of a regression model that involves at least one categorical independent variables with three or more levels. Even if such a categorical variable makes a significant contribution to a dependent variable, this does not mean that all its levels need to be maintained. For example, a researcher may distinguish five levels of animacy in his coding of corpus data – e.g., human and deities, animate but not human, plants, concrete objects, abstract entities – and the corresponding variable ANIMACY might be significant. Then, one can use general linear hypothesis tests to determine whether all five levels of ANIMACY differ from each other or whether Occam's razor would in fact require to conflate the first two and the next two, which would result in a new version of ANIMACY with only three levels.

Interestingly enough and as mentioned above, this logic does not only apply to, and would benefit, complex multifactorial regression models – on the contrary: Instead of submitting a, say, 4 × 2 frequency table to a chi-squared test, one could analyze these data with a Poisson regression and follow up with general linear hypothesis tests to determine whether all four levels of the independent variable are indeed necessary/justified.

Then, there are many statistical tools with huge potential for linguistics. One is *naïve discriminative learning* (NDL) as discussed by Baayen (2010). The overall goal is very similar to that of (generalized) linear mixed-effects modeling but, unlike regression modeling, this algorithm is based on a general model of human probabilistic learning and equations rooted in research on associative learning. As Baayen shows, NDL, while cognitively better grounded than traditional regression models, achieves classification accuracies comparable to other models and can even handle random effects in insightful ways.

Another set of relevant tools includes methods that are designed to address more explicitly the question of causal relations between predictors and responses. Strictly speaking, regression models of the above type only speak to whether there are significant correlations between the predictors on the one hand and the response on the other. However, techniques such as *structural equation modeling* or *Bayesian networks* can help study causal relations better by forcing the researcher to explicitly formulate expectations about how predictors are intercorrelated with each other and correlated with the dependent variable. For example, it is well known that many syntactic/constructional alternations such as the dative alternation or particle placement are correlated with phonological predictors (e.g., stress and syllabic length), syntactic predictors (e.g., definiteness and complexity), and discourse–functional predictors (e.g., givenness/inferrability of referents). However, these predictors are all correlated with each other, which not only poses problems to most regression models but also makes it harder to disentangle cause–effect relations. Applying the above kinds of tools would require the researcher to hypothesize, for example, that discourse-functional characteristics are causes for phonological and syntactic effects, which in turn are causes for syntactic/constructional choices – rather than treating all predictors as mere causes and only the syntactic/constructional choice as an effect. Since usually causal relations are what we are most interested in, these techniques hold high promise.

Finally, a more general recommendation is that linguists may want to become more familiar with the field of robust statistics (cf Wilcox, 2012). Since linguistic data are rarely well-behaved in the sense of being normally distributed (instead, many phenomena are Zipfian-distributed), not suffering from outliers (many linguistic variables such as lengths and frequencies regularly have outliers), and since grouped data often differ in their dispersion, etc., robust statistics offer ways to analyze data with all these properties that undermine the use of the currently more widespread regular statistical tools.

Hypothesis-Generating Methods

The techniques discussed so far are all usually employed to subject hypotheses to significance testing and decide on which hypothesis to adopt based on a *p*-value. A different set of techniques is concerned with generating hypotheses in the first place. That is, these techniques serve exploratory functions and are often applied to data sets whose size and complexity defies an analyst's eyeballing and pattern-matching skills. Again, this section will begin with a discussion of what are arguably the currently most frequently used methods before turning to some refinements and desiderata that would benefit the field; as before, given that the number of techniques is vast, the overview has to be quite selective.

Exploratory Methods in Common Use

The probably most widespread exploratory technique is *hierarchical cluster analysis* (HCA). As most exploratory methods, this one, too, is typically applied with an eye to determining how *n* entities – sentences, NPs, words, subjects, etc. – can be grouped into $m < n$ groups/clusters that exhibit high within-group similarity and low similarity to other groups.

HCAs involve several steps: First, one computes how similar each of the *n* items is to each other item based on a user-defined similarity metric. The results of this step is usually represented in a (dis)similarity matrix. Crucially and unlike other exploratory tools, the items can be compared even if they involve both categorical and numeric characteristics (although numeric characteristics are most frequent). Second, in an iterative process, the pairs of items with the highest similarity values are identified and merged into groups. Once an item enters into a group, it cannot be considered in isolation anymore – what is considered for the next merger is then the newly formed group. In the most frequent type of HCA, this process is repeated until all items have been merged into one group. Third, this grouping process is visualized with a so-called dendrogram and the researcher has to decide how many clusters the dendrogram reflects most revealingly. An example is represented in Figure 4, which is the result of clustering seven nouns on the basis of the frequencies of words they co-occur with in corpus data. Both semantic considerations – the left cluster contains precious-metal nouns, the right one gastronomical terms (although note the ambiguity of *bar*) – and follow-up diagnostics to be mentioned below suggest that assuming the two clusters highlighted with gray boxes is most useful here.

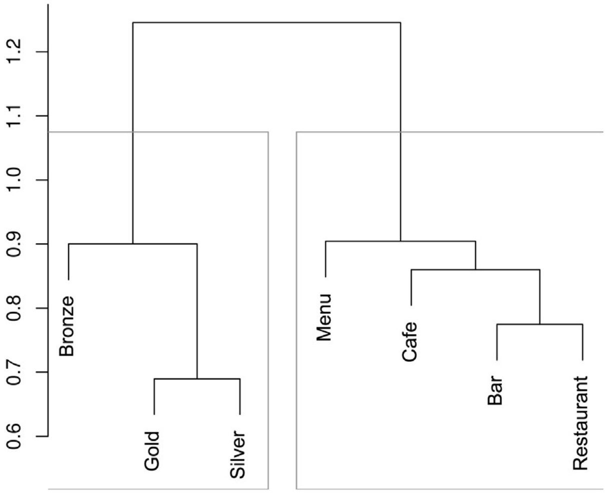

Figure 4 Results of an HCA of seven nouns on the basis of the frequencies of words they co-occur with.

HCA is a very flexible tool, suitable for very many different kinds of both experimental and observational data and often relatively straightforward to interpret, and they are becoming increasingly frequent in linguistic research.

Another frequent technique is *principal component analysis* (PCA, and its 'sibling' principal factor analysis). The overall goal of a PCA is quite similar to that of an HCA, but the underlying mathematics are quite different and PCA is only applied to all-numeric data. The input data is the same as for an HCA and a PCA proceeds by computing all possible intercorrelations between the *n* items/columns being studied. Then, the PCA merges/compresses the *n* items/columns into $m < n$ principal components, such that the items/columns within a principal components are highly intercorrelated but all principal components are usually mutually orthogonal, or independent of each other. This can mean, for example, that a PCA would convert a data set in which NPs are classified according to $n = 20$ characteristics into a new data set where, say, only four principal components retain more than 90% of the information of the originally 20 columns. This kind of result is why PCA and other similar methods – for example, *multidimensional scaling* or *correspondence analysis* – are also referred to as dimension-reduction methods. A researcher can then 'just' analyze the nature of the principal components – what are the linguistic characteristics that make up the principal components and why? – and often this interpretive task is facilitated by a two-dimensional visual representation of the results, as exemplified in Figure 5, where the seven nouns are grouped in a way that resembles the above results of the HCA.

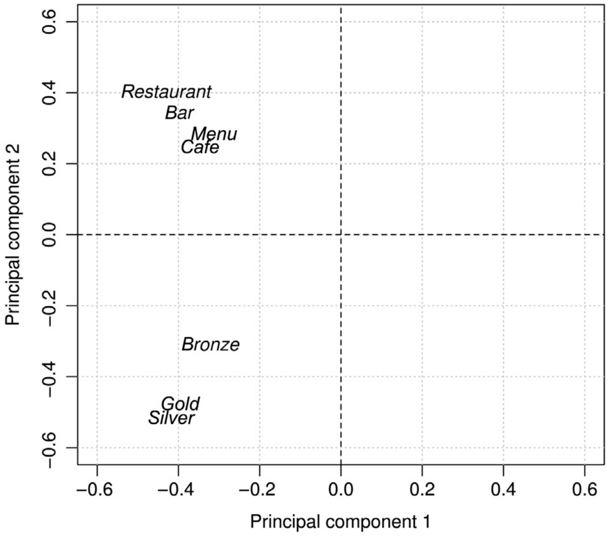

Figure 5 Results of a PCA of seven nouns on the basis of the frequencies of words they co-occur with.

Alternatively, the compressed data set can also be used as input to hypothesis-testing methods, which are now more likely to return useful results, given that instead of 20 intercorrelated variables, now only 4 uncorrelated variables, the principal components, are studied. In corpus linguistics, factor analyses have been extremely influential in, for example, Biber's (1995) work on register/genre variation, but many studies in computational semantics employ mathematically very similar methods.

Desiderata

Just as above, the desiderata regarding exploratory methods come in two kinds: First, as extensions of, or follow-ups to, the methods discussed above; second, methods in addition to the ones discussed above.

As for the former, while cluster analyses enjoy quite some use now, the results they provide are often not yet analyzed to their fullest extent. For instance, not all dendrograms are easy to interpret in terms of how many clusters a researcher should assume but very few studies discuss principled ways of arriving at such decisions. One approach to this question involves a statistic called average silhouette widths, which quantifies how similar elements are to the clusters which they are in relative to how similar they are to other clusters; in the above case of Figure 4, this statistic 'recommends' to assume the highlighted two clusters. Another approach toward the same problem involves resampling approaches that, ultimately, assign a *p*-value to each possible cluster so that researchers can pick the most strongly supported and most robust clusters for interpretation. In spite of the availability of such methods, they are still too rare, which can render HCA results less insightful than they would need to be. Finally, one rarely sees decisions for a particular number of clusters validated by some other statistical technique. For instance, if a cluster analysis really reveals two clusters, then a binary logistic regression (or some other non-clustering technique) should also be able to recognize these two clusters in the data – attempts at validation like these are still hard to find.

A different set of extensions involves the interpretation of the clusters: what in the data is most responsible for a particular cluster; what gives rise to the structures; which variables are reflected most in the clustering? There are diagnostic statistics available to answer such questions, but they have not yet made their way into the relevant research literature.

As for the latter kind of desiderata, there are several relatively new methods that should be highly useful to linguistic research. One of these is actually another cluster algorithm, but one with underutilized properties. This approach, so-called *fuzzy clustering*, allows for items to be members of more than one cluster and to different degrees, which is often a much more realistic assumption than the seemingly clear-cut divisions of items into clusters suggested by traditional dendrograms. Another technique that might be very useful is that of *association rules*, a data mining procedure suitable to uncover co-occurrence and *if … then* relations. The most attractive features of association rules are their conceptual simplicity and that this method can handle extremely large data sets (think Amazon shopping-basket data) consisting of categorical variables, which are not amenable to multiple cross-tabulation anymore. Thus, especially large-scale corpus data might benefit from this method.

Concluding Remarks

It is useful to conclude by mentioning a few, more general desiderata which, if they were more widely implemented or adhered to, would also tremendously help the discipline of linguistics evolve. The first of these is concerned with the fact that we not only need to broaden and deepen our knowledge of statistical methods per se, but we also need more awareness of the distributional assumptions that many methods come with. Many published studies are quite cavalier when it comes to these, but, as discussed at length in the above-mentioned Wilcox (2012) study, failure to deal with violations of such assumptions can lead to hugely anticonservative results.

In addition, the field also needs more and firmer guidelines on what is important in statistical analysis, what is important to report (e.g., checking distributional assumptions), and how methods and results should be reported. Other fields have had long and intense discussions about these issues whereas linguistics, for the most part, has not. We should be prepared to be inspired by how other disciplines with similar kinds of questions and data have come to grips with these challenges and recommendations such as Wilkinson and the Task Force on Statistical Inference (1999) provide many essential tips (e.g., to always include effect sizes to distinguish significance from effect size and make analyses comparable).

Finally, two areas in statistics that linguists should strive to learn about more: (1) developments in *resampling approaches* such as bootstrapping, which can be quite useful to enhance the precision and generalizability of our results; (2) *Bayesian statistics*, which adopts a perspective on statistical analysis that is very different from the predominant null-hypothesis significance testing approach, but which is ultimately more in line with what we as researchers are interested in: the confirmation of reasonable hypotheses and the quantification of effects on the basis of accumulated knowledge. Given the diversity and complexity of linguistic data, any kind of tool can only be a welcome addition to our toolbox.

See also: Bayesian Statistics; Hypothesis Testing in Statistics; Invariance in Statistics; Likelihood in Statistics; Mixture Models in Statistics; Nonparametric Statistics: Advanced Computational Methods; Order Statistics.

Bibliography

Baayen, R. Harald, 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge.

Baayen, R. Harald, 2010. Corpus linguistics and naïve discriminative learning. Brazilian Journal of Applied Linguistics 11 (2), 295–328.

Biber, Douglas, 1995. Dimensions of Register Variation. Cambridge University Press, Cambridge.

Clark, Herbert H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12 (4), 335–359.

Forster, Kenneth I., Dickinson, R.G., 1976. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and min F′. Journal of Verbal Learning and Verbal Behavior 15 (2), 135–142.

Gries, Stefan Th, 2013. In: Statistics for Linguistics Using R: A Practical Introduction, second rev. & ext. ed. De Gruyter Mouton, Berlin & New York.

Hilpert, Martin, 2011. Diachronic collostructional analysis: how to use it and how to deal with confounding factors. In: Allan, Kathryn, Robinson, Justyna (Eds.), Current Methods in Historical Semantics. Mouton de Gruyter, Berlin & New York, pp. 133–160.

Johnson, Keith, 2008. Quantitative Methods in Linguistics. Blackwell, Malden, MA & Oxford.

Larson-Hall, Jenifer, 2012. A Guide to Doing Statistics in Second Language Research Using R. Routledge, New York, London. http://cw.routledge.com/textbooks/9780805861853/guide-to-r.asp.

Oakes, Michael P., 1998. Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh.

Rietveld, Toni, Hout, Roeland van, 2005. Statistics in Language Research: Analysis of Variance. Mouton de Gruyter, Berlin & New York.

Theijssen, Daphne, Louis ten, Bosch, Lou, Boves, Bert, Cranen, Hans van, Halteren. to appear. Choosing Alternatives: Using Bayesian Networks and Memory-based Learning to Study the Dative Alternation. Corpus Linguistics and Linguistic Theory 9 (2), 227–262.

Wilcox, Rand, 2012. Introduction to Robust Estimation and Hypothesis Testing, third ed. Elsevier, Amsterdam.

Wilkinson, Leland, The Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: guidelines and expectations. American Psychologist 54 (8), 594–604.

Woods, Anthony, Fletcher, Paul, Hughes, Arthur, 1986. Statistics in Language Studies. Cambridge University Press, Cambridge.

Zuur, Alain F., Ieno, Elena N., Smith, Graham M., 2007. Analysing Ecological Data. Springer, Berlin & New York.

Zuur, Alain F., Ieno, Elena N., Walker, Neil, Saveliev, Anatoly A., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin & New York.