2 Corroborating evidence: Data and methodology

2.1 ‘What counts as evidence in Linguistics’

As Penke and Rosenbach point out, ‘nowadays most linguists will probably agree that linguistics is indeed an empirical science’ (2004: 480). However, while the importance of empirical evidence is generally acknowledged by most researchers, the following quotations from Sampson and Chomsky show that there is no agreement among linguists as to the type of data that is to be analysed empirically:

We do not need to use intuition in justifying our grammars, and as scientists, we must not use intuition in this way. (Sampson 2001: 135)

You don’t take a corpus, you ask questions … You can take as many texts as you like, you can take tape recordings, but you’ll never get the answer. (Chomsky in Aarts 2000: 5–6)

If both Sampson’s position on introspection and Chomsky’s views on corpora were correct, there would obviously be no valid data base left for linguists to investigate. Fortunately, however, Sampson and Chomsky are only extreme proponents of their respective schools of linguistics. Nevertheless, when investigating a particular syntactic phenomenon, many linguists still only draw on either corpus or introspection data (though there seems to be an increasing number of exceptions such as Gries, Hampe and Schönefeld 2005, the collected volume by Kepser and Reis 2005 or the special issue on corpus and experimental techniques of Corpus Linguistics and Linguistic Theory 5.1 – in particular see Gilquin and Gries 2009). In the literature this preference for either of the two types of data is often attributed to different epistemological approaches (e.g. Lemnitzer and Zinsmeister 2006: 14–32).

Linguists like Sampson are said to be influenced by empiricism, a philosophical school which advocates the prime importance of experience and favours an inductive scientific approach. Followers of Chomsky, on the other hand, are said to be influenced by rationalism, which emphasizes rational hypothesizing and is characterized by a deductive approach. While the

preference for a particular type of data by an individual school might be explained by its philosophical background, I consider this fact immaterial for the present study. Instead, I claim that in order to qualify as scientific evidence it must only be ensured that a specific linguistic data type meets the major constraints normally imposed on empirical research, i.e. that data

(1) must be objective, i.e. interpersonally observable (cf. Sampson 2001: 124),

(2) allow for valid and reliable measurements (cf. Kline 1993).

As I will show, while the criticism of a specific type of data is not always couched in these terms, it is in fact the objectivity, validity and reliability of introspection and corpus data that is questioned by proponents of the alternative schools of linguistics.

Interestingly, advocates of both introspection and corpus data usually follow the same line of argument: the weaknesses of introspection/corpus data are x, y and z. Yet x, y and z are exactly the advantages of the competing methodology. That is why you should never use the former, but only stick to the latter type of data.

The argument for introspection data usually runs like this: corpora only exhibit a speaker’s ‘performance’, which is influenced by ‘memory limitations, distractions, shifts of attention and interest and errors (random and characteristic)’ (Chomsky 1965: 3). Thus a speaker’s performance, i.e. corpus data, is only an indirect and partly flawed reflection of his competence. As a result, corpus data are haunted by the ‘performance’ problem: just because a sentence appears in a corpus doesn’t mean that it is grammatical. In addition to this, it is generally accepted that linguistic competence enables a speaker to create an infinite number of sentences. Yet, how should a finite corpus contain all the examples relevant for the analysis of a particular problem (cf. McEnery and Wilson 1996: 4–10)? This obviously leads to the well-known ‘negative data’ problem: just because a construction does not surface in a corpus it does not follow that it is ungrammatical. Therefore, the intuition of a native speaker drawing on his competence has to be preferred over the examination of corpus data.

The argument for corpus data, on the other hand, usually runs like this: the sentences used for introspective judgements are ‘unnatural’, invented data which lack a communicative context. Judgements on these sentences are then collected in an unsystematic, unscientific way: most of the time the linguist will only rely on his or her own intuitions. Thus linguists who use introspective data ‘produce theory and data at the same time’ (Labov 1972: 190). If anyone then casts doubts on their judgements, these linguists resort to the claim that judgements might vary but that in their idiolect the sentence is in fact grammatical/ungrammatical (Sampson 2001: 137). Since introspection thus yields data which cannot be refuted, it must be considered ‘un-scientific’ (e.g. Sampson 2001: 124). Finally, the
intra- and inter-speaker stability of introspection data is questioned (for an overview, see Cowart 1997: 4f.) and sometimes speakers even say things that they believe sincerely that they would never produce (Sampson 2001: 136). Due to these problems one should always stick to authentic data provided by corpora.

A closer look at the above arguments reveals that they imply that corpus and introspection data violate the objectivity, validity and reliability constraints on empirical research.

The main criticism of corpus data obviously concerns their validity. An empirical measurement technique can only be ‘said to be valid if it measures what it claims to measure’ (Kline 1993: 15). Now the opponents of corpus data claim that the object of study in linguistics is the linguistic competence of a speaker. Since corpus data are ‘fl awed’ by performance factors, they do not constitute a valid means of investigating linguistic competence. Before evaluating this argument, it needs to be pointed out that not only corpus data are measurements of linguistic performance. As Schütze (1996: 6, 14–15) has argued, introspective judgements are also subject to performance factors. Thus, introspection offers ‘a different access path from language use to competence, [but is itself] just another sort of performance’ (Schütze 1996: 6). Accordingly, if performance was in fact only a flawed mirror of competence, neither corpus nor introspective data could be considered valid.

Yet, as Leech noted, ‘the putative gulf between competence and performance has been overemphasised’ (1992: 108), since the latter clearly is the product of the former. Following Occam’s razor, it would therefore be much more reasonable to assume that, under normal circumstances, performance should actually be a rather good mirror of a speaker’s competence. Second, the guiding principle of modern corpus design is representativeness: nowadays, corpora are designed as statistically representative samples of populations (cf. McEnery and Wilson 1996: 63–6). Although a corpus can never contain all sentences of a language, it will at least be a carefully constructed miniature model. Thus, both corpus data and introspection can be considered valid ‘access paths’ to competence. Nevertheless, the two are obviously not measuring exactly the same phenomenon: while the strength of the former lies in constituting positive data which can be analysed for frequency and context effects, the latter allows the investigation of the grammaticality of negative data. It is this complementary nature of the validity of these different types of data that calls for them to be treated as corroborating evidence. Note that complementing introspection data with corpus data also has the advantage of minimizing the one problem affecting the validity of the former: since ‘informant judgements don’t always agree with the informant’s own linguistic behaviour’ (Cowart 1997:5; cf. e.g. Labov 1975), corpus data can be used to check the validity of subject’s grammaticality judgements.

As can be seen above, one of the main criticisms of introspection data concerns their reliability, i.e. their intra- and inter-subject consistency (cf. Kline
While informal methods of introspection data elicitation might indeed fail to be reliable, recently several studies have provided statistical proof that judgements elicited via carefully constructed experiments are in fact intra- and inter-subject consistent (cf. Bard, Robertson and Sorace 1996; Cowart 1997: 12–28; Keller 2000). The reliability of corpus data – at least with respect to test–retest reliability (Kline 1993: 5) – on the other hand, has never been questioned: since most corpora, with the exception of monitor corpora, are finite samples, repeated measurements of a phenomenon automatically yield reliable results. Yet, the question of intra- and inter-subject consistency also applies to corpus data: however balanced and representative a corpus might be, there is no guarantee that all informants have contributed the same number of tokens of a particular linguistic phenomenon. In fact, apart from highly frequent phenomena (such as NPs) it is very likely that some informants might not have produced any relevant data while others have contributed two or more tokens to a corpus. As long as it can be entertained that linguistic theory should be ‘concerned primarily with an ideal speaker–listener, in a completely homogenous speech community’ (Chomsky 1965: 3f.), this obviously does not constitute a problem: if linguistic knowledge were in fact distributed homogeneously across a speech community, then it would not matter how many tokens an individual contributed to a corpus. However, while the mutual intelligibility of speakers of a single dialect does indeed imply a certain degree of homogeneity within a speech community, ever since Labov’s groundbreaking New York study (1966) it has become apparent that all speech communities display variation on all linguistic levels (for an overview see e.g. Hudson 1996; Milroy and Gordon 2003; Trudgill 1995). Consequently, since not all informants contribute an equal amount of data on a particular phenomenon, it cannot automatically be assumed that overall corpus results reflect the linguistic behaviour of all informants – let alone of all speakers of the speech community in question.

This problem is even more acute for second-language-learner corpora: it is well known that language learners differ greatly with respect to the rate at which and the degree to which they attain a second language (cf. e.g. R. Ellis 1995: 99–126, 1998: 73–78; Granger 1998: 5–6; Leech 1998: xviii). Therefore if a more advanced learner adds little or no data to a corpus while a less proficient one contributes a great number of tokens, the resulting skewed data set clearly can not be considered representative of all second language learners. Now one way to preclude such a confounding effect is of course the choice of a representative set of informants which one expects to be linguistically as homogeneous as possible. In addition to this, however, it is also necessary to test statistically whether the data set in question is in fact homogeneous. This requires statistically analysing corpus data, ensuring that the resulting statistical models have a good fit for the data and minimizing the undue influence of individual informants (in the case of preposition-stranding and pied-piping by e.g. subjecting corpus data to a Generalized Linear Model
(GLM) analysis, checking the model fit and the error structure of the final GLM model and establishing the cross-validation accuracy of the model; see Maindonald and Braun 2003: 209f. and section 2.2 below for details). Finally, another way to ensure the intra- and inter-subject consistency of results based on corpus data is to corroborate them in an introspection experiment in which all subjects provide equal information on all tested conditions. In such experiments the homogeneity of the subjects’ judgements – and via statistical inference the homogeneity of the speech community – can thus explicitly be tested.

The fact that both types of data can in fact be tested in a reliable way is also important for the third criterion of empirical research mentioned above: objectivity. Since results of empirical studies should be independent of a particular researcher, it is obviously essential that repeated measurements of a data source by different researchers yield similar results. Objectivity in turn guarantees that results obey the meta-principle of empirical research: the fact that conclusions drawn from data analysis must be refutable/falsifiable. As Sampson correctly emphasizes: ‘All that matters is that any feature of the theory which is doubted can be confirmed or refuted on empirical grounds’ (2001: 137). As mentioned above, this is indeed a problem if in the face of competing judgements, a linguist argues that in his/her idiolect a construction is (un-)grammatical. While this may or may not be the case, it is a virtually unfalsifiable claim. Again it must be stressed that nowadays an increasing number of linguists refrain from only relying on their own introspection. Instead introspection data is collected obeying the standards of psychological experiments: via the selection of a representative sample of subjects, careful design of experimental materials, randomization of the order of stimuli, the use of fillers/distractors and the employment of statistical analysis (cf. Bard, Robertson and Sorace 1996; Cowart 1997; Schütze 1996). Clearly, introspection data collected in such a way can be considered both objective and falsifiable. Since corpora easily allow hypotheses to be tested by different researchers, the objectivity and falsifiability of this data source is also uncontroversial.

Thus, both corpus and introspection qualify as good data sources for empirical research. Both are valid, reliable and objective data which allow for the generation of falsifiable hypotheses. Their individual strengths derive from the fact that they are measuring slightly different phenomena: corpora yield good data for the statistical analysis of positive data with respect to frequency and context phenomena. Single tokens or the absence of a construction in a corpus are, of course, interesting and important findings. Any hypothesis based on such data, however, needs to be corroborated by experimental introspection data in order to ensure empirical validity. Introspection, on the other hand, allows the investigation of negative data and, given the right experimental method, an elicitation of subtle grammaticality judgements. In addition to this, it should be noted that due to the
creative aspect of language – a classic argument for this type of data – introspection experiments can also profit from corpus studies when it comes to test design: if the set of possible sentences is infinite, how should a linguist be expected to come up with all relevant examples of a construction? Or as Jan Aarts puts it: ‘Only linguists who use corpus data themselves will know that a corpus always yields a much greater variety of constructions than one can either find in the literature or think up oneself’ (1991: 46). Thus, a corpus study can also yield data for further experimental studies which otherwise would not have been detected. For while linguistic competence might be homogeneous in a speech community – a claim which requires much further empirical research – creativity clearly varies greatly from one individual to another.) As a result, I disagree with Sampson and Chomsky, and instead completely agree with McEnery and Wilson, who ask

[w]hy move from one extreme of only natural data to another of only artificial data? Both have known weaknesses. Why not use a combination of both, and rely on the strengths of each to the exclusion of their weaknesses? A corpus and an introspection-based approach to linguistics are not mutually exclusive. In a very real sense they can be gainfully viewed as being complementary. (1996: 16)

Elsewhere (Hoffmann 2006), I have called the use of multiple sources of different types of data ‘corroborating evidence’ (an approach independently advocated as ‘converging evidence’ by Gries, Hampe and Schönefeld 2005). The basic idea behind this term is that, just like in a criminal investigation, linguists have to amass enough corroborating evidence to convince a jury of their peers. Accordingly, for the present study I will draw on introspection and corpus data to come up with a case on preposition placement in British and Kenyan English that will hopefully convince my peers.

In a criminal case, however, it is also important that the adduced evidence is gathered and interpreted in a forensically sound way. For the present study I take this to mean that, whenever possible, statistically analysed quantitative data are preferred over ‘hearsay’, i.e. qualitative data. The latter might no doubt provide interesting insights, but I argue that the former helps to present a much stronger case. In the following chapters I will elaborate on my various sources of evidence as well as the forensic tools used, i.e. the statistical analyses with which the data were interpreted.

2.2 Exhibit A: Corpus data

When contrastively investigating two or more varieties it is crucial to ensure that the data which are to be analysed have been sampled in a comparable way. Only then is it possible to claim that differences between samples actually reflect differences between the varieties in question (e.g. Leech 2007: 141–2).
For corpus-based contrastive studies on varieties of English it is therefore of vital importance to employ ‘matching’ or ‘comparable corpora’, i.e. a set of two or more corpora whose design differs, as far as possible, in terms of only one parameter; the temporal or regional provenance of the textual universe from which the corpus is sampled’ (Leech 2007: 141–2).

For the synchronous varieties of English around the world the International Corpus of English (ICE) project is such an attempt to compile a set of comparable corpora. The project was initiated by Sidney Greenbaum in 1988, and in addition to British English, ICE corpora have, for example, been compiled for varieties such as East African (Kenyan and Tanzanian English), Hong Kong, Indian, Irish, Jamaican, New Zealand, Philippines or Singaporean English. Based on identical design principles, all these corpora employ a common annotation scheme ‘in order to ensure maximum comparability’ (Nelson, Wallis and Aarts 2002: 3). Furthermore, all corpora will not only be tagged for part-of-speech but will also be parsed for syntactic structure. At present, however, only the British English component, the ICE-GB corpus, is available fully tagged and parsed.

Since all other ICE corpora, including the East African one (ICE-EA), are to be modelled on the British English component, I will now first give an overview of the ICE-GB before detailing the specific properties of the ICE-EA corpus.

2.2.1.1 The International Corpus of English (ICE) corpora

2.2.1.1.1 ICE-GB

ICE-GB was ‘compiled and grammatically analysed at the Survey of English Usage, University College of London, between 1990 and 1993’ (Nelson, Wallis and Aarts 2002: 3), with all texts dating from 1990 to 1993 inclusive. It is a one-million-word corpus, consisting of spoken (about 637,000 words) as well as written (about 423,000 words) material. Every text in the corpus has been assigned a unique text code which identifies its text category (e.g. S1A-001 to S1A-090 are face-to-face conversations; see Nelson, Wallis and Aarts 2002: 309–31). Concerning their stylistic level, both spoken and written texts in ICE-GB range from less formal (private face-to-face conversations and social letters) to rather formal (public legal cross-examinations and printed academic writings). Furthermore, some categories, such as dialogues and student examination scripts, are obviously produced more spontaneously than printed texts, which allow more planning time and have undergone an extensive editorial process (Nelson, Wallis and Aarts. 2002: 5ff.).

All the speakers and writers in the corpus are adults (age 18 or over), and, with few exceptions, were born in England, Scotland or Wales. All informants have completed secondary-level schooling, with many having ‘received tertiary education as well’ (Nelson, Wallis and Aarts 2002: 5). ICE-GB is
thus intended to be a representative sample of educated British English in the 1990s (Nelson, Wallis and Aarts 2002: 4f.). As a result, although the corpus contains texts from all levels of formality (from private conversations to academic writings), all data from the ICE-GB corpus undeniably constitute samples of the most educated end of the British English sociolect. If, as is sometimes claimed, the pied-piping preference is indeed to a great extent dependent on education (McDaniel, McKee and Bernstein 1998: 309), then, regardless of the level of formality, the current study can be said to investigate the sociolect with the greatest pied-piping tendency.

ICE-GB is fully tagged for part-of-speech and parsed for syntactic structure. Thus each word in the corpus has been assigned a word-class label (‘tag’) and in many cases also additional features, providing information such as (e.g.) a verb’s transitivity. Furthermore, its syntactic function has also been identified. (For an overview of the various tags, features and syntactic functions encoded in ICE-GB see Nelson, Wallis and Aarts. 2002: 22–68.)

(2.1) I very much enjoyed the work that I was involved in<PS, PREP>

The word *in* in (2.1), for example, carries the tag ‘PREP’, which means that it has been classified as a preposition (Nelson, Wallis and Aarts 2002: 34–5). Furthermore, if a preposition is not followed by a complement it is assigned the syntactic function ‘PS (stranded preposition)’ (cf. *in* in (2.1); see also Nelson, Wallis and Aarts 2002: 53). With the help of the ICE-GB’s retrieval software (a program called ICECUP) it was thus possible to extract all relevant stranded tokens for the present study via the stranded preposition ‘PS’-tag. Note however that prepositions which are part of incomplete utterances (2.2) or unintelligible fragments (2.3) also do not exhibit an overt complement, and accordingly these are also parsed as ‘stranded prepositions’ in the corpus. The output of the ICECUP query thus still had to be inspected manually and tokens such as (2.2) and (2.3) were excluded from further analysis.

(2.2) It’s**[like]** <ICE-GB:S1A-053 #257:1:C>
(2.3) <unclear> three-or-four-words </unclear> soft **[in]** <unclear>
<unclear-words> </unclear> <ICE-GB:S1A-018 #57:1:B>

---

2 For detailed information about the tagging and parsing procedure, see Nelson, Wallis and Aarts (2002: 13–17).
3 For expository purposes the tags of the other words in (2.1) have been omitted.
4 All information from the ICE-GB corpus contained within angle brackets < > represents part of the corpus’s structural markup, i.e. additional information added by the compilers. In (2.3), for example, the string ‘<unclear> three-or-four-words </unclear>’ means that there were three or four words in the text which could not be identified by the compilers. A complete list of all structural markup symbols employed in ICE-GB can be found in Nelson, Wallis and Aarts (2002: 333).
2.2 Exhibit A: Corpus data

The second advantage of ICECUP which facilitated the extraction of the relevant data from ICE-GB was that ICECUP has a so-called ‘Fuzzy Tree Fragment’ (FTF) option which allows the user to search the corpus for abstract syntactic structures (cf. Nelson, Wallis and Aarts 2002: 116ff.). Instead of having to limit the search for pied-piped constructions to, for example, specific preposition+\textit{which}-word constructions (e.g. ‘in which’, ‘to which’, ‘of which’, etc.), it was therefore possible to design FTFs which for a given \textit{which}-item found all instances in the corpus where it was governed by a preposition (i.e. ‘\textit{P} + \textit{which}’, ‘\textit{P} + \textit{who}’, etc.).

In light of the above features of ICE-GB, it should have become apparent that the corpus constituted the perfect data source for the present study: it includes texts from all levels of formality and is the model based on which several comparable corpora have been or are currently being compiled. In addition to this, the accompanying software ICECUP allowed the retrieval of all relevant stranded and pied-piped data from the corpus. Unfortunately, as I will show next, the source of the Kenyan corpus data, the East African component of the ICE project (ICE-EA), does not possess several of these advantages.

2.2.1.2 ICE-EA

The Kenyan corpus data were retrieved from the ICE-EA corpus, which consists of spoken and written texts from Kenya and Tanzania, all dating from between 1990 and 1996 (cf. Hudson-Ettle and Schmied 1999: 5). In line with the general ICE sampling scheme, all informants are ‘adults (over 18) who have received formal education up to at least secondary level schooling through the medium of English or have a public status that makes their inclusion appropriate’ (Hudson-Ettle and Schmied 1999: 5). Thus concerning the criteria of the sampled time period as well as the included informants the ICE-EA corpus can be said to match ICE-GB.

In total, the Kenyan subcorpus of the ICE-EA corpus comprises 791,695 words (289,625 words in the spoken component + 100,207 words in the written-as-spoken section and 401,863 words from written texts; see Hudson-Ettle and Schmied 1999: 53–63). The ICE-EA sampling scheme differs somewhat from that of ICE-GB. The reason for this is that the specific situation in East Africa required several changes in order to guarantee the representativeness of the data (as being typical of Kenyan and Tanzanian English; see Schmied 1996). In particular, it turned out to be impossible to sample the text types described below, all of which feature in the ICE-GB corpus (cf. Hudson-Ettle and Schmied 1999: 4).

While the spoken part of ICE-GB contains ‘phonecalls’, ‘business transactions’ and ‘unscripted monologues’, none of these could be sampled for either Kenyan or Tanzanian English. While it was also impossible to record unscripted ‘legal presentations’, the researchers at least were able to obtain handwritten versions of such texts, which they included in the...
written component. Finally, no recordings from parliament or the courts could be made in Kenya or Tanzania (i.e. there are no ‘parliamentary debates’ or ‘legal cross-examinations’ texts in the ICE-EA corpus), but transcripts of spoken dialogue from both institutions in Kenya were made available to the researchers. Since these transcripts had been composed by third parties and not by the ICE-EA team, it was decided to create a new category in addition to ‘spoken’ and ‘written’, the ‘written-as-spoken’ component (Hudson-Ettle and Schmied 1999: 6–8). Finally, unlike the written section of ICE-GB, ‘skills/hobbies’ texts do not appear to be of particular cultural importance in East Africa, so that no texts were sampled for this category.

As the above discussion shows, while the ICE-EA is modelled as closely as possible on the ICE-GB corpus to ensure comparability, the situation in Kenya and Tanzania required several changes to the sampling scheme to ensure the representativeness of the data (see Schmied 1996). Nevertheless, due to their balanced and comparable composition the ICE-GB and the ICE-EA corpora can be considered the most reliable data sources available for any contrastive study on British and Kenyan English.

The greatest disadvantage of the ICE-EA corpus for the present study was that so far it has not been tagged or parsed. Yet, in order to allow the corpus results to be generalized to the underlying population it was considered essential to retrieve all instances of stranded and pied-piped prepositions. The only way to guarantee full retrieval meant printing out and manually reading through the entire Kenyan subcorpus.

2.2.2 Forensic tools I: Goldvarb, R, Coll.analysis 3 and HCFA

After having described the corpus evidence which has been investigated for the present study, I will now present the forensic tools, i.e. the programs used for its statistical analysis.

The choice of the correct statistical test for a set of data greatly depends on what type of response/dependent variable one is investigating and what kind of explanatory/independent variables one suspects of affecting the response/dependent variable (see Crawley 2005: 1–2). In the main corpus studies I carried out (see chapter 4), the response / dependent variable ‘preposition placement’ was binary, i.e. it only had two variants, ‘stranded’ or ‘pied-piped’. Furthermore, all explanatory variables were categorical, i.e. had two or more levels but no continuous intermediate levels (for the apparent exception of the factor complexity, see section 3.5). For such cases the correct statistical test is a subtype of the Generalized Linear Model called binary logistic regression analysis (cf. Baayen 2008: 195–208; Crawley 2005: 2, 117–18, 260–80; Gries 2008: 284–94; Paolillo 2002). The statistical research tool deemed appropriate for the present study was
the Goldvarb 2001 computer program for Windows\(^5\) (Robinson, Lawrence and Tagliamonte 2001), which provides descriptive statistical information as well as a multivariate binary logistic regression analysis of the data.

Now while Goldvarb might seem like the straightforward choice to readers with a sociolinguistic background, many people familiar with other statistical packages often argue against using the program for a number of reasons (cf. e.g. Szmrecsanyi 2006: 88–9). Before addressing the main points of criticism, let me stress that I see statistical analysis as a tool and not an end in itself. Furthermore, to be honest, my main motivation for using Goldvarb was that preliminary analyses on parts of the British English data had already been carried out with this program (Hoffmann 2005, 2006). It is true that Goldvarb is a single-purpose software that only carries out binary logistic regression, while software suites such as the commercial program SPSS or the free, open-source program R (R Development Core Team 2008) offer a much greater choice of statistical tests and come with much better graphical facilities. In fact, this wider range of options, together with the recent publication of a number of great introductory textbooks on statistics with R for linguists (such as Baayen 2008; Gries 2008; K. Johnson 2008), probably means that in the future I will abandon Goldvarb altogether and opt for R instead. Nevertheless, from a statistical point of view all that matters for the present book is that the corpus data required a binary logistic regression analysis, which is exactly the analysis that Goldvarb provides. (Though, as the following discussion will show, all Varbul models were also checked in R and the information from these analyses is included in the Appendix for readers who are more familiar with logistic regression in R.)

One point of criticism that is occasionally levelled at Goldvarb is the fact that it reports its parameters on a different scale from most other statistical software.\(^6\) Logistic regression parameters can be reported on either the logit or the probability scale (Paolillo 2002: 162). While other statistical programs, such as SPSS or R, report parameters on the logit scale (ranging from –∞ to ∞ with a neutral value of 0), Goldvarb gives factor weights on the probability scale. As a result, the neutral value for Goldvarb factors is 0.5, with factors ranging from 0 to <0.5 having an inhibiting and those from >0.5 to 1 having a favouring influence on the investigated variant of the dependent variable. Yet, since the logistic function is the inverse of the logit function, Goldvarb parameters share the same properties as the SPSS/R logit parameters (Paolillo 2002: 160–2): values that are equidistant from the neutral value in either direction have the same magnitude of effect (thus Goldvarb factors of

---

\(^5\) Source: www.york.ac.uk/depts/lang/webstuff/goldvarb.

\(^6\) E.g. an anonymous reviewer objected to the use of Goldvarb, stating, amongst other issues, that the program’s ‘outputs are less natural to interpret than log odds’.
Corroborating evidence: Data and methodology

0.2 and 0.8 have the same magnitude of effect, but in the opposite direction. Furthermore, ‘small differences in parameter values have a smaller effect around the neutral value’ (Paolillo 2002: 162; i.e. around 0.5 in Goldvarb and around 0 in SPSS/R). On top of that, Goldvarb’s probability weights can easily be transformed into logit coefficients (or to be more precise sum contrast log odds) via the formula \( \ln(p/(1-p)) \). Thus Goldvarb weights of 0.4 and 0.6 correspond to logit log odds of -0.405 and + 0.405, respectively (D. Johnson 2009a: 361). I understand that, depending on individual preferences, people might find one or the other way of reporting regression coefficients/factor weights easier to interpret.\(^7\) In order to facilitate the interpretation of the logistic regression analyses, it was therefore decided to re-run the Goldvarb analysis in R using Rbrul (www.ling.upenn.edu/~johnson4/Rbrul.R; D. Johnson 2009a, b). Amongst other options, Rbrul emulates Goldvarb analyses in R and outputs the coefficients on the probability as well as the logit scale. Consequently, all logistic regression models presented in this book will give the effects of independent factors as Varbrul probability weights as well as logit log odds. This should enable all readers to interpret the effects of independent factors, regardless of whether they have a Varbrul or SPSS/R background.

Next, independent variables with a categorical effect (e.g. contexts which obligatorily induce stranding or pied-piping) are always problematic for binary logistic regression analyses. Goldvarb, for example, cannot compute such categorical effects or ‘knockout constraints’ at all (Young and Bayley 1996: 272–4; Sigley 1997: 240). Consequently, tokens exhibiting such factors either have to be eliminated from the data, or grouped together (‘recoded’) with other non-categorical factors from the same factor group, provided there are sufficient linguistic reasons supporting such a regrouping\(^8\) (Paolillo 2002; Sigley 1997: 240; Young and Bayley 1996: 272–4). Again this might be construed as a specific disadvantage of Goldvarb since SPSS, R and other software packages can calculate coefficients for such categorical effects (cf. Baayen 2008: 196). However, the parameters for knockout effects reported by these programs can be unreliable and are often accompanied by various warning messages (this phenomenon is usually discussed as (quasi-) complete separation in the statistical literature; see e.g. Allison 2008; Field 2009: 274–5; Hosmer and Lemeshow 2000: 138–40). Thus the issue of how to deal with knockout constraints is one that has to be addressed by all researchers using logistic regression, regardless of the software they are working with.

\(^7\) Note, however, that it’s not only people working with Varbrul that find probabilities easier to interpret than logits (see e.g. Gelman and Hill 2007: 80).

\(^8\) A third possibility would have been to add a ‘fictitious token’ (cf. Paolillo 2002) coded only for the categorical environment, and for the dependent variant. However, such a fictitious token might distort model results if the categorical environment is not very frequent (cf. Hoffmann 2005).
Even more importantly, from a linguistic point of view, I agree with Daniel Johnson (2009b) that truly ‘invariant contexts … should be excluded from an analysis of variation’. As I will show, preposition placement is constrained by several such categorical effects (e.g. obligatory stranding with that-relativizers: cf. the data that he talked about vs *the data about that he talked). In cases where it was suspected that knockout effects were due to genuine categorical constraints, the affected tokens were therefore dropped from the logistic regression analysis. In order to be able still to further investigate these constraints and their interaction with other variables, alleged categorical effects were then tested experimentally and the corpus data in question were subjected to various other statistical tests (see below).

Szmrecsanyi (2006: 88) voices two further points of criticism of Goldvarb, namely that the program ‘cannot easily handle interactions’ and ‘does not report the extent of variance explained’. While I consider the former unproblematic (since all that matters is that interactions can be modelled at all; see e.g. Sigley 2003), the latter is definitely more serious: even if the best logistic regression model has been identified, it still has to be investigated how well this model actually fits the data. Model selection in Goldvarb proceeds via ‘step-up/step-down regression analysis’, which tests the contribution of the individual factor groups for significance via a \( \chi^2 \)-test (the standard test for maximum likelihood models which can also be used to assess whether a recoded model with fewer parameters is as good as the original, more complex model; see Crawley 2005: 103–24; Paolillo 2002: 140–2; Sigley 2003). Once the perfect, ‘minimal adequate model’ (Crawley 2005: 104) for a data set has been found, Goldvarb offers two parameters that indicate whether it is a good fit for the data: (1) the ‘Fit: X-square’ test (a \( \chi^2 \)-test that checks whether the maximum possible, i.e. the log-likelihood of the data, and the model log-likelihood can be said to approximate each other) as well as (2) a chi-square value for the differences of the actual and expected realizations for each cell created by all factor combinations (‘Error’) and the overall chi-square for all cells. Szmrecsanyi, however, is right that none of these parameters specifies the amount of variation accounted for by the model. Fortunately, however, Rbrul calculates an \( R^2 \) value for logistic regression models (namely Nagelkerke-\( R^2 \)) which allows one to quantify the amount of variation explained by the best Goldvarb model (with Nagelkerke-\( R^2 \) ranging from 0 to 1, which correspond to 0 per cent and 100 per cent of variation explained by a model, respectively).

For all logistic regression models presented in this book I will thus report raw frequencies, probability and log odds logit coefficients as well as model fit parameters (Nagelkerke-\( R^2 \) and the ‘Fit: X-square’ test). On top of this, it will also be checked whether the number of included factors is justified considering the sample size. Following Sigley (2003: 251), the threshold value for the maximum number of \( S \) parameters per \( n \) number of tokens used in this study is that of Freeman (1987): \( n > 10 \) (\( S+1 \)). While this value is likely to be
too optimistic for linguistic studies (since it assumes complete independence of tokens), it at least provided a maximum threshold value against which the results of the corpus studies could be compared. As long as the number of employed parameters was considerably lower than the resulting Freeman threshold, a model was deemed fully acceptable.

Finally, as pointed out in section 2.1, apart from the fit of a model it is also important to preclude the undue effect of individual tokens and to guarantee the homogeneity of the data on which a model is based. Unfortunately, this is something that indeed cannot be tested in Goldvarb. For this reason, the best model identified by Goldvarb was always fed into the R 2.7.1 for Windows software (R Development Core Team 2008). This made it possible to compute the so-called cross-validation parameter (cf. Maindonald and Braun 2003: 121–3, 209–10). This test assesses the predictive accuracy of a model by randomly splitting up the data into a number of subsets (so-called ‘folds’; I always use Maindonald and Braun’s 2003 cv.binary() function for this test, which by default creates ten folds). Each fold then becomes a test set against which the model’s accuracy is assessed. Only if this procedure yields a high value for the cross-validation parameter can it be guaranteed that a model fits well to all of the created folds. Consequently this procedure helps to ensure that individual tokens as well as undue influence of data from single speakers can be factored out.

As a final note on the logistic regression chapter, let me say that I am aware that some readers might find it excessive that I thus report the findings from both Goldvarb and R analyses. Nevertheless, as pointed out above, this approach should make the results accessible to researchers from both traditions (and, since the statistical output of both R and Goldvarb are provided in the Appendix, allow them to check the validity of the proposed models).

Next, while the above statistical tools were used to test the overall distribution of preposition placement, it was also deemed important to check whether specific syntagmatic lexicalizations were significantly associated with particular syntactic constructions. Take for example the relative clause structure in (2.4):

\[(2.4) \text{ I like the way in which he did it.}\]

As will be seen in section 4.4.2, *way in* is actually the most frequent antecedent–preposition syntagm in manner adjunct relative clauses in the ICE-GB. This raises the question, however, of whether the two words are also significantly associated – or in other words whether there is statistical support for the claim that *way in*-manner relative clauses are a lexically stored syntagm for British English speakers. In order to answer questions such as these I will draw on Stefanowitsch and Gries’s covarying-collexeme analysis...
2.2 Exhibit A: Corpus data

Table 2.1 Covarying-collexeme analysis (adapted from Stefanowitsch and Gries 2005: 9)

<table>
<thead>
<tr>
<th>way in slot1 (word way in slot 1)</th>
<th>in slot2 (word in in slot 2)</th>
<th>¬ in slot2 (all other words in slot 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency (way + in)</td>
<td>Frequency (way + ¬in)</td>
<td>Frequency (¬way + in)</td>
</tr>
<tr>
<td>Frequency (¬way + in)</td>
<td></td>
<td>Frequency (¬way + ¬in)</td>
</tr>
</tbody>
</table>

Table 2.1 shows that the underlying data structure of such a covarying-collexeme analysis is always a two-by-two table. The association which is to be tested is the grey shaded cell in Table 2.1, i.e. the syntagm way in. Now in order to statistically assess the association of way and in in manner relative clauses this syntagm must be compared to all other logically possible lexicalizations of manner relative clauses: i.e. the frequency of the antecedent way occurring with another preposition than in (way + ¬in), the frequency of in co-occurring with another antecedent noun (¬way + in), as well as the frequency of all other antecedent noun and preposition combinations (¬way + ¬in).

Normally, the standard test applied by linguists to two-by-two tables such as Table 2.1 is the chi-square test. However, chi-square tests always require the expected frequencies of all cells to be greater than 5 (Woods, Fletcher and Hughes 1986: 144f.), a requirement that is often violated by collocational data such as Table 2.1. In the covarying-collexeme analysis the association of two slots of a construction is therefore instead tested via the Fisher-Yates Exact test. Unlike chi-square tests, the Fisher-Yates Exact test does not require large frequencies and places no particular restrictions on its input data (see Baayen 2008: 113). The only disadvantage of the Fisher-Yates Exact test is that it is computationally expensive and thus extremely time-consuming when calculated manually. Yet this could be avoided by using the Coll. analysis script (Gries 2004a), which allows for the automatic calculation of the Fisher-Yates Exact test via R.

For the interpretation of the output of a covarying-collexeme analysis it is important to note that the Coll. analysis script does not report simple \( p \)-values for the association of two words, but log-transformed \( p \)-values. Thus ‘values with absolute values exceeding 1.30103 are significant at the level of 5% (since \( 10^{-1.30103} = 0.05 \)), and values exceeding 2 and 3 are significant at the levels of 1% and 0.1% respectively’ (Stefanowitsch and Gries 2005: 7). One reason for the log-transformation of the \( p \)-values is that ‘the most interesting values are only located in the small range of 0.05 and 0 (and

\(^{10}\) Since \( \log(0.01) = -2 \) and \( \log(0.001) = -3 \).
24 Corroborating evidence: Data and methodology

many linguists are unfamiliar with the scientific format employed for representing such small numbers) (Stefanowitsch and Gries 2005: 7). Another is that unlike \( p \)-values, the base-ten logarithm of the \( p \)-value (called ‘\( p_{\log_{10}} \)’) can be more easily interpreted as a measure of association strength, especially since in the covarying-collexeme analysis the sign of the resulting \( p_{\log_{10}} \) value is changed ‘to a plus when the observed frequency is higher than the expected one’ (Stefanowitsch and Gries 2005: 7). Thus a \( p_{\log_{10}} \) value of \(-2\) indicates that a pair of words appears significantly less often in a construction than expected, while a \( p_{\log_{10}} \) value of 2 implies that the two words in question are significantly associated (both results being significant at a \( p \)-value of 0.01).

Finally, as was pointed out above, Goldvarb cannot compute categorical factors. Yet, it would, of course, be interesting to see whether specific categorical contexts (say, prepositional passives, which only allow stranding; cf. *He was talked about* vs *About he was talked*) exhibit different preferences in British and Kenyan English with respect to the PP types they license (e.g. whether affected-location PPs such as *this bed has been slept in* are significantly more frequent in British than in Kenyan English). In cases such as these when the correlation of three or more nominal variables was investigated, the data in question were subjected to a ‘configural frequency analysis’ (CFA; Bortz, Lienert and Boehnke 1990: 155–7; Gries 2008: 242–54).

In a CFA each combination of factors is labelled a ‘configuration’. Consequently, a factor arrangement like British English × Affected location PP × Passives would qualify as a configuration. Table 2.2 gives a (fictitious) example of the output of a CFA. It shows that for each configuration of factors Factor _1 × [...] × Factor _N (BritishE × [...] × Passives and Kenyane × [...] × Passives in the table) the observed frequency (‘Freq’) as well as its expected frequency (‘Exp’) is given. Based on these a specific chi-square (‘Cont.chisq’) value is calculated. So far, a CFA thus resembles a normal chi-square test. Yet, while in a simple chi-square test only a single test is performed on the data, in a CFA the significance of each configuration is calculated (‘P.adj.bin’). Since multiple tests are thus carried out over the same data set, the significance values for each configuration have to be adjusted accordingly (e.g. as indicated by ‘bin’ in Table 2.2, via the Bonferroni correction \( p_{\alpha'} = p_{\alpha}/n; \) cf. Bortz 2005: 129; Gries 2008: 244; Sigley 2003). If a configuration then turns out to be significantly more frequent than expected (‘Obs-exp’ = ‘>’) it is called a ‘type’, and if it is less frequent than expected (‘Obs-exp’ = ‘<’) it is said to be an ‘antitype’ (cf. Gries 2008: 245–6). Since the \( p \)-values of configurations can be fairly small (in Table 2.2 2.898e-06 = 0.000002898), the significance value is furthermore often indicated by a series of asterisks (‘Dec’ in Table 2.2, with \( ** = p<0.05, *** = p<0.01 \) and \( **** = p<0.001 \)). For the sake of readability, however, in this book I will only give the significance thresholds in the CFA tables (i.e. \( p<0.05, p<0.01, <0.001 \)), with the precise \( p \)-value figures being...
Table 2.2  *Fictitious output of a CFA*

<table>
<thead>
<tr>
<th></th>
<th>FACTOR_1</th>
<th>[...]</th>
<th>FACTOR_n</th>
<th>Frequ</th>
<th>Exp</th>
<th>Cont.chisq</th>
<th>Obs-exp</th>
<th>P.adj.bin</th>
<th>Dec</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BritishE</td>
<td>.</td>
<td>passives</td>
<td>621</td>
<td>709.5</td>
<td>11.039</td>
<td>&lt;</td>
<td>2.898e-06</td>
<td>***</td>
<td>0.125</td>
</tr>
<tr>
<td>2</td>
<td>KenyanE</td>
<td>.</td>
<td>passives</td>
<td>798</td>
<td>709.5</td>
<td>11.039</td>
<td>&gt;</td>
<td>2.898e-06</td>
<td>***</td>
<td>0.125</td>
</tr>
<tr>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
<td>[...]</td>
</tr>
</tbody>
</table>
made available in the Appendix for each CFA. Finally, $p$-values are affected by sample size and therefore CFAs also include a parameter called ‘coefficient of pronouncedness’ (‘Q’), which is a sample size-independent measure of effect size ranging from 0 to 1 (thus equivalent to $r^2$; cf. Bortz, Lienert and Boehnke 1990: 156; Gries 2008: 252).

If CFAs employ chi-square tests, however, they obviously require the expected frequencies of configurations to be greater than 5 (see above; also Woods, Fletcher and Hughes 1986: 144f.). In cases where this criterion cannot be met, as will often be the case in linguistic studies, the significance of configurations must be calculated by the exact binomial test (cf. Bortz, Lienert and Boehnke 1990: 157). Due to the fact that the HCFA 3.2 script (Gries 2004a) for R automatically carries out an exact binomial test, it was decided to use this program for the present study. Another advantage of the HCFA 3.2 script was that it does not run a simple CFA but a hierarchical CFA (thus the name HCFA) over the data. While CFAs only test the effect of a set of specified configurations (say, FACTOR $1 \times$ FACTOR $2 \times$ FACTOR $3$), HCFAs generally also check whether there are effects of simpler configurations (e.g. FACTOR $1$ alone or FACTOR $1 \times$ FACTOR $2$; cf. Gries 2008: 249–54).

After having presented the first type of evidence used in the present study together with the statistical tools for its interpretation, next I will focus on the second, corroborating type: experimental introspection data.

2.3 Exhibit B: Introspection data

Following Schütze (1996) and Cowart (1997), it was decided that introspection data can only be considered valid and reliable if they are collected and interpreted in a scientifically sound way, following the design of psycholinguistic experiments.

2.3.1 Experiment design

A first important prerequisite for the scientific design of an experiment is that the materials used have been created with the help of so-called ‘paradigm-like token sets’ (Cowart 1997: 13). The underlying idea behind this approach is that it is well known that ‘… an informant’s response to an individual sentence may be affected by many different lexical, syntactic, semantic, and pragmatic factors, together with an assortment of extralinguistic influences that become haphazardly associated with linguistic materials and structures’ (Cowart 1997: 46). In order to minimize these confounding factors, paradigmlike token sets ensure that all these factors are uniformly spread across all the tested items and that differences of judgements of two items can thus solely be attributed to the syntactic phenomenon under investigation. The first step to achieve this goal is to take a particular lexicalization of
2.3 Exhibit B: Introspection data

Table 2.3  Token set example laugh at: You wouldn’t believe the things …

<table>
<thead>
<tr>
<th>Token set</th>
<th>Pied-piping</th>
<th>Stranded</th>
</tr>
</thead>
<tbody>
<tr>
<td>which</td>
<td>... at which Bill laughs</td>
<td>... which Bill laughs at</td>
</tr>
<tr>
<td>that</td>
<td>... at that Bill laughs</td>
<td>... that Bill laughs at</td>
</tr>
<tr>
<td>Ø</td>
<td>... at Ø Bill laughs</td>
<td>... Bill laughs at</td>
</tr>
</tbody>
</table>

a phenomenon and cross all tested conditions until all theoretically possible variants have been created.

Take, for example, an experiment testing preposition placement with prepositional verbs in English relative clauses. Since there are two variants of PREPOSITION PLACEMENT (stranded vs pied-piped) and three types of RELATIVIZER (wh- vs that vs Ø), the resulting token set contains six possible structures (2 types of PREPOSITIONAL PLACEMENT × 3 types of RELATIVIZER; see Cowart 1997: 46–50). Table 2.3 illustrates all these six structures for the prepositional verb laugh at in the sentence You wouldn’t believe __ the things __ Bill laughs __.

Nevertheless, before running an experiment there is still no way to preclude the possibility of particular lexical effects. So if only the six sentences from Table 2.3 were presented to a single subject, the results might be skewed, for example if for some idiosyncratic reason speakers favoured pied-piping with laugh at but not with other prepositional verbs. In addition to this, it is well known that each confrontation with a particular stimulus changes an informant, which might also influence his/her judgement. For these reasons all experimental stimuli must also be ‘counterbalanced’ (i.e. every subject sees all conditions, but never with the same lexical material; Cowart 1997: 93). Thus token sets with six different lexicalizations (e.g. apologise for, dream of, rely on, sleep with, talk about) for every PREPOSITION PLACEMENT × RELATIVIZER factor combination must be created. Then a different factor combination can be taken from each of the six token sets to generate a set of stimuli which contains all conditions but only one sentence from each token set, yielding a so-called ‘material set’ such as Table 2.4.

Furthermore, it is necessary that all lexicalizations of a phenomenon be judged by a different subject, and all informants judge all factor combinations but that no informant see more than one sentence from a single token set (Cowart 1997: 93).

While counterbalancing thus avoids many potentially confounding factors, further precautions are still necessary to ensure objective introspection data elicitation. One additional problem that can affect psycholinguistic experiments is the formation of implicit hypotheses. If an informant were only exposed to experimental stimuli he/she might become aware of the aim of the experiment which in turn might affect and distort his/her judgements.
Corroborating evidence: Data and methodology

In order to preclude such effects it is important to include at least as many balanced, i.e. grammatical and ungrammatical, fillers as experimental stimuli. These fillers then act as distractors and prevent informants from forming implicit hypotheses (cf. Cowart 1997: 93).

In addition to this, ‘[f]atigue, boredom, and response strategies the informant may develop over the course of the experiment can have differing effects on sentences judged at various points in the entire procedure’ (Cowart 1997: 94). It is therefore essential to randomize the order in which the stimuli and the fillers are presented to the informant. This is also important since earlier research (cf. Bock 1987, 1990; Cowart 1997: 51–2) has shown that the preceding sentence can influence the judgement of a following sentence. Only the randomized presentation of the experimental items can guarantee that such order effects do not ‘systematically distort effects to the targeted differences among sentence types’ (Cowart 1997: 51).

As a result of the above considerations, all the experiments presented in this book were designed using token sets which were counterbalanced and supplemented by at least as many fillers and the order of all stimuli was randomized for each informant (for information on the total number of stimuli used per experiment see chapter 5). Finally, since the corpus data for British English as well as Kenyan English come from speakers from the upper end of the sociolinguistic continuum, the informants recruited were all adults (over 18) who have at least completed secondary education (for details of the sampled subjects, see chapter 5).

### Methodology: Magnitude Estimation

After having outlined the design of the experimental stimuli, I now want to give an overview of the specific methodology employed to elicit introspection data. The method used in the experiment was based on the experimental paradigm of Magnitude Estimation (see Bard, Robertson and Sorace 1996; Keller 2000). As psychophysical experiments have shown, human beings are not really good at making absolute judgements, e.g. saying whether a line is 10 or 15 cm long. Instead, in Magnitude Estimation studies

---

**Table 2.4 Counterbalanced material set sentence list**

<table>
<thead>
<tr>
<th>Factor combination</th>
<th>Lexicalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>pied-piping + wh-stranding + wh</td>
<td>I know the man on whom Jane relied.</td>
</tr>
<tr>
<td>pied-piping + that</td>
<td>Sally fancies the guy who Steve talked about.</td>
</tr>
<tr>
<td>stranding + that</td>
<td>You wouldn’t believe the things at that Bill laughs.</td>
</tr>
<tr>
<td>pied-piping + Ø</td>
<td>Brad did something that he apologised for.</td>
</tr>
<tr>
<td>stranding + Ø</td>
<td>Sarah never achieved the fame of Ø she dreamt.</td>
</tr>
<tr>
<td></td>
<td>Jennifer never calls the groupies Ø she sleeps with.</td>
</tr>
</tbody>
</table>
subjects are asked to judge stimuli relative to a reference item. Thus sub-
jects judge whether a given line is longer or shorter than a reference line and
try to express this difference in numerical terms (e.g. saying that a stimulus
is half as short as the reference line). Since such relative judgements seem
easier for humans to make, this approach allows gathering of far more reli-
able results. Recently, several studies (e.g. Bard, Robertson and Sorace 1996;
Featherston 2004, 2005; Keller 2000; Keller and Alexopoulou 2005) have
applied this methodology to sentence-judgement experiments. Thus they
asked subjects to give numerical judgements on sentences proportional to a
constant reference sentence. The results from these Magnitude Estimation
studies indicate that eliciting linguistic judgements via this method allows
‘reliable and fine-grained measurements of linguistic intuitions’ (Keller and
Alexopoulou 2005: 1120). The experiments for this study were all conducted
using the WebExp software (Keller et al. 1998), which includes a cross-
modality (judgement of line length) as well as a linguistic training session
and automatically randomizes the order of presentation of stimuli in the
main experiment (for more information on WebExp, see Keller 2000; Keller
and Alexopoulou 2005).

The WebExp software allows running Java-based on-line acceptability
experiments as well as creating printed versions of these experiments (Keller
et al. 1998: 7, 12). The greatest advantages of online experiments are clearly
that subjects cannot go back and change earlier answers and that they must
respond to all experimental items (since otherwise the software will not
allow them to proceed). On the other hand, notable disadvantages are that
subjects are self-selecting (i.e. that only a limited part of a speech commu-
nity uses the internet and is willing to participate in such studies) and that
subject authentication is not as good as under laboratory conditions (Keller
et al. 1998: 6). Furthermore, it turned out that due to technical problems and
limited internet access no online data could be gathered from the Kenyan
speakers at all.

For the present study it was decided to use the online version of an experi-
ment whenever possible since the advantages clearly outweigh the disadvan-
tages (especially since the self-selecting subjects willing to participate in
such studies turn out to be members of the targeted population: educated
adults). Whenever this was not possible – either due to technological prob-
lems or low response rates – the experimenter personally distributed the
printed questionnaires to a set of subjects, explicitly informing them about
the restrictions of the experiment (i.e. not to spend more than a few seconds
on a single item, not to go back and change earlier answers and to check that
they have judged all items).

After this short introduction to the WebExp experiments, the reader
might wonder whether Magnitude Estimation is in fact a valid way of eliciting
linguistic introspection data. On the one hand, extending the Magnitude
Estimation method to linguistic introspection experiments might seem
Corroborating evidence: Data and methodology

straightforward: since ‘linguistics is a branch of psychology that studies a specialized kind of human perception, it is a sister field to psychophys- ics’ (Bard, Robertson and Sorace 1996: 38). On the other hand, ever since Chomsky’s Syntactic Structures much of mainstream linguistics has taken it for granted that ‘[t]he fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical ones’ (Chomsky 1957: 13). However, if sentences can simply be divided into either grammatical or ungrammatical ones, then the question arises why one should carry out Magnitude Estimation introspection experiments which allow subjects to make far more intermediate judgements?

One obvious reason is that introspection judgements are not judgements on the grammaticality of a structure as such but on its acceptability. Grammaticality only pertains to the ontological status of a linguistic stimulus, i.e. in a technical sense, whether a string is generated by a grammar or not. An acceptability judgement, on the other hand, is how a speaker rates a linguistic stimulus (see Bard, Robertson and Sorace 1996: 33). The acceptability of a sentence is considered to depend not only on its grammaticality but also its naturalness in discourse or its immediate comprehensibility (see Chomsky 1965: 10). Furthermore, unlike grammaticality, ‘acceptability will be a matter of degree’ (Chomsky 1965: 10), i.e. that ‘[t]he more acceptable sentences are those that are more likely to be produced, more easily understood, less clumsy and in some sense more natural’ (Chomsky 1965: 11).

From the description of the design of the introspection experiments in section 2.3.1 it should have become apparent, however, that the aim of the present study was not to establish the discourse naturalness of a particular preposition placement construction. Indeed, the employment of counterbalanced token sets was used explicitly to relate different acceptability judgements to grammatical phenomena (e.g. preposition-stranding with different relativizers; cf. section 2.3.1). Since grammaticality does at least play some part in the acceptability judgement of a sentence, statistical differences in judgements of conditions of counterbalanced token sets can therefore not simply be explained away as merely the effect of discourse naturalness.

Interestingly, some support for fine-grained acceptability studies as a window on fine-grained grammaticality differences even comes from mainstream generative linguists: already in the 1960s Chomsky recognized, for example, that there are ‘degrees of grammaticality’ (1965: 148). While he still maintained that the class of grammatical sentences constituted a single set, he identified various different degrees of ungrammaticality (e.g. distinguishing subcategorization violations such as John found sad from selectional rule violations such as Colorless green ideas sleep furiously; see Chomsky 1965: 148–53). Even in a non–technical sense the idea that there are degrees of ungrammatical- icity can be made intuitive; cf. e.g. the following sentences (2.5a–c):
While (2.5a) is grammatical, in Standard English the string in (2.5b) is ungrammatical since it contains an agreement error (*they is*). In addition to this agreement violation, (2.5c) also lacks the required *-ing* participle marker on the verb *kill*. Accordingly, (2.5c) can be said to be more ungrammatical than (2.5b) (for a more technical explanation, see Chomsky 1961).

Since violations of grammaticality are cumulative and also of differing ontological status, it can be taken for granted that there are at least degrees of ungrammaticality. In order to properly investigate such degrees of ungrammaticality it takes a fine-grained and carefully planned introspection experiment as described in section 2.3.1. In fact, as Sorace and Keller have shown, Magnitude Estimation experiments are particularly well suited for distinguishing, for example, semantic from syntactic violations (or in their terminology ‘soft’ and ‘hard constraint violations’; Sorace and Keller 2005: 1502–3): while the former only result in slightly lower acceptability scores, the latter lead to significantly decreased scores.

Moreover, despite the fact that it is still widely assumed that grammatical sentences are a uniform categorical category, there is an increasing number of linguists who contest this view. Ever since the advent of prototype theory (Rosch 1978; Wittgenstein 1953) it has become clear that speakers often conceptualize linguistic categories not as clear-cut but as fuzzy, with some entities being more central and others more peripheral (see also e.g. Labov 1973; Lakoff 1987; Taylor 1995; and for an excellent reader on the topic of linguistic categorization, B. Aarts et al. 2004). As mentioned in chapter 1, there is a growing body of work within usage-based construction grammars that considers language acquisition an input-driven bottom–up process (cf. Croft 2001; Croft and Cruse 2004; Langacker 2005; Tomasello 2003; and chapter 6). In such approaches more frequent exemplars of a syntactic entity are also more likely to become more deeply entrenched cognitively, thus qualifying as better exemplars of the emerging prototypical mental concept. Drawing on an empirical investigation of particle movement in English, Gries has furthermore argued that better exemplars of a prototypical concept can also be expected to receive higher acceptability judgements than less typical ones (2003: 132–9). Accordingly, statistically significant differences between conditions in experiments employing counterbalanced token sets can be said to be indicative of different degrees of cognitive entrenchment, or in other words of prototypicality effects.

Magnitude Estimation experiments allow speakers to differentiate as many intermediate levels of acceptability as they deem necessary. Yet, if statistically significant differences in judgements can either be interpreted as differences in the degree of entrenchment of grammatical constructions or...
the degree of ungrammaticality, then how can these two be distinguished? As I have argued elsewhere (Hoffmann 2006), it is important in such cases to contrast the judgement scores of the experimental items with those of the fillers. Since all judgements are relative within Magnitude Estimation experiments the set of grammatical and ungrammatical fillers constitute the background against which the effects of the experimental stimuli need to be interpreted. Besides, as mentioned above, the use of a corroborating source of evidence, i.e. corpus data, is of paramount importance.

2.3.3 Forensic tools II: SPSS

All Magnitude Estimation data elicited for the present study obviously had to be ‘forensically’ examined, i.e. subjected to a statistical analysis. For this, the data were entered into the SPSS 12.0 for Windows program. Due to the fact that subjects had employed different scales and values for their judgements, the data were then first normalized by transformation to \( z \)-scores (Featherston 2004, 2005). This procedure ‘effectively unifies the different scales that the individual subjects adopted for themselves, and allows [one] to inspect the results visually’ (Featherston 2005: 1533).

Magnitude Estimation yields response variables which are measured on an interval scale (Bard, Robertson and Sorace 1996: 39), while the experiments presented in this book only tested categorical explanatory variables. In addition to this, all experiments crossed several factors so that more than two means had to be compared. Furthermore, as a result of counterbalancing the token sets (see section 2.3.2) all subjects were tested on all conditions of the experiment. Such a so-called within-subject design is said to use repeated measures and the scores it yields had to be analysed via a ‘repeated measure Analysis of Variance’ (ANOVA; for details see Bortz 2005: 331–60; Crawley 2005: 154; Field 2003, 2009: 457–505; Gravetter and Wallnau 1992: 203).

Repeated measure ANOVAs require that the differences between all pairs of treatment levels have approximately equal variances (Field 2009: 459). If this assumption, known as ‘sphericity’, is not met, the statistical computation of a repeated measure ANOVA has to be corrected accordingly. In SPSS the so-called ‘Mauchly’s test of sphericity’ tests the data for this assumption. If the Mauchly’s test yields a significant \( p \)-value (below 0.05), the results of the following repeated measure ANOVA have to be corrected using an appropriate correction. SPSS offers several options in such cases, but for the present study the conservative Huynh and Feld correction was always applied (see Field 2009: 460–1).

In ANOVA tests, the critical value for identifying significant effects is the \( F \)-ratio calculated by dividing the differences between treatments by the differences within treatments (see e.g. Gravetter and Wallnau 1992: A–90). For each experiment two repeated measure ANOVAs had to be carried
2.3 Exhibit B: Introspection data

out: one testing whether the effects are significant by subject \( (F_1) \), and a second one to see whether they were also significant by item \( (F_2) \). This was necessary since just as the informants actually tested in an experiment are (usually) seen as representatives of the entire population from which they are selected, the token sets are likewise seen as representatives of all the relevantly similar token sets one might construct in the same language (or perhaps any language). Just as statistical tests on data for informants test the reliability of patterns seen in those results, so tests on data for token sets test the reliability of patterns seen in the summaries of the token set data. (Cowart 1997: 122)

Once the \( F \)-ratio score of an explanatory variable (also called ‘factor’ in ANOVAs) was identified as significant in either the by-item or the by-subject analysis, the next question to be addressed concerned the informativeness of this result: i.e. how much variation in the data can actually be explained by a significant factor? The statistical value which provides this information for ANOVAs is the so-called eta-square parameter \( \eta^2 \). This value is calculated by dividing the ‘[s]um of squares (SS) for each factor or interaction in the ANOVA result table … by the total sum of squares for the set as a whole’ (Cowart 1997: 136). In non-technical terms, \( \eta^2 \) can therefore be seen as the proportion of data explained by a particular factor.

Note that a significant ANOVA result only indicates that there is an effect within a factor. As long as there are only two levels or conditions within a factor this does not cause any problems since a straightforward comparison of the means of the two levels will immediately allow identification of the effect. Whenever there are three or more levels, however, the locus of the significant effect cannot be found that easily. Instead, the standard procedure in such cases is to run a post-hoc test over the ANOVA model. Such post-hoc tests involve a pairwise comparison of all levels which is corrected for multiple testing (Gravetter and Wallnau 1992: 372). For the present study the Tukey post-hoc test (Bortz 2005: 325–8; Gravetter and Wallnau 1992: 406) was used whenever a significant factor had more than two levels.

As argued in section 2.3.2, in order to decide whether a particular effect should be regarded as reflecting the grammaticality or the ungrammaticality of a condition it was important to compare these with the set of grammatical and ungrammatical filler items. Because the fillers were not part of the repeated measure ANOVAs, such comparisons had to be carried out by performing a set of dependent \( t \)-tests.\textsuperscript{11} To preclude an inflation of the \( \alpha \)-error in these tests the \( p \)-values for multiple \( t \)-tests were adjusted using

\textsuperscript{11} ‘Dependent’ here has the same meaning as ‘repeated measures’ in the ANOVA: since all subjects judge the same experimental conditions and the same set of fillers, the two treatment conditions tested do not qualify as independent.
the standard, albeit rather conservative, Bonferroni-correction $p_{\alpha}^* = p_{\alpha}/n$. (see Bortz 2005: 129; Sigley 2003; and section 2.2.2).

Finally, in order to visualize the results from the statistical analysis of the introspection data, graphs displaying the mean judgements of significant results together with standard error bars were created using Excel for Windows.

In this chapter I have argued that linguists should treat the analysis of a particular syntactic phenomenon as a criminal case which draws on corroborating evidence from corpora and introspection. In addition to this, I have laid open the kind of evidence as well as the forensic tools (a.k.a. statistical programs) used for the present case. Before turning to an in-depth forensic analysis of the evidence, however, it is important to first survey witness statements and the case notes of other detectives. In other words, the discussion of preposition placement in the linguistic literature has to be examined in the next chapters.