8

Statistics for learner corpus research

Stefan Th. Gries

1 Introduction

Over the last decades, second/foreign language acquisition (S/FLA) has become an ever larger, more diverse, and more productive discipline. This evolution notwithstanding, for most of that time SLA research seems to have favoured experimental and introspective data over the exploration or analysis of corpus data (cf. Granger 2002: 5). Fittingly, Mackey and Gass (2005), for example, devote not even two pages to the topic of corpora in learner corpus research (LCR) (in Chapter 3, which is nearly sixty pages long), and in Chapter 9, which covers quantitative methods of analysis, corpus data play no role (the later Mackey and Gass (2012) includes a chapter on LCR, however). Similarly, Tyler (2012) discusses many experimental results in great detail but summarises a mere handful of corpus studies. Despite this neglect, corpus data have now become a major source of data in S/FLA research, both on their own and in combination with experimental data. This is in particular due to the increasing availability of corpora of learner language (most of them on learner English), which offer researchers the opportunity to study a wide range of questions regarding:

- how learners from different mother-tongue (L1) backgrounds use English in speaking and writing
- how the use of English by learners with a particular L1 differs from that of learners with other L1s
- how the use of English by learners differs from that of native speakers.

However, corpora contain nothing but frequency data: they reveal whether linguistic element *x* does or does not occur in a corpus ($n_x > 0$ or $n_x = 0$), whether *x* occurs in a part *a* of a corpus (e.g. a register, dialect, variety, speaker group) or not ($n_{x \text{ in } a} > 0$ or $n_{x \text{ in } a} = 0$), whether *x* occurs with *y* or not ($n_{x \text{ and } y} > 0$ or $n_{x \text{ and } y} = 0$). Thus, whatever a (corpus) linguist is interested in needs to be (i) operationalised in terms of frequencies of (co-)occurrence and (ii) analysed with the tools of the discipline that deals with quantitative data, statistics.

In this chapter, I will survey the ways in which corpus-based research in SLA has utilised, or has yet to utilise, statistical methods. While I will attempt to cast a wide net and cover a variety of different approaches and tools, this survey can, of course, only be selective.

2 Core issues

2.1 Statistical methods in LCR

The simplest kind of statistics in (corpus) linguistics is general *descriptive statistics*, i.e. statistics describing some state of affairs in the data. The most frequent ones in LCR include:

- *frequencies* of occurrence of linguistic elements as observed frequencies, as normalised frequencies (per cent, per thousand words, per million words), ranks of such frequencies or statistics computed from such frequencies (e.g. type–token ratios, vocabulary richness/growth statistics)
- frequencies of co-occurrence or association measures that do not involve statistical significance testing like *mutual information* (*MI*) or odds ratios; such measures quantify the association of one linguistic item (typically a word) to another (typically a word or a syntactic pattern/construction), in which case we talk about *collocation* or *colligation/collostruction*, or the association of a word to one of two corpora (which is what, statistically, the method of *keywords* boils down to)
- *measures of central tendencies* such as means or medians
- *dispersion measures*, which should accompany averages, such as standard deviations, standard errors, median absolute deviations, or interquartile ranges
- *correlation measures* such as Pearson's *r* or Kendall's τ.

Second, there are tools from the domain of *inferential statistics* in the form of statistical tests returning *p*-values (determining how likely an obtained result is due to chance variation alone) or in the form of confidence intervals (providing likely ranges into which observed results may fall); the most common ones involve:

- *significance tests of two-dimensional frequency tables* involving chi-square tests or, much more rarely, Fisher-Yates or similar exact tests
- *association measures* that do involve significance tests (e.g. log-likelihood ratio G^2, *z*, *t*, see Evert 2009)

- *significance tests for differences* between measures of central tendencies involving *t*-tests, *U*-tests, or Kruskal-Wallis tests as well as *significance tests for correlations*.[1]

Currently, most of the statistics used in LCR are covered by these categories, but more advanced tools are available. First, in inferential statistics, there is the area of *multifactorial regression modelling*. A multifactorial regression is a statistical model trying to predict a dependent variable/response (often the effect in a hypothesised cause–effect relationship, either a numeric variable or a categorical outcome such as a speaker's choice of one of two or more ways of saying the same thing) on the basis of multiple independent variables/predictors (usually the potential causes of some effect), using a regression equation. Such a regression helps to quantify each predictor's significance and/or importance ('does this predictor help make the prediction more accurate or not and how much so?') and direction of effect ('which of the possible outcomes does this predictor make more likely?'). Thus, a regression equation is little more than the mathematical way of expressing something such as *If a possessor is animate and a possessee is inanimate, then the speaker is* x *times more likely to encode this relation with a possessive* s*-genitive than with an* of*-genitive*. This type of approach – as well as its 'sister approaches' of classification trees and other classifiers – is extremely powerful in allowing researchers to investigate the impact of multiple predictors on a linguistic choice simultaneously, but it is still very much underutilised; this method and its advantages will be discussed in detail below.

Second, there is the area of *multivariate exploratory tools*, such as hierarchical cluster analysis, principal components analysis, correspondence analysis, multidimensional scaling, and others. These methods do not try to predict a particular outcome such as a speaker's choice on the basis of several predictors and typically do not return *p*-values from significance tests – rather, they find structure in variables with an eye to allowing researchers to detect groups of variables/expressions that are similar to each other but different from everything else. Such results can then either be interesting in their own right or inform subsequent (regression) modelling. I will discuss these techniques very briefly below, too.

The next section will survey some studies that have utilised some of these methods.

2.2 Applications involving simple descriptive statistics

2.2.1 Frequency data

Just about every empirical learner corpus study reports some kind of frequency data. Consider, as a first example, Hyland and Milton (1997),

[1] See Gries (2013a) for detailed hands-on explanation of how these statistics are computed in the context of LCR and Gries (2013b) for statistics in linguistics in general.

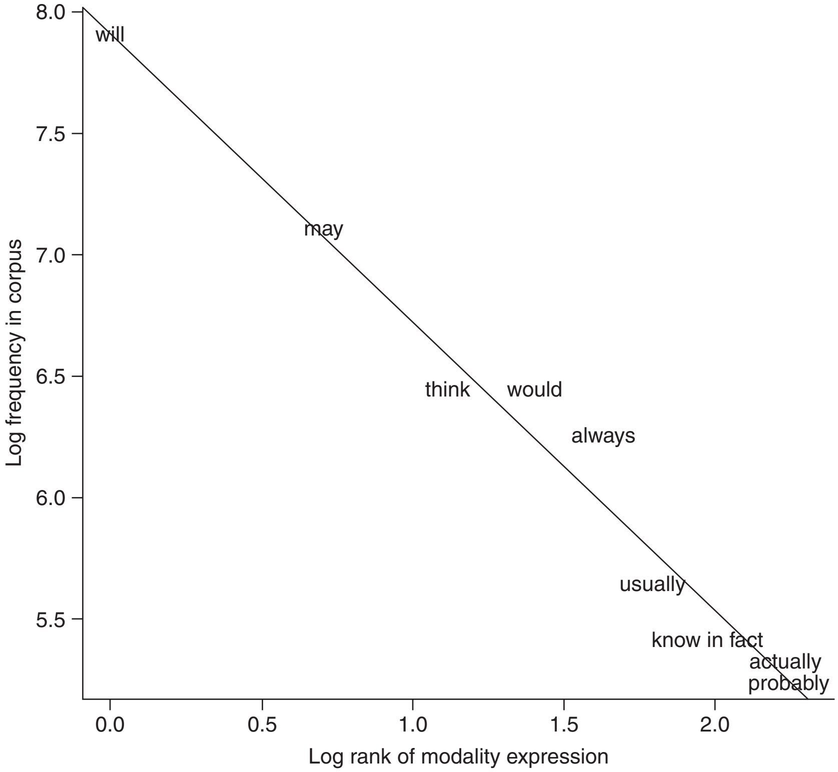

Figure 8.1 Visualisation of the data in Hyland and Milton (1997: 189)

who compare ways in which native speakers (NS) and non-native speakers (NNS), here Cantonese-speaking learners of English, express modality. Among other things, they report overall frequencies of expressions of epistemic modality, finding that NS and NNS exhibit considerable similarities of usage. In addition, they used sorted top-10 frequency lists of epistemic modality expressions for NS and NNS. While they do not dwell on this, especially for the NNS they find a very Zipfian distribution, i.e. a distribution that is highly typical of linguistic data where a small set of the types (here, the top ten types) accounts for a large proportion (here 75%) of the total tokens. Their NNS data are visualised in Figure 8.1 with the log of the frequency of an expression on the *y*-axis, the log of the rank of the frequency of an expression on the *x*-axis, and the expressions plotted at their coordinates.

Further statistics they report include normalised frequencies (per thousand words) for different grammatical ways of expressing epistemic modality (e.g. with modal or lexical verbs, adverbials, adjectives or nouns) or these frequencies grouped into different ability grades.

An example whose orientation is representative of much current LCR is Hasselgård and Johansson (2011). Like many learner corpus studies, they report different kinds of frequency data with an eye to

Table 8.1. Raw/normalised frequencies per million words (pmw) of *quite* (from Hasselgård and Johansson 2011: 46)

	LOCNESS	*ICLE-SP*	*ICLE-FR*	*ICLE-NO*	*ICLE-GE*
Frequency	67	63	78	92	147
Frequency pmw	205	318	380	437	623

Table 8.2. Laufer and Waldman's (2011: 660) extended frequency data on V–N collocations

	LOCNESS	*ILCoWE*: advanced	*ILCoWE*: intermediate	*ILCoWE*: basic	Totals
V–N collocations	2,527	852	162	68	3,609
Non-collocations	22,242	12,953	2,895	1,465	39,555
Totals	24,769	13,805	3,057	1,533	43,164

over-/underuse in the learner data, as well as results of simple significance tests. For instance, one of their case studies is concerned with the frequencies of *quite* in the *Louvain Corpus of Native English Essays* (*LOCNESS*) and four components of the *International Corpus of Learner English* (*ICLE*) (Norwegian, German, Belgian-French and Spanish) shown in Table 8.1. They report the results of chi-square tests comparing each frequency from the *ICLE* components to the *LOCNESS* frequency and state that '*quite* is overused in all the learner groups', that all learners but the Spanish ones differ significantly from the NS data and that 'the overall frequency distribution … thus seems to reflect the Germanic–Romance distinction' (pp. 45–6).

Then, there is interesting work bridging the gap from frequency statistics to association measures, namely research on collocations that does not involve measures of collocational strength but, for instance, collocational dictionaries. One such example is Laufer and Waldman (2011). They compare the use of verb–noun (V–N) collocations by NS (*LOCNESS*) with that of learners in the *Israeli Learner Corpus of Written English* (*ILCoWE*); verb–noun candidates were considered a collocation if they were listed in at least one combinatory/collocational dictionary. Laufer and Waldman then test whether NS and NNS differ with regard to the number of V–N collocations with a chi-square test and find that the NS produce significantly more V–N collocations. They then proceed to group the NNS into three proficiency groups as represented in Table 8.2 and conduct eight different chi-square tests on this table. They summarise their results by stating that the NS produce significantly more collocations than the learners and that, within the NNS, only the advanced and basic NNS differ from each other significantly.

Other studies also largely based on raw and normalised frequencies of elements and basic statistical comparisons of frequencies include Altenberg (2002) and Götz and Schilk (2011). The former is concerned with the uses of causative *make* by American NS as well as French and Swedish NNS. Altenberg reports many tables of observed frequencies and percentages that indicate how learners differ in their use of causative *make* from the American NS and each other, which Swedish equivalents of causative *make* are used how often, which English equivalents of causative *göra* are used how often, etc. Götz and Schilk (2011) contrast the frequencies of 3-grams in spoken L1 English from the British component of the *International Corpus of English* (*ICE-GB*) and the *Louvain Corpus of Native English Conversation* (*LOCNEC*), in spoken L2 English from the Indian component of the *ICE* (*ICE-IND*), and in spoken learner English from the German component of the *Louvain International Database of Spoken English Interlanguage* (*LINDSEI-GE*), and then perform G^2-tests to determine which observed frequencies differ from each other. Another similar example is Gilquin and Granger (2011), who explore the use of *into* across four *ICLE* subcorpora – the Dutch, French, Spanish and Tswana components – and in NS English. They, too, report relative frequencies of *into* per 100,000 words and comparisons using the G^2-statistic.

2.2.2 Association measures and other (monofactorial) significance tests

A different group of studies involves frequency data but uses them more as a basis for association measures quantifying how much two elements are attracted to, or repelled by, each other; as mentioned above, some of these association measures involve statistical significance tests (see Evert 2009). Sometimes, such studies also utilise simple monofactorial statistics, e.g. significance tests for measures of central tendencies or correlations.

One study that fruitfully combines different statistical tools is Zinsmeister and Breckle (2012), who explore the annotated learner corpus (*Annotiertes Lernersprachenkorpus*, *ALeSKo*) of German essays produced by NS and NNS (Chinese) learners. Apart from a comparison of frequent 3-grams, they also discuss frequencies of part-of-speech 3-grams. As a cut-off point for over-/underuse, they do not use the G^2-test ('because of the small size of the *ALeSKo* corpus' p. 84, n 25) but the difference between ranks in frequency lists. However, they also use several more sophisticated tools: to study the lexical complexity of their corpora, they compute and test type–token ratios and vocabulary growth rates for both subcorpora. Similarly, they compute summary statistics for both and test for significant differences using non-parametric *U*-tests.

Durrant and Schmitt (2009) is another interesting case in point. They compare the use of adjective–noun and noun–noun collocations by Bulgarian learners of English with that of NS, which were extracted from essays and whose strength was quantified using the *t*-score (to highlight more frequent collocations) and *MI* (to highlight less frequent collocations). These values were classified into seven and eight bands, respectively, so that the authors could explore how much NS and NNS use collocations of particular strengths with *t*-tests. Results for the *t*-scores indicate that NNS make greater use of collocations in terms of tokens (but that this is in part due to their overuse of some favourite collocations), whereas results for *MI* indicate that NNSs make less use of collocations in terms of tokens.

Apart from collocational studies, S/FLA research has also begun to target colligations/collostructions, i.e. the association of words to syntactic patterns. One of the first studies to explore verb-construction associations is Gries and Wulff (2005), who compare the attraction that verbs exhibit to the ditransitive and the prepositional dative constructions in NS corpus data (based on Gries and Stefanowitsch's (2004) distinctive collexeme analysis) to advanced German learners' sentence-completion behaviour in a priming study (finding a significant positive correlation), but also to the constructional preferences of the German translational equivalents of these verbs (finding no correlation). This is interesting because it suggests that the tested German learners have internalised the frequency distributions of English verbs in constructions rather than falling back on what their L1 would have them do.

A final related example is Ellis and Ferreira-Junior (2009a). They study six different hypotheses regarding the acquisition of three verb-argument constructions: the verb-locative construction (e.g. *My squirrel walked into the kitchen*), the verb-object-locative construction (e.g. *My squirrel carried the nuts into the kitchen*), and the ditransitive construction (e.g. *My squirrel gave the other squirrel a nut*). More specifically, their study is concerned with the distribution of verbs in the verb slots of these constructions (is it Zipfian?) and whether the first-learned verbs in the constructions are more frequent in, more strongly attracted to, and more prototypical of, the construction. Their study is one of the first to use a directional measure of association, ΔP, i.e. an association measure that does not quantify the association between two elements x and y in a bidirectional fashion, but makes it possible to distinguish the association from x to y from the one from y to x (cf. Gries 2013c). Their exploration is based on the *European Science Foundation Second Language Database* and shows that the type–token distributions in the verb slots are Zipfian and that first-learned verbs are highly frequent in, strongly attracted to, and prototypical of, the respective constructions.

2.3 Applications involving multifactorial statistics (regression modelling)

Very recently, LCR has begun to recognise the power of regression approaches and researchers are becoming familiar with the basic logic underlying regression-analytic approaches. Regression approaches of the type mentioned above offer many advantages:

- as mentioned above, they allow us to include multiple predictors in an analysis
- with multiple predictors, one can explore interactions between variables, i.e. one can test whether one variable has an effect on how another variable is correlated with the dependent variable; also, non-linear effects can be explored
- regression modelling provides a unified framework to understand many seemingly unrelated tests. For instance, instead of trying to learn many monofactorial tests (e.g. chi-square tests, *t*-tests, Pearson correlations, *U*-tests) and then regression modelling separately, it is useful to understand that monofactorial tests can often be seen as the simplest possible cases of a monofactorial regression
- while regression modelling is typically used in a hypothesis-testing context, there are extensions that allow the researcher to also perform (guided) exploration of the data
- regressions generate predictions (with confidence intervals) of how a response will behave, which often allows for seamless integration of results from different studies of whatever type (observational, experimental, simulations, etc.).

The remainder of this section is devoted to exemplifying these advantages. As a first simple example, let us return to Hasselgård and Johansson's (2011: 46) data shown in Table 8.1. Even a simple one-dimensional frequency list such as this one can benefit from a regression-analytic approach. Here, where one is interested in frequencies, one useful kind of regression is a monofactorial Poisson regression (cf. Gries 2013b: Section 5.4.3). Such a regression tries to predict, or model, the frequencies of *quite* (the dependent variable) in each of the different corpora (the independent variable), as shown in (1).[2]

(1) a. Freq ~ Corpus (L1 vs SP vs FR vs NO vs GE)
b. dep. variable as a function of predictors

If this approach is applied to Hasselgård and Johansson's data, one finds that indeed all learner varieties are significantly different from the *LOCNESS* baseline. However, as mentioned above, one can now undertake more detailed exploration using so-called general linear hypothesis tests – a method that allows the researcher to test, for instance, whether

[2] An offset was included to account for the fact that the corpus sizes differ, but this does not affect the general logic.

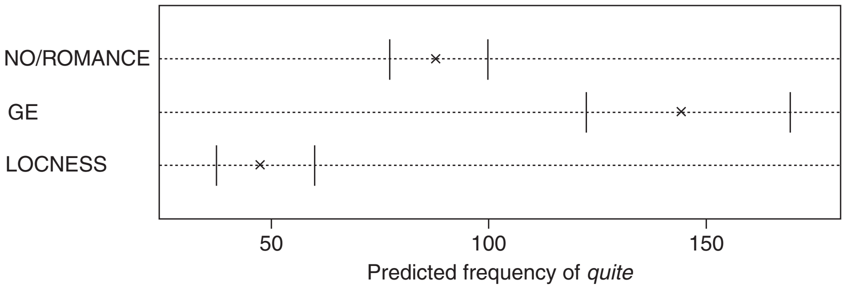

Figure 8.2 Visualisation of the final model on Hasselgård and Johansson's (2011: 46) data

different L1 data differ from each other significantly (cf. Bretz et al. 2010). The results suggest that the two Romance languages can indeed be conflated without a significant loss of accuracy, but that the two Germanic languages cannot. If one followed Occam's razor, one would therefore conflate the two Romance languages in a second regression model, which then reveals that (i) the two Germanic languages do not behave similarly, but that (ii) the Norwegian data are not significantly different from the two Romance languages' frequencies. The final results of a third model that conflates the Norwegian and the two Romance data points shows results quite different from Hasselgård and Johansson (see Figure 8.2): the postulated Germanic–Romance distinction collapses because (i) the two Germanic languages do not behave identically and (ii) the two Romance languages are not different from Norwegian.[3]

More interesting applications involve additional complexity and result in powerful explorations of learner corpus data. As discussed above, this 'additional complexity' can result both from different independent linguistic variables or their interactions. However, another crucial level of complexity arises when the corpus source, or speaker group or L1, is not only included as a predictor but also allowed to interact with all others. This step is simultaneously the most important and most underutilised one; the present discussion borrows from Gries and Deshors (2014).

Imagine a regression where one tries to predict the choice of *may* or *can* – let's call this variable FORM – on the basis of two linguistic predictors: NEGATION (whether the clause in which the speaker has to choose *may*/*can* is negated or not) and ASPECT (whether the clause in which the speaker has to choose *may*/*can* features neutral or perfect/progressive aspect). In addition, there is another predictor CORPUS, which specifies the L1 of the speakers (let's say, native vs French vs Chinese). Several models are conceivable:

[3] For advanced readers who object to sequential model simplification or the use of *p*-values of the above kind, it should be noted that a single regression with planned contrasts also reveals that the alleged Germanic cluster is not homogeneous.

(2) Form ~ Corpus + Aspect + Negation
(3) Form ~ Corpus + Aspect + Negation + Aspect : Negation
(4) Form ~ Corpus + Aspect + Negation + Corpus : Aspect +
Corpus : Negation + Aspect : Negation +
Corpus : Aspect : Negation

The model in (2) already goes beyond much previous work because it embodies a multifactorial regression where several predictors, not just one, are studied simultaneously. However, it may still be lacking because it does not include interactions: one will not learn, say, whether the effect of Negation is the same for both levels of Aspect (or vice versa). For instance, if negated clauses in general have a higher probability of *may*, is that equally true for both aspects? The model in (3) answers this question by including the interaction Aspect : Negation and returning a regression coefficient and a *p*-value for this interaction. However, the model that should really be fit is that in (4) because here not only the two linguistic predictors interact with each other, but *all* predictors – including Corpus – do. These interactions, which contrast different speaker groups, are what most work in contrastive analysis and Contrastive Interlanguage Analysis is implicitly about, but which are too rarely tested explicitly:

- the interaction Corpus : Aspect tests whether the effect that Aspect has on Form (*can* vs *may*) is the same in the three L1 speaker groups
- the interaction Corpus : Negation tests whether the effect that Negation has on Form is the same in the three L1 speaker groups
- the interaction Corpus : Aspect : Negation tests whether the interaction of Aspect and Negation has the same effect on Form in the three L1 speaker groups.

Thus, only this type of regression will quantify whether any linguistic predictor does different things in NS vs NNS as well as in NNS_1 (e.g. French) vs NNS_2 (e.g. Chinese). On the basis of such a first regression model, one can then trim the model to the minimally adequate one by (i) weeding out independent variables that do not contribute enough predictive power to a model one by one and (ii) conflating levels of predictors that do not merit enough to be distinguished.

One recently developed approach (Gries and Deshors 2014) adds a new exploratory twist to regression modelling, namely a way to study in detail the following questions: (i) 'given the linguistic/contextual situation the NNS is in right now, what would a NS do?' and (ii) 'what determines the degree to which NNS do not make the choices a NS would have made?'. These are central questions raised a long time ago, but hardly studied accordingly: we need to be 'comparing/contrasting what non-native and native speakers of a language do *in a comparable* *situation*' (Pery-Woodley 1990: 143, quoted from Granger 1996: 43, my emphasis). Most previous LCR has adopted a very lax interpretation of 'comparable situation', namely that the NS and the NNS data were produced, e.g., in a 'similar essay-writing/speech situation'; some better ones do at least control for topics (see below). However, with a regression-analytic mindset, a much more realistic and revealing approach can be pursued: Gries and Deshors (2014) develop a protocol called MuPDAR (for *Multifactorial Prediction and Deviation Analysis with Regressions*); see also Gries and Adelman (2014) and Wulff and Gries (2015). First, one uses a multifactorial regression to see why NS make a particular choice. Second, if the fit of that regression is good, then that regression equation is applied to the NNS, which is the statistical way of asking (i) above, 'what would a NS do here?' Then, one determines where the NNS did not make the choice that a NS would have made and explores, with a second regression, which of the annotated factors explain when NNS do not behave as NS would. The authors show that French NNS often make non-NS choices with negated clauses as well as with *can* in perfective/progressive and *may* in neutral aspect, plus they have more difficulties with *may* with animate than with inanimate subjects. Similarly, Gries and Adelman (2014) show that NNS have difficulties in making NS-like subject realisation choices in Japanese precisely when the subject referents are not completely discourse-new or completely discourse-given but in the grey area in between. Such results are nearly impossible to obtain with mere over-/underuse counts and require methods with a fine-grained and contextualised view of the data.

3 Representative studies

3.1 Gries, St. Th. and Wulff, S. 2013. 'The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research', *International Journal of Corpus Linguistics* 18(3): 327–56.

Gries and Wulff (2013) is a study involving the above-mentioned regression approach. They study the genitive alternation in English as represented in (5) by comparing the constructional choices of native speakers of British English to those of Chinese and German learners of English.

(5)	a.	the squirrel's nut	*s*-genitive	possessor*'s* possessed
	b.	the nut of the squirrel	*of*-genitive	possessed *of* possessor

Previous studies of the genitive alternation in native speaker data have uncovered a large number of factors that co-determine which genitive speakers choose. As with many other alternations, these factors are from many different levels of analysis and include:

- morphosyntactic and semantic features: number, animacy and specificity of possessor and possessed as well as the semantic relationship between possessor and possessed (e.g. possession, attribution, participant/time and event)
- processing-related features: length and complexity of possessor and possessed, the previous choice made by a speaker, information status of possessor and possessed
- phonological features such as rhythmic alternation (the preference to have stressed and unstressed syllables alternate) or segment alternation (the preference for CV structures).

They retrieve approximately 3,000 examples of *of*- and *s*-genitives from the *ICE-GB* (for the NS) data and from the Chinese and German components of the *ICLE* (for the learner data), and annotate them for the above features. Given the fact that the genitive alternation is obviously a multifactorial phenomenon, they adopt a regression-analytic approach along the lines discussed in the previous section, and since the dependent variable – the choice of genitive – is binary (*of* vs *s*), they perform a logistic regression analysis, i.e. an analysis that determines (i) which of the annotated features and their combinations predict all speakers' genitive choices best, (ii) if/how the NNS differ from the NS in their genitive choices, and (iii) if/how the two NNS groups differ from each other. In order to determine the most parsimonious model of the genitive alternation, they undertake a manual model selection process that weeds out predictors that do not significantly help predict the genitive alternation while at the same time controlling for collinearity, i.e. the omnipresent and potentially dangerous phenomenon that predictors are too highly related to each other and thus do not allow the researcher to identify which predictor has what kind of effect. Their final model is then shown to explain the data very well based on a correlation coefficient and on the high accuracy with which the model can classify the speakers' genitive choices (>93%).

Several interesting findings emerge – once the results are visualised: multifactorial regressions are usually much easier to understand when represented graphically; visualisation should nearly always be provided for results. One is an effect across all three speaker groups such that segment alternation patterns are indeed weakly preferred. Another is an interaction of PossessorNumber and LengthDifference (between the possessor and the possessed), which is represented in Figure 8.3. The *x*-axis represents the predicted probability that a speaker would use the *s*-genitive, the *y*-axis represents the difference LengthPossessor minus LengthPossessed (in characters), and the small *s*'s and *p*'s represent the predicted probabilities of the *s*-genitives for **s**ingular and **p**lural possessors; the left panel highlights the curve for singular possessors (and

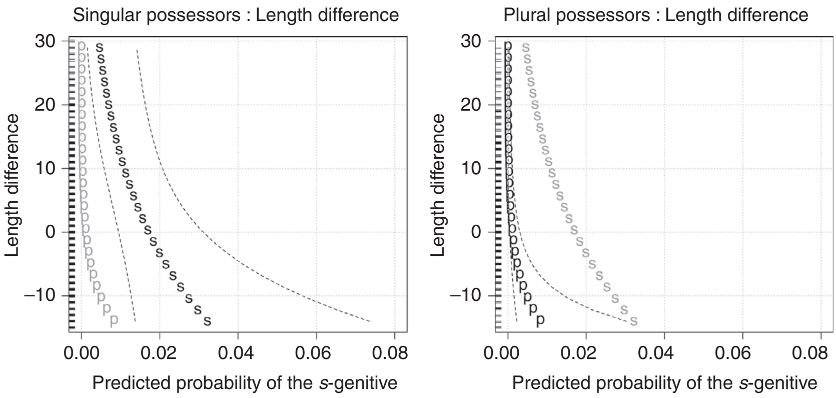

Figure 8.3 The effect of the interaction of PossessorNumber : LengthDifference on the choice of *s*-genitives in Gries and Wulff (2013)

plots the plurals in grey for the sake of easy comparison), the right panel focuses on the curve for plural possessors.

This is an interesting finding (that normal chi-square test analyses could not really make) because, while nearly every analysis of the genitive alternation has found that LengthDifference matters (in a general short-before-long tendency), this interaction shows that this is more pronounced with singular than with plural possessors. This is presumably because there is a general avoidance of *s*-genitives with plural possessors for articulatory reasons, which means that for LengthDifference to still have any impact at all, the length difference has to be quite large to still 'overpower' that avoidance.

Finally, one interaction that shows how the NNS differ from the NS involves the specificity of the possessed: all speakers prefer *s*-genitives with non-specific possesseds, but the German NNS do so only weakly while the Chinese NNS do so strongly. In sum, this study is instructive in how it showcases the power of regression-analytic approaches, viz. the ability to study multiple determinants of a phenomenon and their interactions as they affect the language of learners from different L1s.

3.2 Paquot, M. 2014. 'Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing', in Roberts, L., Vedder, I. and Hulstijn, J. (eds.), *EUROSLA Yearbook 14*. Amsterdam: Benjamins, pp. 240–61.

A study that is interesting for its use of monofactorial tests and frequency-based observations is Paquot (2014). She explores English 2/3/4-grams containing a lexical verb produced by French learners of English to determine how much of the learners' idiosyncratic use of n-grams is due to L1 transfer and what kinds of transfer effects can be found. She retrieves all 2/3/4-grams from the *ICLE* that occur 5+ times and computes their normalised frequencies per 100 words in order to control for the fact that corpus parts are not always equally large, which makes it impossible to conduct frequency comparisons based on raw frequencies. She then compares the mean French learners' n-gram frequencies with those of the other nine learner groups; laudably, she

- uses pairwise Wilcoxon rank sum tests for this rather than the more commonly but often incorrectly used alternatives of *t*-tests or ANOVAs; in this case, Wilcoxon tests are more appropriate because the data on which the tests are run will violate the assumptions that *t*-tests/ANOVAs make
- applies corrections for multiple testing rather than use the traditional significance level of 0.05 for all these tests; in this case, this is appropriate, or even required, because she studies one and the same data set with multiple (pairwise) significance tests.

On the basis of these tests, a variety of n-grams whose frequencies in the French data differ significantly from five other learner groups are identified: 228 n-grams (154 2-grams, 59 3-grams and 15 4-grams) showing intra-L1-group homogeneity and inter-L1-group heterogeneity. A more qualitative route is used to determine the degree of congruity between French and the French learners' English, whereby each n-gram's use in the learner data is compared to what the translation equivalent in French would be (also controlling for the effect that topic might have).

Paquot finds that the large majority of the 228 significantly overused n-grams identified as described above are referential expressions (>86%), but also that many of these are more likely due to the choice of essay topic by the learners because a large majority of these n-grams only appear in French learners' essays discussing one particular topic (the creation and future of Europe); these n-grams include both statistically overrepresented content words and function words marking tense. However, a variety of n-grams whose function was classified as 'discourse organisers' or 'stance markers' exhibited significant overuse by the French learners that could not be attributed to the essay topic. Of these, several are part of a longer chunk, which allows for a subsequent analysis of French translation equivalents: some overused n-grams turn out to result from their use in teaching materials, but many others can be shown to be due to several categories of transfer effects such as transfer of

- semantic properties (cf. *on the contrary* and *au contraire*)
- collocational/colligational properties (cf. *according to me* and *selon moi*)
- functions and discourse conventions (cf. *let us not forget that* and *n'oublions pas que*)
- L1 frequency (cf. *from this point of view* and *de ce point de vue*).

In sum, this study is instructive in its use of statistical tools (the use of non-parametric statistics given the non-normal data it studies) and the ways in which the statistical results – while not multifactorial per se – are carefully controlled for potentially epiphenomenal effects (topic choice) and include careful comparisons with frequency effects in L1s other than the targeted French learners.

4 Critical assessment and future directions

LCR has made an important contribution to S/FLA in that it showed how corpus frequencies are correlated with many central notions in S/FLA research and has thus raised awareness of the important role that all kinds of frequency information play in language acquisition and learning (cf. Ellis 2002). This development coincided with the general recognition in linguistics that corpus data – long shunned while generative linguistics was the dominant linguistic framework – have a lot to offer. However, while this positive development has drastically increased the number of corpus studies in LCR, the somewhat obvious fact that corpus methods are by definition distributional and quantitative has not yet led to an analogous increase in the statistical sophistication of LCR – neighbouring fields such as sociolinguistics, psycholinguistics or corpus linguistics in general exhibit an overall larger degree of sophistication. After having discussed a variety of core issues and representative studies in the previous sections, I will now discuss a range of problems that LCR studies often manifest (regardless of any insights they may still offer) as well as make a variety of suggestions as to how LCR needs to evolve to come to grips with the immensely complex nature of its questions and data. Specifically, I will first turn to a few problems that arise from how data and analyses are reported and make a few easy-to-implement suggestions to address these problems (Section 4.1), before I turn to problems in how analyses are conducted and what needs to be done instead or on top of current practices (Section 4.2). Crucially, the goal is *not* to dismantle any studies mentioned here, but to anticipate the common objections that the points to be made are neither as frequent nor as harmful as I claim they are.

4.1 Problems with how statistical analyses are reported

The simplest problems to fix pertain to how data and methods are characterised. For the former, many studies unfortunately only provide normalised frequencies of the phenomena studied. A case in point is Altenberg (2002: 44), who provides only normalised frequencies of *make*/*göra* in source texts and their translations; another is Connor et al. (2005), who report a variety of statistical results without any representation of the data thus analysed. This is problematic because it rules out follow-up analyses or replications because one can neither replicate the tests nor conduct additional ones without knowing the exact distributions of the data or, minimally, the sample sizes. As for the latter, often the statistical tests undertaken are not described comprehensively enough for readers to understand, let alone try and replicate/extend the analysis. For example, many LCR studies report using chi-square tests, but they do not report what kind of chi-square tests they used (goodness-of-fit or independence), whether they used a correction for continuity (yes or no), they do not report the chi-square values themselves (only *p*-value ranges such as <0.05 or <0.01), etc. Neff and Bunce (2012: 75–6) is a case in point, which also contains seemingly contradictory information: their Table 7 reports a *p*-value as '<0.05', but the very first sentence discussing that result states '(*p* between 0.10 and 0.05)'.

Obviously, the field would benefit much if authors were required to, wherever possible, provide at least summaries of observed data (as in Hundt and Vogel (2011), who usefully provide the exact observed frequencies in an appendix) as well as detailed descriptions of which statistics are used and how exactly they have been computed. Additionally, more statistical information should be provided to make analyses more revealing and useful; the most attainable improvements have long been discussed elsewhere:

- measures of dispersion should be provided for all measures of central tendencies
- confidence intervals should be provided for percentages
- meaningful effect sizes should be provided (see Norris and Ortega 2003: 746–7; Plonsky and Gass 2011 and, more generally but very usefully, Wilkinson and the Task Force on Statistical Inference 1999).

4.2 Problems with how statistical analyses are done

4.2.1 Analyses are done incorrectly/incompletely

The most straightforward way in which statistical analyses can be problematic is if the chosen statistical technique – whether it is in fact the appropriate technique or not – is applied incorrectly. A simple example of such problems is Laufer and Waldman (2011). The real chi-square value for their Table 1 is either 257.6 or 258.16, depending on whether one uses a continuity correction or not, respectively.[4] Second, probably because of this discrepancy, they also report the wrong effect size for their data: the real effect size is not 0.082 but – without the continuity correction, given the sample size – only 0.077, and all the results they report for their data in Table 8.2 are similarly fraught with mistakes; finally, they do not correct for the right number of post hoc tests.

Another kind of problem not uncommon in LCR is the application of multiple tests on the same data without an explicit correction that takes the number of tests performed under consideration. For example,

[4] *R* code for all computations is available from the author upon request.

Aijmer (2005) studies frequencies of possibility modals in three NNS corpora and one NS corpus and presumably performs twelve chi-square tests on one and the same data set, which is dangerous because it dramatically increases the risk of erroneously claiming significant findings. This is because in a single test, the probability of erroneously assuming that a finding is significant is typically 5%, but if one does twelve tests on one data set, the probability of erroneously assuming that at least one finding is significant increases to $1\text{-}0.95^{12}\approx0.46$, i.e. 46%. To felicitously perform twelve tests, one needs to correct one's significance level (recall the discussion of Paquot (2014) above) from 0.05 down to $1\text{-}0.95^{1/12}\approx0.0043$ (according to one frequently used correction).

Another important problem of some existing work is the convention of grouping numeric data in several different categories. For example, Durrant and Schmitt (2009) explore percentages of collocations as a function of the association measures *t* and *MI*. However, instead of correlating the numeric percentages of collocations with the numeric values of the association measures, they grouped the association measures into seven/eight groups respectively. This practice of grouping – especially its extreme case, dichotomisation – has long been known to result in potentially huge losses of power and precision (cf. Cohen 1983; Baayen 2010) and should be avoided unless it can be shown that grouping is not harmful and in fact even necessary.

A final important point of critique mentioned earlier is that too much of LCR is still *mono*factorial in nature. That is, it explores a particular phenomenon from a perspective which involves maximally one potential cause–effect relationship, as if L1 is the only reason why particular words are over-/underused. Nothing in linguistics is truly monocausal, so this perspective is impoverished because it implicitly assumes that (i) the one predictor a study included in its exploration of a particular response is among the most important ones and (ii) the effect of that predictor is not unduly amplified/downplayed by other factors not included in the design, or that all other factors' effects cancel each other out in some way. This implicit assumption is usually unwarranted – I have yet to see a multifactorial LCR study without at least one significant interaction – which means that LCR should embrace methods that can handle *multi*factoriality more.

4.2.2 Desiderata involving state-of-the-art corpus-linguistic statistics

Another way in which LCR can evolve is by learning from developments in corpus linguistics or neighbouring fields. One simple example involves the question of which association measures one uses to study co-occurrence. Measures such as *MI* and *t* are still widely used – ideally together, given that they return very different results (Durrant and Schmitt 2009 is exemplary in this regard) – but there have been plenty of studies using other measures, which in turn have yielded converging evidence from experimental work. For instance, most studies exploring the association of words and constructions have used-log $p_{\text{Fisher-Yates exact}}$ as an association score, and Gries et al. (2005, 2010) as well as Ellis and Ferreira-Junior (2009a) have found that this is a good measure in terms of predicting experimental results with NS (sentence-completion and self-paced reading times) and NNS (learner uptake). As mentioned above, Ellis and Ferreira-Junior (2009a) test one of the very few directional association measures, Δ*P*. While Δ*P* in their results is highly correlated with $p_{\text{Fisher-Yates exact}}$ but a slightly less good predictor of learner uptake, the directionality of the measure makes it a very appealing tool on theoretical grounds.

LCR should also evolve by exploring the only existing association measure that includes type frequencies. That is, virtually all measures are based on (co-)occurrence frequencies of tokens, but do not take into consideration how many different types the (typically) words in question co-occur with. Just as in corpus linguistics, this is probably due to the ease with which token-based measures can be computed, but since we know that type frequencies of co-occurrence are important – to assess productivity and as an indication that a slot in a construction is (not) variable – there is no real reason not to explore a measure such as lexical gravity *G* (see Daudaravičius and Marcinkevičienė 2004; Gries 2010b). The computation of *G* includes the number of different words with which a word occurs: all other things being equal, the association of two words *a* and *b* is stronger than the association of two words *x* and *y* when *a* occurs with more different types than *y*.

Related to this question is the increasingly popular topic of n-grams. Currently, these are usually studied on the basis of some *n* and (ranked) frequency lists or *MI* scores. However, the choice of *n* often seems quite arbitrary – counterproductive in fact, since one needs *n*=2 to find *according to*, *n*=3 for *in spite of*, *n*=4 for *on the other hand*, *n*=5 for *as a matter of fact*, and *n*=6 for *the fact of the matter is*. It would therefore be more useful to adopt bottom-up strategies that vary *n* to identify the best candidates for n-grams. Also, with regard to the statistics, *MI* scores are usually computed on the basis of complete independence of the words in the n-gram. This way of computation is easiest computationally, but at the same time it is a shortcut yielding results that are unintuitive and volatile compared to other approaches. For instance, if one computes *MI* for the 3-gram *in spite of* in *ICLE* (Version 1) on the basis of complete independence – i.e. assuming that the occurrences of *in*, *spite*, and *of* are not correlated at all – the resulting *MI* score is more than double the *MI* of *in spite* and *of*, and nearly double the *MI* score of *in* and *spite of*. Plus, the above examples do not consider the possibility of discontinuous n-grams and, as one might expect, *spite* is extremely highly predictive of *in spite of*. Researchers interested in n-grams should consider measures such as Kita's cost criterion (Kita et al. 1994), extensions of lexical gravity *G* (Gries and Mukherjee 2010), O'Donnell's (2011) adjusted frequency list, and a pertinent special issue of *Language Resources and Evaluation* (Rayson et al. 2010).

Finally, LCR has not been concerned much with dispersion, i.e. the fact that the frequencies of two words *x* and *y* may be identical, but that *x* may only occur in one small part of a corpus (i.e. be underdispersed) while *y* occurs everywhere (i.e. is evenly dispersed). Imagine the string of letters below to be a corpus consisting of five parts and altogether fifty word tokens (each letter represents a word, | indicates boundaries of corpus parts):

> y b c d e f g h i y | y t e f q v b g u t | e f t y c q w e d e | y r t h j o f e w y | x x x x y x x x x y

In this case, both *x* and *y* occur eight times, but *x* is only attested in one of the five corpus parts, whereas *y* is attested in every corpus part. While frequencies cannot distinguish between the two words, dispersion measures can. For instance, the measure *DP* (for *deviation of proportions*, a measure ranging between 1 and 0 for clumpily and evenly distributed words respectively) for *x* and *y* are 0.8 and 0.15, reflecting the former's clumpiness and the latter's fairly even distribution in the corpus. Gries (2008b) shows that neglecting dispersion can lead to spurious generalisations, such as when high frequencies of linguistic expressions mask the fact that they may be distributed very clumpily and, thus, be highly specialised (in terms of mode, register, topic, or speaker population).

In sum, nearly every kind of corpus statistic that LCR depends on could benefit from being more open to newer developments; exploration of these techniques can only increase the robustness of the field's findings.

4.2.3 Multivariate approaches

Apart from the type of hypothesis-testing examples discussed above, LCR can also benefit from more application of exploratory tools such as hierarchical cluster analyses and others. As far as I know, these tools have not been widely applied in LCR, so I will mention three examples here.

The first of these involves a very simple cluster-analytic application to a two-dimensional representation of relative frequencies. As an example, we will use Hasselgård and Johansson's (2011) Table 3 (shown in Table 8.3), which shows the patterns of *quite* across corpora (relative frequencies per 100,000 words).

One interesting alternative approach to the stacked barplot used by Hasselgård and Johansson would be to try to visualise the structure in Table 8.3 with a hierarchical cluster analysis, a technique that computes user-defined (dis)similarity measures between, in this case, the corpora, and then presents these (dis)similarities in a so-called dendrogram. In such dendrograms, elements (here, corpora) that are clustered together early (i.e. in the bottom of the graph) are highly similar to each other

Table 8.3. Hasselgård and Johansson's (2011) Table 3

	+ADJ	+ADV	+PRED	+INDEF NP	+PP	+DEF NP	Other
ICLE-NO	24.7	4.3	1.9	10.5	1.0	1.0	0.5
ICLE-GE	38.5	6.8	2.1	12.3	0.4	2.1	0.0
ICLE-FR	25.4	3.9	1.0	5.9	0.0	0.5	1.5
ICLE-SP	25.2	2.5	1.0	0.0	0.5	1.0	1.5
LOCNESS	12.6	3.4	0.9	2.8	0.6	0.3	0.0

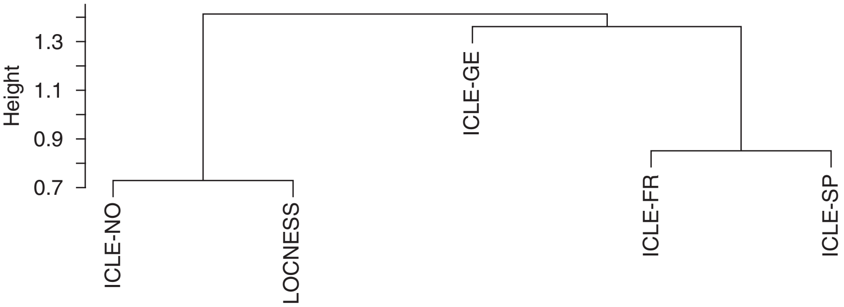

Figure 8.4 Dendrogram of Hasselgård and Johansson's (2011) Table 3

and dissimilar to the others. If one performs a particular kind of cluster analysis (using 'centred Pearson (1-*r*)' as a similarity measure and Ward's amalgamation rule; see Gries 2013b for discussion), one obtains the dendrogram in Figure 8.4, which in a very recognisable way groups the two Romance languages together, the Norwegian NNS with the NS, while the German NNS are fairly dissimilar to both these groups. (Other kinds of analysis might yield different results, a phenomenon not untypical of exploratory methods.)

A more interesting application of cluster analysis bridges the gap between the study of learner varieties and indigenised L2 varieties. Szmrecsanyi and Kortmann (2011) study, among other things, English corpus data from *ICLE*, *ICE* and, as a NS control, the *British National Corpus* with an eye to quantifying and comparing the analyticity and syntheticity of all L1s. In a first step, they plot analyticity against syntheticity of all L1s and show, with *t*-tests, that learner varieties are significantly more analytic than indigenised varieties. Their next step, however, involves a cluster analysis, whose dendrogram 'confirms that on the whole, *ICLE* varieties and *ICE* varieties indeed split up quite nicely into two different clusters' (p. 175).

As a final example of the exploratory use of multivariate tools, consider Jarvis (2011). This paper discusses the contrast between different kinds of classification methods as well as aspects of how they are used and tweaked, and then proceeds to test how well twenty supervised classifiers can identify the L1s of authors of 2,033 texts in L2 English on the basis of the frequencies of 722 n-grams. Interestingly, he finds that, while chance accuracy would be as low as 9.4 per cent, the best classifiers (Linear Discriminant Analysis and Sequential Minimal Optimisation) achieve accuracies more than 5.6 times as good. Even though these approaches all use decontextualised frequencies – unlike the regression-analytic approaches discussed before – their goals of predicting an L1 or distinguishing/comparing different L1s are, of course, very different from the goal of comparing NS and NNS with regard to particular words or patterns. Nevertheless, the overall good accuracy points to strong frequency patterns that would be impossible to recognise without multivariate methods.

4.2.4 Concluding remarks

Recently, some LCR has begun to catch up with experimental research in SLA and the frequent use of multifactorial techniques in that field (cf. Plonsky and Gass 2011) and in corpus linguistics in general. However, it has also become clear that much of LCR still has to evolve in its use of statistics. In addition to the most central issues discussed above – a move towards regression-analytic methods with multiple predictors and interactions – the following are what I believe are the most pressing issues and obvious next steps.

First, the statistical techniques used need to be not just appropriate to the general task at hand but also appropriate given the specific data set, and researchers need to become more familiar with a wider range of techniques, their assumptions, limitations and implementation. The following kinds of approaches are particularly noteworthy:

- One is the approach of *robust statistics*, a set of methods useful for handling the types of data that SLA researchers often deal with, namely data that violate distributional assumptions such as normality, variance homogeneity and others (see Larson-Hall and Herrington 2009; Wilcox 2012).
- Other important areas are *exact statistics* and *bootstrapping/resampling* approaches, which can also help with data that violate standard assumptions of more traditional tools as well as with smaller sample sizes.
- Finally, and maybe most importantly, very little work in LCR has concerned itself with the fact that the data points in studies are often not independent because, for instance, multiple data points are contributed by individual speakers/writers, multiple data points may be nested into individual files, which may be nested into different corpus sources, which may be nested into the sets of files representing different L1s of speakers, etc. (Durrant and Schmitt (2009) is one laudable exception, which takes at least individual texts into consideration).

> These types of issue render all sorts of traditional statistics – including all of the above suggestions for regressions – risky and require currently underutilised methods such as *mixed-effects models* or *multi-level models*, which can handle such repeated measurements and nested data structure (see Rietveld et al. 2004; Baayen 2008: Chapter 7; Gries 2015 for some discussion).

A second important development is the use of multimodel inferencing. Rather than doing regression modelling as discussed above, i.e. developing an initial model and then trimming it down until only significant predictors are included, the approach chosen in most regression modelling in linguistics, *multimodel inferencing* involves fitting a potentially large number of different models on the same data and aggregating the results such that results from better-fitting models contribute more to the final results. Given the large numbers of variables that linguistic data often involve and their high degrees of interrelatedness (a.k.a. collinearity), this approach is an extremely promising development (see Burnham and Anderson (2002) for a general introduction, and Kuperman and Bresnan (2012) for an application in linguistics).

Thirdly, all of the above also implies that learner corpus researchers need to become more familiar with *data analysis software*. The fastest-growing statistical software environment in linguistics is the open source programming language *R* (*R* Core Team 2014), which provides all the functionality corpus linguists will ever need (see Gries 2009, 2013b).

LCR is making its first steps in all these directions and, hopefully, developments and tools like these will ultimately put the findings within our discipline on a more solid foundation.

Key readings

Larson-Hall, J. and Herrington, R. 2009. 'Improving data analysis in second language acquisition by utilizing modern developments in applied statistics', *Applied Linguistics* 31(3): 368–90.

This paper is an overview of how statistical methods in S/FLA research can and should be improved to take into consideration the distributional particularities of language with (in particular) so-called robust statistics.

Wulff, S., Ellis, N. C., Römer, U., Bardovi-Harlig, K. and Leblanc, C. J. 2009. 'The acquisition of tense–aspect: Converging evidence from corpora and telicity ratings', *The Modern Language Journal* 93(3): 354–69.

This paper is a detailed study of the effects of frequency, (Zipfian) frequency distributions, contingency/association of form and function, and prototypicality on the acquisition of tense–aspect patterning in English using both corpus data (multiple distinctive collexeme analysis) and experimentally obtained telicity ratings.

Ellis, N. C. 2012a. ‘Formulaic language and second language acquisition: Zipf and the phrasal teddy bear’, *Annual Review of Applied Linguistics* 32: 17–44.

This is a comprehensive survey article on the topic of language learning with a focus on formulaic sequences, their statistical definition, their processing and its sensitivity to statistical co-occurrence information, and individual differences in language learning.

Gries, St. Th. 2013b. *Statistics for Linguistics Using R,* 2nd edn. Berlin: De Gruyter Mouton.

This book is an accessible introduction to statistics for linguists that covers nearly all statistical techniques referred to in this overview, using corpus and experimental data from a wide range of linguistic fields.