Chapter 4: Grammar and corpus methodology – Sean Wallis

4.1 Introduction

This chapter explores the potential of natural language *corpora* , databases of text samples, for grammatical research. These samples may be entire texts or excerpts, and may be composed of written text, transcribed speech, handwriting, or even sign language. Texts are sampled using a set of criteria termed the *sampling frame* , which specifies the genres, contexts, meta-data and participants sampled.

These 'texts' contain more than words and punctuation. A *plain text* corpus with minimum annotation identifies sentences and, where relevant, speaker turns; it may mark other phenomena, such as headlines, pauses and overlapping speech.

Grammatical research requires either grammatically annotated corpora or 'on-the-fly' grammatical interpretation of plain text corpora. In practice, due to the availability of effective algorithms, most corpora are *tagged* so that every word is grammatically categorised by its word class. In corpus linguistics, tagged corpora predominate. Such corpora can be exceedingly large (hundreds of millions of words upwards), or may be drawn from specialised data sources.

Tagged corpora are an improvement on plain text, but they are rather limited. Grammar is fundamentally structural (see Chapter 23). Word and word class sequences are meaningful in the context of structural grammatical relationships. It follows that corpora with the greatest potential benefit for grammarians are likely to be those where every sentence has been given a full grammatical tree analysis.

Creating this kind of *parsed* corpus (also known as a 'treebank') is difficult and time-consuming, and consequently resources tend to be smaller: typically, a million words or so. Opting to parse a corpus means selecting a particular framework and applying it consistently across diverse naturally-occurring data. Which scheme should we choose? Once selected, are future researchers constrained by our parsing decisions? How do we apply the scheme consistently? We discuss these questions in this chapter.

Corpora have been used to study, *inter alia* , semantics, syntax, morphology and pragmatics. The distinction between applying categorical labels ('tagging') and specifying structural relationships ('parsing') is extensible to other linguistic levels. Indeed, a richly annotated corpus might allow frameworks to be *related* . For example, pragmatic analysis might apply speech act categories (such as 'request' or 'assertion') to main clauses independently categorised by structural type (such as 'interrogative' or 'declarative': see Chapter 18)..

Researchers have approached natural language text data at multiple levels, in different ways, with various software tools. This combination of purposes, approaches and tools falls within the scope of the *methodology* of corpus linguistics. This chapter does not attempt a complete review of tools, or to enumerate the broad range of fields of enquiry, from stylistics to social geography and pedagogy, to which corpus linguistics methods have been applied. For such a review, see McEnery and Hardie (2012), O'Keefe and McCarthy (2012) or Biber and Reppen (2015). This chapter has a particular focus: *which corpus linguistic methods are likely to be of the greatest benefit for the study of grammar?*

Corpus linguistics has grown in popularity over the years, but not all linguists agree about the relevance of corpus data for their subject. Some, like John Sinclair (1992), argue that all linguistic knowledge resides in the text. Others, notably Noam Chomsky have argued (see, e.g. Aarts 2001) that corpus data represents at most a

collection of performances and epiphenomena, the study of which tells us little about the internal linguistic processes that give rise to grammar.

Corpus linguistic data has important strengths and weaknesses for the grammarian. Although corpora can be drawn from a wide range of sources, most corpora are constructed from daily life rather than artificial contexts, and responses are not cued artificially by a researcher. Corpus data is raw primary data, unselected by linguistic introspection. The task of the corpus linguist is to use linguistic insight to interpret this data after it has been collected.

In this chapter, I argue that the optimum position for grammatical researchers is at the intersection of linguistic theory and corpus data analysis. This means working with grammatically annotated corpora, while recognising that the annotation is based on a necessarily partial knowledge of grammar.

This chapter is organised as follows. Section 4.2 considers what a corpus could potentially tell us about language – the classes of evidence that corpus linguistics can offer a grammatical researcher. Armed with these distinctions, the next section returns to the Chomsky–Sinclair dichotomy outlined above. In section 4.4 the proposed solution is further subdivided into levels of knowledge and process.

This set of distinctions allows us to relate different grammatical research programmes with corpora, including top-down corpus parsing, cyclic treebank exploration and bottom-up clustering. To conclude, I discuss how corpus research raises practical problems of experimental design and analysis.

4.2 What types of evidence can a corpus offer a linguist?

Corpus linguistics arrived late to the grammar party. Corpus linguists tend to date the advent of their field with the compilation of the *Standard Corpus of Present-Day Edited American English* (popularly known as the 'Brown Corpus'; Kučera and Francis 1967). This was followed by the Survey of English Usage Corpus of spoken and written British English (the 'Quirk Corpus'), begun on paper in 1959, the spoken part of which was published electronically in 1990 as the *London-Lund Corpus* .

What made these corpora different from simple collections of texts was a focus on scale and sampling. Firstly, these corpora were *substantial* collections of naturally-occurring text samples. They were of the order of a million words – small by today's standards, but larger than contemporaneous resources. Secondly, they were consciously collected with the aim of creating a *representative* sample of the relevant English language variety. Generalisations from such a sample might be said to be 'representative of the language', or more precisely, a well-defined subset of it.

The novel contribution of corpora lay not in their being a source of natural-language examples. Indeed, linguists had long drawn insight from, and accounted for, real-world examples. Thus in 1909, Otto Jespersen wrote "I have tried… to go to the sources themselves, and have taken as few facts and as few theories as possible at second hand" (Jespersen, Preface to Part I: 1954:VI). Pre-corpus sources tended to be limited to the library of a particular grammarian, the Bible, or the works of Chaucer or Shakespeare.

Moreover, from a grammatical perspective, representativeness has one major drawback. Test cases that expose theoretical distinctions between analyses may be too infrequent to be found in a typical corpus. Constructing artificial conditions to elicit examples from speakers may be a more effective approach (see Chapter 3).

The benefits of a corpus for grammatical theory lie elsewhere. There are three types of evidence that may be obtained from a corpus (Wallis 2014), which can apply

to many different linguistic phenomena. These phenomena, instances of which we will term a 'linguistic event', x , might be (for example) an individual lexeme, a group of words, a prosodic pattern, a grammatical construction, a speech act, or any *configuration* of these, such as a question speech act employing a rising tone. The three types of evidence are **factual evidence** (event x took place, written 'Exists(x)'), **frequency evidence** (x occurs '$f(x)$' times) and **interaction evidence** (x tends (not) to co-occur with another event y).

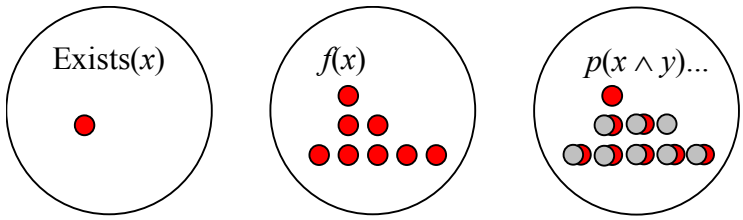

Figure 1: Three types of evidence: left, factual evidence: event x is found in a corpus; middle, frequency evidence: x is distributed in the corpus by frequency; and right, interaction evidence: y co-occurs with x more/less often than chance would predict.

Factual evidence is simply evidence that an event occurred in the corpus, i.e., it was expressed at some point in the past. One application of this is the identification of novel events not predicted by a framework, discussed in 4.2.1 below.

Corpus linguistics is most strongly associated with **frequency evidence**. A single observed frequency, $f(x)$, is an observation that a linguistic event appears in a corpus a certain number of times. More useful is a frequency *distribution* : a set of frequencies of related events. Thus the first application of the Brown Corpus was a word frequency list and set of distributional analyses, Kučera and Francis (1967).

Distributions allow researchers to compare related frequencies. For example: the word *pretty* is conventionally considered first as an adjective in dictionaries.[1] However, Nelson *et al* . (2002: 233) report that in the *British Component of the International Corpus of English* (ICE-GB), *pretty* is found 85% of the time (112 times out of 132) as an adverb.[2] Another type of frequency distribution might reveal variation in the frequency or rate of the same phenomenon over a sociolinguistic contrast, such as genre, speaker gender, or time (see e.g. Chapter 27).

The third type of evidence is **interaction evidence**. These are observations that two events tend (or tend not) to be found together, sometimes referred to as 'association', 'attraction' or 'co-location' statistics. This evidence is interesting because if two apparently independent events coincide more frequently than expected, there may be a deeper connection between them. It is also possible to find evidence of negative association or 'horror aequi' (Rohdenburg 2003: 236). We discuss interaction evidence in 4.2.2.

4.2.1 Factual evidence and the validation of frameworks

When corpora are constructed, one of the first tasks of the compilers is to choose a framework and completely annotate the corpus with it.

---

[1] In the OED, *pretty* (adjective) has between seven and eight times the space afforded for the description of the adverb form (OED, 1961: VIII 1332-1333).

[2] Since corpora are of different sizes, many corpus linguists cite frequencies 'normalised' by scaling per thousand or million words before reporting. However, the number of words per text is often not the optimum basis for comparison. A better approach is to pick a meaningful baseline. See 4.5.2.

In a tagged corpus, every word must be classified, including *novel words* , i.e. words that have not previously been given a word class. Thus if we applied a scheme developed with a standard language source to a corpus of regional dialect speech we might find words with no agreed classification, which a human linguist would need to determine. The same principle extends to parsing or any other framework.

A corpus can be used to validate a framework through a process of identifying 'gaps' and encouraging reappraisal of the scheme. First, the corpus is subjected to an annotation process, e.g. the texts are tagged and parsed. Second, a search process is applied to the entire corpus to identify unannotated cases.

Let us take a real example. Quirk *et al.* 's (1985: 54) transitivity framework describes the following complementation patterns:

TRANSITIVE VERBS →
- MONOTRANSITIVE VERBS occur in type *SVO*
- DITRANSITIVE VERBS occur in type *SVOO*
- COMPLEX TRANSITIVE VERBS occur in type *SVOC* and *SVOA*

Quirk *et al.* offer *My mother enjoys parties* (where *parties* is a direct object) as a monotransitive pattern. The vast majority of *SVO* patterns (64,482 in ICE-GB) are of this type. Yet in parsing ICE-GB, the compilers found some 271 clauses where the single object was an indirect object, like *she told me* (cf. ditransitive *she told me a story* , with indirect object + direct object). They had two options:

- **Create a new category.** The ICE-GB team decided to create a new 'dimonotransitive' type for these patterns. The distinction between direct and indirect objects was considered sufficiently important to be recorded as a transitivity feature.
- **Modify an existing category** to include the novel pattern. A literal reading of Quirk *et al.* might treat 'monotransitive' as encompassing all single-object complementation patterns: *she told me* is monotransitive. Alternatively, perhaps *she told me* is considered to have an ellipted direct object, in which case we might argue it is ditransitive. Irrespective of which approach is taken, the distinction is not recorded as a transitivity feature.

The process of incorporating novel events into a framework is not merely one of applying a label, but one of *grammatical argumentation* (Chapter 2).

Contrast this process with the *ad hoc* classification of novel examples without a corpus. The non-corpus linguist, initially at least, finds a single unexplained example. However, it can be difficult to determine a single differentiating factor from just one case, or to decide if the instance is simply anomalous. Obtaining further attested examples of infrequent phenomena requires exhaustive manual reading.

It is commonly said that a corpus cannot tell us what is impossible in a language. Corpora do tell us what *is* possible – occasionally, contrary to our expectations. Validating frameworks against data makes them empirically *more robust* , so the probability of finding new novel events in the future tends to decline.

4.2.2 Interaction evidence

Interaction evidence is little discussed in corpus linguistics textbooks. It is empirical evidence that theoretically independent events actually tend (or tend not) to co-occur. This evidence is relevant to grammarians at both abstraction and analysis levels (see 4.4).

Interaction evidence may be *associative* (bi-directional) or *directional* . The simplest type is associative. Consider two events, x and y , assumed to arise independently, each with probabilities $p(x)$ and $p(y)$. The *expected* joint probability (the chance that they will co-occur independently) is simply the product, $p(x) \times p(y)$. We measure the *observed* rate of co-occurrence, $p(x \wedge y)$, and compare figures with a statistical test. The result is a statistical correlation that can occur for a variety of reasons. This principle is used in many computer algorithms, from part of speech tagging and probabilistic parsing to collocation analysis. It can also be used in interaction experiments (see 4.5.3).

Collocation analysis (see also 4.4.5 below) finds word pairs tending to co-occur (co-locate) next to each other. In the *British National Corpus* (BNC), *askance* tends to be immediately preceded by the lemma LOOK. The chance of the next word being *askance* jumps up dramatically if the current word is *look* , *looks* , *looked* or *looking* . This example is directional. The probability that *askance* is preceded by LOOK is around 65% in the BNC (LOOK *askance* is idiomatic). This is far greater than the probability that LOOK is followed by *askance* , which is around 0.03% (LOOK is much more likely followed by *up* , *forward* , etc).[3]

Directionality is taken further in word class tagging algorithms. These use *transition probabilities* , $p(y \mid x)$, 'the probability of y given x '. A training algorithm analyses a previously-tagged corpus and creates a database, which is then used by a tagging algorithm to tag new texts.

Let us take a simple example. In ICE-GB the word *work* appears 982 times: 288 times as a verb and 694 times (70%) as a noun. This is a frequency distribution. However, if *work* is immediately preceded by an article, the chance of *work* being a noun jumps to 100%. (In all 115 cases of article + *work* in ICE-GB, *work* is a noun.) We can collect an *interaction distribution* of transition probabilities from the corpus, and use it to make tagging decisions in the future, thus:

article	*the work*	115	out of 115	(100%)
preposition	*to work*	149	149	(100%)
adjective	*recent work*	137	138	(99%)
adverb	*presently work*	4	30	(13%)
…				

Note how adverbs predict the opposite outcome, i.e. 26 times out of 30, *work* is a verb.

Events x and y need not be adjacent for interaction evidence to be derived, merely found in the same text. The exact relationship depends on the research question. The same principle can be used to study *grammatical priming* using a corpus (Gries 2005). In this model, events x and y are instances of the same grammatical structure (Speaker A says x , and 'primes' Speaker B to say y later on), potentially separated by many utterances.

Patterns of interaction may arise for several reasons. Some co-occurrence tendencies are due to a grammatical relationship, such as the known relationships between articles, adjectives and nouns illustrated above. Others may be due to psycholinguistic processes of attention and memory operating on grammatical rules (Wallis 2012).

[3] Despite the highly directional evidence, the bi-directional 'mutual information' score places *askance* first in collocates of LOOK. See also https://corplingstats.wordpress.com/2017/03/28/direction.

Unfortunately, interaction can also arise for more trivial reasons: a text concerns a particular topic (Church 2000), or patterns reflect semantic associations and idioms. What at first sight appears to be grammatical 'priming' may be mere lexical repetition. Detecting patterns is only a starting point. We may need multiple experiments to distinguish plausible causes.

Frequency and interaction evidence are open to *statistical inference* . If the corpus is representative of the language from which it is drawn, and instances are drawn from many participants and texts, it becomes possible to argue that detected preferences are not due to an individual writer or speaker but are typical of the language community. We can make sound statistical claims about the population of sentences from which the corpus is drawn.

4.3 Approaches to corpus research

Commonly, a contrast is drawn between 'corpus-based' and 'corpus-driven' linguists (Tognini-Bonelli, 2001). A corpus-based linguist is one whose research is primarily theoretical, who uses a corpus for exemplification and hypothesis-testing. For them, a corpus is a source of knowledge about grammar that must be interpreted by theory.

While widely used, the term 'corpus-based' is far too general. It includes any linguist who uses a corpus but does not claim to rely on corpus data exclusively, i.e. a corpus linguist who is not 'corpus-driven'.

This seems unsatisfactory. There are many possible approaches to research, ranging from an extreme theory-driven approach that avoids corpus data (typified by Chomsky) to an extreme corpus-driven one that sees all theory as inevitably incorrect (typified by Sinclair). Instead of 'corpus-based' vs. 'corpus-driven', it is clearer to say there is a continuum of corpus research perspectives between 'theory-driven' and 'corpus-driven' poles.

4.3.1 Corpus-driven linguistics

'Corpus-driven' linguists argue against what they perceive as a necessarily selective approach to the corpus. John Sinclair (1992) objected that the grammatical tradition of top-down research (with or without a corpus) resulted in a plurality of grammatical frameworks with no agreed way to select between them. Corpus-based linguists are vulnerable to research bias. They discover what they expect to find and explain away counterevidence as 'performance errors'. 'Research' is reduced to categorising data under a pre-existing theory, rather than an attempt to critically engage with theories.

Consider the 'dimonotransitive' verb category we discussed earlier. The compilers of ICE-GB did not overturn the category of transitivity, but extended it to account for problematic examples. Sinclair's point is that corpus-based researchers tend to 'patch' their framework rather than reconstruct it from first principles – and risk reappraisal of their previous research.

Corpus-driven linguists adopt a different starting point. Research should start from the plain text, and researchers should derive theoretical generalisations from the corpus itself. Let us make a minimum set of assumptions and see where this takes us. We can harness computational power to process many millions of words and identify new generalisations.[4]

[4] See http://corplingstats.wordpress.com/2016/12/16/pos-tagging.

Sinclair's achievement was the construction of the proprietary *Bank of English* ™ corpus of over 650 million words, and a grammatical framework (published as the *Collins Cobuild Dictionary* and *Grammar* , Sinclair *et al*. 1987 and 1990) that his team reportedly compiled 'bottom-up' from the corpus.

However, the Cobuild project raises an obvious objection. If it is possible to obtain a grammatical framework from a corpus, how should we refine or extend *this* grammar? Are we obliged to start again (as true corpus-driven linguists), or may we take this work as a starting point, thereby incorporating generalisations found in the first stage as theoretical assumptions for the next? But if we adopt the latter course, are we not becoming corpus-based?

4.3.2 Theory-driven linguistics

Probably the most famous example of a theory-driven position is found in Noam Chomsky's comments on corpus linguistics (Aarts 2001, Chomsky 2002). His explanation of corpus evidence as ultimately constituting evidence of performance, rather than indirect evidence of some kind of internalised language faculty, might appear to place him outside of corpus linguistics altogether.[5] Linguistics is then principally an exercise in deduction (Chapter 2). Nonetheless, many linguists strongly influenced by Chomsky's theories have engaged with corpus evidence (e.g. Wasow 2001). Instead of starting with a corpus and generalising upwards, they have attempted to test hypotheses against corpus data.

The principal point of reference for a top-down, theory-driven researcher is their theoretical framework. Corpus data is interpreted by the theory, so contrary evidence could be interpreted as epiphenomena and exceptions, or even ignored.[6] It is ultimately not possible to formally disprove a theory by this method.

Yet a key goal of all science lies in attempting to improve theories. We have already seen how a corpus can identify phenomena unanticipated by a grammatical framework. Using a corpus, is it possible to identify and test the kinds of theoretical predictions that might cause us to choose between competing frameworks?

4.3.3 Transcending the dichotomy

The resolution of the two positions – 'theory-driven' and 'corpus-driven' – starts with the following observation: neither extreme is necessary. Instead, we can agree that any current linguistic theory is almost certainly incorrect, but a theory is necessary to make progress – including identifying where it fails.

In other sciences, theories are understood to be necessarily partial – indeed, potentially untrue and misleading – but also a necessary part of the scientific process (Putnam 1974). It is not possible to avoid 'theory', thus compiling a word list requires us to define a 'word'. If theories are unavoidable, we must state their assumptions. See also Chapter 2.

All theories include *auxiliary assumptions* , i.e. assumptions outside of the theory proper that are necessary to obtain data. Thus, in astrophysics, one cannot directly measure the chemical composition of stars. Instead researchers collect spectrographs, where each light 'spike' matches the characteristic wavelength of a

[5] See http://corplingstats.wordpress.com/2016/11/02/why-chomsky.

[6] Labov (1966: 27) writes "[I]n listening to everyday speech, we tend to hear only those linguistic features that have already been described, and it takes a major effort to hear the new variables that are being generated in the speech community."

fluorescing atom. But this raw data is further distorted by Doppler shifts (due to the star moving relative to the viewer), bent by gravity, etc. Auxiliary assumptions are found in calculations to calibrate equipment and interpret measurements.

How are observations ultimately interpreted? By comparing spectrographs from distant and local stars, and by relating observations to an overall theory of stellar decay: in short, by comparing expected and observed results and making sense of what remains. Moreover, a systematic difference between expectation and observation can obtain *indirect observations* , i.e. observations that cannot be perceived directly but may be inferred by another observed effect. Famously, Pluto was first detected by perturbations (wobbles) in the orbit of Uranus.[7]

Naïve falsification (Lakatos 1970), where hypotheses are rejected on a single piece of counterevidence, is unusual in science (the exception being fields obliged to have a low tolerance for error, such as clinical trials). Counterevidence may even be revelatory. Rather, progress is often made by *triangulation* , that is, support for a position builds up when multiple independent approaches tend to converge on the same conclusion, and by *competition* between alternative theories to explain observations. In short, Lakatos (1970) and Kuhn (1977) argue, parallel research programmes coexist and progress to a certain point, whereupon the dominant theory fails to explain new phenomena or is supplanted by a more effective theory.

What does this mean for linguistics? Since neither theory nor data can be dispensed with, extreme top-down and bottom-up positions are untenable. Instead, linguists should adopt a position where corpus data and linguistic theory are engaged in a critical dialogue or *dialectic* : testing the theory against the corpus while enriching the corpus theoretically to make sense of data. Researchers must commit to a particular theoretical framework (albeit temporarily) to make progress.

Finally, if all theories are necessarily partial, then we must transcend the corpus-driven / theory-driven dichotomy by a *cyclic methodology* .[8] Knowledge ultimately derives from the attempt to explain data by theory. But researchers can work from the data up, or from the theory down, as the need arises.

4.4 Tools and algorithms for corpus research

Modern corpus linguistics could not exist without computation, and a mini-industry of tools and algorithms, 'toolkits' and platforms, has grown up alongside corpora. Some tools help build and annotate corpora; others explore existing corpora. Some tools, such as automatic taggers and parsers, perform different tasks. Others, like automatic parsers and manual tree editors, perform comparable tasks with different methods.

To make sense of diverse algorithms and approaches we need finer distinctions than 'top-down' and 'bottom-up'. Wallis and Nelson (2001) propose a *3A perspective in corpus linguistics.* This identifies three processes that generalise from raw text to linguistic hypothesis: 'annotation', 'abstraction', and 'analysis'. Figure 2 summarises the idea.

The 3A model is fundamentally cyclic. Each process is capable of upward and downward application. Thus 'annotation' is usually considered top-down (applying a framework to text), but, as we observed in 4.2.1, detecting novel elements in a text may generate new terms in a scheme, bottom-up.

---

[7] See www.discoveryofpluto.com for an introduction to this topic.

[8] The idea that complex systems of knowledge undergo cyclic development is found in computer science (Boehm 1986) and other fields.

The model also identifies three more general knowledge layers above the source text. These are the 'corpus' (text enriched with annotation); 'dataset' (example set extracted from the corpus); and 'hypothesis space' (set of hypotheses testable on the dataset).

Abstraction (and its reverse, 'concretisation') maps abstract terms to particular examples in the annotated corpus. Consider a linguist concerned with investigating verb phrase complexity. She uses a concept of a 'complex verb phrase', possibly graded by complexity. Her notion of complexity is theory-laden and is not represented in the annotation. However, she can translate it into queries (Wallis 2008; see also 4.4.4) applicable to the annotated corpus.

Abstraction performs a crucial task: mapping specific, concrete example structures, patterns or elements in observed sentences to general conceptual terms. Like annotation, these concepts form part of a systematic framework, but this framework belongs to the researcher, not the annotator.

Usually researchers wish to extract examples of a set of linguistic events, rather than a single one. Multiple related queries are typically required, so abstraction should also collect data together in a dataset amenable to analysis, for example to determine if variable X and Y interact.

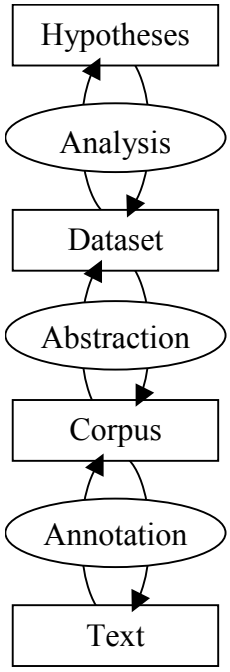

Figure 2: The 3A perspective in corpus linguistics (after Wallis and Nelson 2001).

The final stage, Analysis, is a process of evaluating this abstracted dataset for generalisations ('hypotheses' in the experimental paradigm).[9] A single hypothesis, such as 'spoken data contains a lower proportion of complex VPs than written data', can be tested against the dataset, or multiple hypotheses may be evaluated together. Analysis can also be cyclic. The results of one experiment may trigger further refinements of the research design.

The 3A model contains six process arcs – three up, three down – and four levels of linguistic knowledge (Figure 2). Different algorithms (taggers, collocation analysers, search tools, etc.) operate on one or more of these arcs. The same principle applies to manual processes under a linguist's direct control, such as browsing examples and intuiting new concepts and queries.

Conceiving of corpus linguistics in this way has two advantages. First, it allows us to integrate different tools in a single platform and identify tools performing parallel tasks. Second, it offers options to linguists who might otherwise fail to see a way forward. For example, analysis software may generate impressive visualisations, but if the linguist has no access to the underlying sentences they cannot verify patterns. Which set of sentences does a datapoint reflect? Are results genuine or an artefact? What follow-up experiments are necessary to further distinguish hypotheses? Only by returning to the corpus and the original sentences is it possible to find the underpinning of analytical results. Returning to the source data should not be a mere afterthought.

The following sections exemplify this perspective. We discuss concordancing in section 4.4.1, lexical search in 4.4.2, parsing in 4.4.3, software for exploring parsed corpora in 4.4.4, and bottom-up exploratory methods in 4.4.5.

[9] To take a radically different example, an 'analysis' stage in an automated telephone answering service might classify input patterns and trigger a response.

4.4.1 Concordancing tools

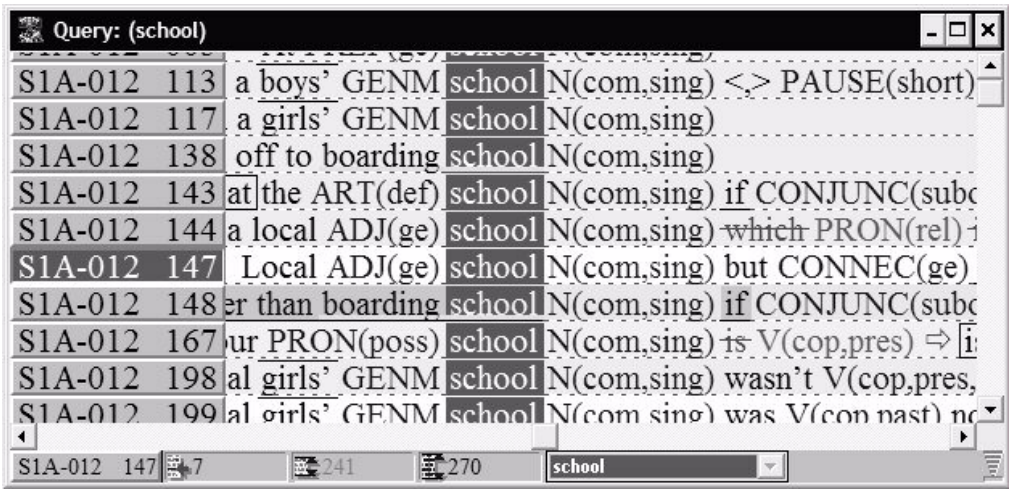

Figure 3: Example of a Key Word in Context concordance for the lexical item *school* in ICE-GB, showing adjacent word class labels.

As corpora have grown, the need for specialised software has also grown. Although some early corpora, such as the Quirk Corpus, were built without computers, the benefits of computerisation soon became obvious. Initially, corpus developers tended to use existing computer programs, such as standard text editing software and databases, to construct early corpora.

The first set of software tools developed specifically for corpus linguistics were concordancing tools. Popular current examples include *AntConc* (Anthony 2005) and *WordSmith* (Scott 2012). 'Concordancing' originated in Bible studies, where citations of words in context were deployed in theological discussions.

Key Word In Context (KWIC) concordances display the results of a corpus search – a particular word, morpheme, part-of-speech tag, etc. – in their immediate source sentence context. A simple search for a word can generate a large set of results for a researcher to browse.

Concordancing is a bottom-up exploratory technique *par excellence* . A simple search can uncover patterns of use 'emerging' from the text by exposing contrasts in a set of examples. Compare the concordance view in Figure 3 to the conventional 'literary' way we consider language in a narrative context where contrasts with other examples are unavailable. Figure 3 reveals that in the ICE-GB text 'S1A-012', *school* is often prefigured by *boy's* , *girl's* , *local* or *boarding* – a fact that is immediately apparent, but might not be guessed by simply reading the text.

4.4.2 Lexical-grammatical search in tagged corpora

To perform deeper research, we must exploit linguistic distinctions. In the introduction I suggested that a parsed corpus (or 'treebank') offered the greatest opportunities for grammatical research.

The dominant movement in corpus linguistics has been to compile ever-larger word-class-tagged corpora. The 650 million-word Bank of English is a fraction of the commercial *Collins Corpus* of 6.5 billion words. Mark Davies has compiled and published several multi-million word corpora totalling 1.9 billion words and growing (see http://corpus.byu.edu).

Although these corpora are not parsed, it is possible to perform searches for small phrases and clauses using sequences of word class tags and words. This method combines the benefits of large corpora with a 'quasi-parsing' approach using search strings and careful review. But we cannot guarantee to accurately find all examples.

Thus, using the tagged *Corpus of Historical American English* (COHA), Bowie and Wallis (2016) employed a search string to find examples of the *to* - infinitival perfect (as in *to have forgotten*) occurring as complement of a preceding governing verb (as in SEEM *to have forgotten*). The search pattern "<verb>, *to* , *have* , <past participle verb>" finds cases like *seems to have forgotten* and *[was] considered to have begun* , but excludes others.

For instance, noun phrases of varying length can appear between some governing verbs and the particle *to* , as in *considers [the space race] to have begun* . Without parsing, extensive manual analysis is the only way to reliably identify such examples: additional search strings can be devised to allow for varying numbers of intervening words, but these retrieve numerous 'false positives' (instances of irrelevant structures).

4.4.3 Corpus annotation and parsing

Parsing is a considerably more complex problem than tagging. In a review that has stood the test of time, Briscoe (1984) estimated that automatic parsing algorithms correctly parse English sentences approximately 75% of the time.[10] If we apply an automatic parser to a corpus, we can expect to perform considerable manual annotation. Word class tagging algorithms achieve a 5% error rate for English, which we might accept, but a 25% parsing error rate is not tolerable for linguistic purposes. Moreover, the errors are likely to include interesting problems, like the dimonotransitive (see 4.2.1).

Since their goal is to obtain a better parsing algorithm, natural language processing (NLP) researchers treat their algorithm and database as primary. The corpus is simply test data. Improvement occurs by modifying the algorithm and re-testing it against the corpus. Correcting the *corpus* seems irrelevant to improving algorithms.

When developing a treebank corpus, the cost of hand-correcting parsing is substantial. However, the exercise has benefits. We have already seen how completing the annotation of a corpus can validate and extend a framework. 'Corpus parsing' is not merely an exercise in applying a theoretical description to language data, but a cyclic exercise causing the framework to be validated and revised (Leech and Garside, 1991; Wallis and Nelson 1997).

The parsing-and-correction process obtains a *corrected treebank* , where each tree has been comprehensively reviewed by multiple linguists.[11] These trees contain a set of 'situated parsing decisions', i.e. decisions about how best to analyse *this* sentence in *this* context. Errors will exist, but they are less likely to be systematic than errors obtained by deterministic parsing algorithms. This is an important criterion

[10] This estimate is probably optimistic. It depends on scheme complexity (simple schemes are more 'accurately' applied than complex ones) and variation in input sentences. Estimates for computational performance of parsing algorithms are often obtained with limited linguistic data (e.g. computer software manuals or dictionary examples). A broad-coverage ('balanced') corpus, by contrast, can include speech and writing in many contexts.

[11] This exercise requires supervisory 'knowledge management' protocols (Wallis and Nelson 1997).

when it comes to statistical analysis (see 4.5 below): we want to know about the grammar of sentences, not merely the performance of the parser.

Once we have decided to parse a corpus, we next must decide on the annotation scheme and how it should be applied to sentences.

- **Which framework should we choose?** Whereas there is a certain consensus in word class definitions (see Chapter 14), linguists have adopted many different frameworks to describe grammatical relationships.
- **Can we avoid methodological 'over-commitment'**, i.e. that a corpus parsed with Framework X compromises research in Framework Y? The answer must involve*abstraction* , mapping terms in the researcher's Framework Y to terms in Framework X. We discuss a solution in 4.4.4 below.
- **What should be done with missing elements**, such as ellipted subjects, verbless clauses, and incomplete sentences?

The approach taken to the last question depends in part on whether we consider annotation from a 'top-down' or 'bottom-up' perspective. If we believe that a particular theoretical internal language is primary, and missing elements are considered to be performance errors, we might insert 'null elements' (Marcus *et al* . 1993: 321) or words 'recovered from the context'.

In the spirit of being true to the data (bottom-up), corpus linguists have tended not to introduce implied content. Superfluous elements such as self-corrected words and 'slips of the tongue', common in conversation, are often marked out by annotation (cf. struck-out '~~which~~' in Figure 3).

Not all ellipsis involves single words. Natural language, especially conversational language, includes grammatically incomplete utterances as well as complete ones (Chapter 30). Consider this dialogue sequence from ICE-GB.

> B: *Didn't there used to be deer in Richmond Park?*
> A:*There still are.* [S1A-006 #230, 232]

A's 'clause fragment' appears perfectly intelligible to B (and readers). The conversation continues without recapitulation or repair. If we wished to insist on identifying the 'missing' elements of the structure, A's fragment would need further annotation, e.g. by linking to the relevant material in the previous sentence (*deer in Richmond Park*), or inserting *some* .[12]

4.4.4 Grammatical exploration with ICECUP

Tools for working with parsed corpora are relatively rare (see Wallis 2008 for a discussion). The *International Corpus of English Corpus Utility Program* (ICECUP, Nelson *et al.* 2002) is a 'corpus workbench' (a suite of closely connected tools on a single software platform) designed for parsed corpora. It is not the only such tool, but it is arguably the most integrated example of its type.

Central to ICECUP is a query representation called a *Fuzzy Tree Fragment* (FTF): a diagrammatic representation of a grammatical query. It consists of

[12] This issue illuminates a subtle methodological distinction between *speaker parsing* and *hearer parsing* . Conventionally, corpus researchers parse the structure as a model of the speaker's construction process, rather than how the hearer might have reconstructed it.

information about grammatical nodes and/or words, and their possible structural relationships in the tree. A researcher constructs an FTF like the one in Figure 4, and ICECUP uses it to find multiple matching examples in the corpus. Results may be concordanced (Figure 5), so that each example structure is highlighted and aligned visually. If the FTF matches more than once in the same sentence, each matching case is highlighted on a separate line.

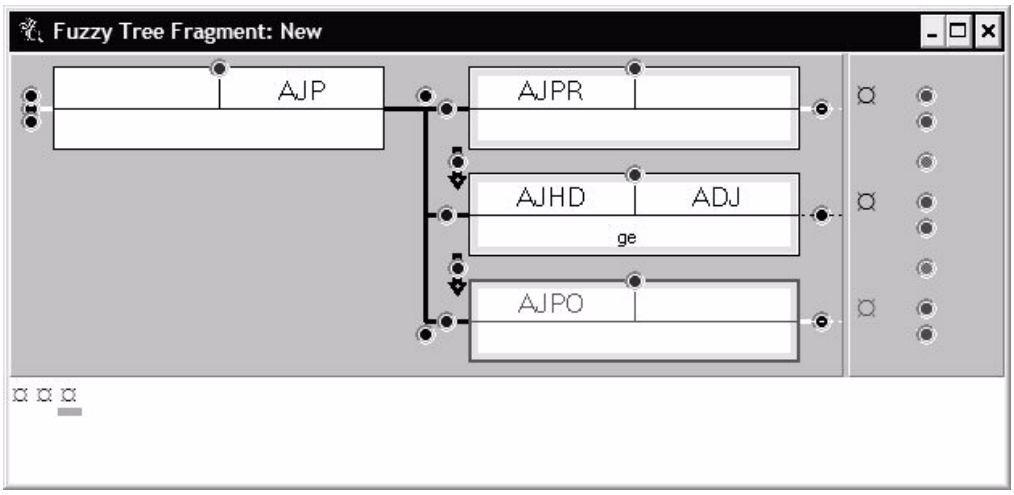

Figure 4: A Fuzzy Tree Fragment for an adjective phrase (AJP) containing a general adjective head (AJHD,ADJ(ge)) with at least one premodifier (AJPR) and one postmodifier (AJPO). Black lines and arrows mean that elements are immediately connected to each other; white lines mean that they are eventually connected. (For reasons of space, trees are typically drawn left-to-right. See also Figure 6.)

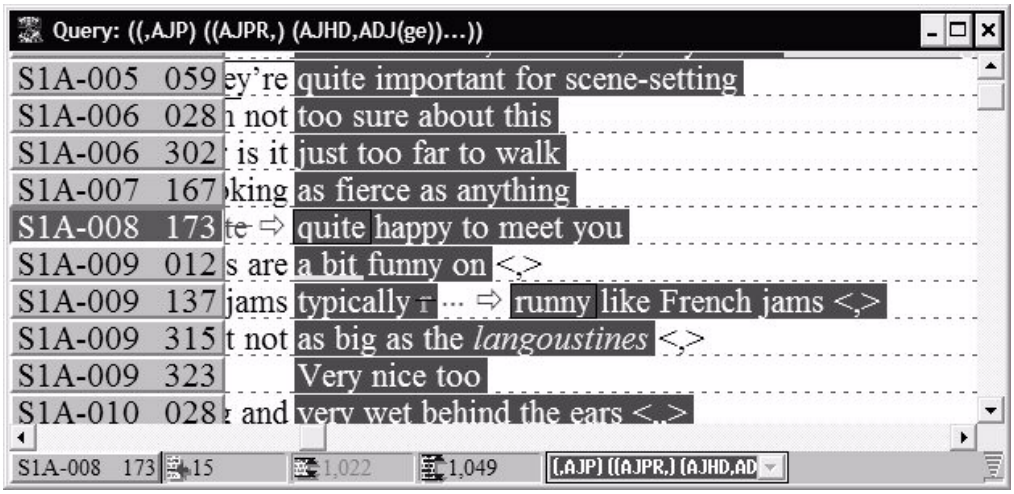

Figure 5: A simple grammatical concordance: highlighted constructions in ICE-GB, such as *quite happy to meet you* , match the FTF in Figure 4.

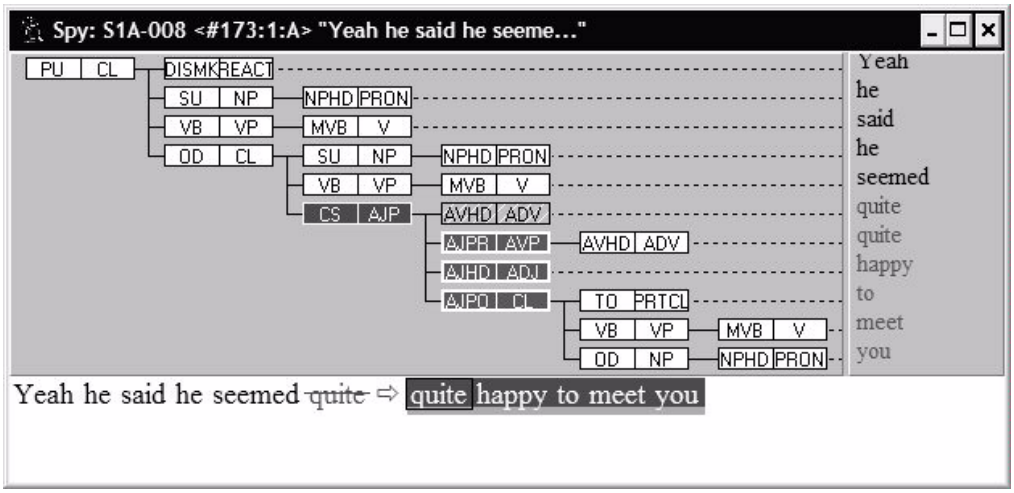

Figure 6: Examining any line in the concordance displays the sentence and phrase structure tree, showing how the FTF matches a tree in the corpus. In this case the adjective phrase premodifier is realised by the adverb phrase *quite*, the head by the adjective *happy* , and the postmodifier by the subordinate clause *to meet you* .[13]

The benefits of a visual representation of a 'tree fragment' (rather than, say, a logical expression) are threefold. First, the FTF appears as a coherent, intuitive whole. Second, the tree metaphor helps us identify how it matches part of a particular tree and sentence, helping researchers refine their queries. See Figure 6. Third, users can build an FTF from a tree. They select nodes they want the FTF to match, and make some simple choices in a 'Wizard' tool to form the FTF.

This combination – a visual representation that maps from query to tree, a user interface that makes search results easy to browse, plus algorithms working in the background – produces an exploratory research platform that allows researchers to explore the rich parse annotation in the corpus and carry out experiments.

What about the risk of 'over-commitment' noted earlier? Are ICE-GB users committed to the Quirk *et al* . (1985) grammar? Must they learn every last detail?

ICECUP addresses this problem by an *exploration cycle* (Nelson *et al.* 2002: 85). Users often begin with a text search or by browsing part of the corpus; find a construction of the type they are interested in; and then use the tree to create an initial FTF. The linguist does not have to 'agree' with the grammatical framework used: the parse analysis is simply 'a handle on the data' (Wallis 2014).

Strictly, this exploration cycle combines the bottom pair of cycles in Figure 2 (abstraction and annotation). In practice, exploration often consists of the abstraction cycle, i.e. annotation is taken as given. Occasionally, new annotation may be required. Thus Aarts *et al.* (2013) ultimately subdivided cases of modal verbs *will* and *shall* into semantic subcategories Epistemic and Root. We must be able to add distinctions if required.

[13] Gloss: PU = parse unit, CL = clause, DISMK = discourse marker, REACT = reaction signal, SU = subject, NP = noun phrase, NPHD = NP head, PRON = pronoun, VB = verbal, VP = verb phrase, MVB = main verb, V = verb, OD = direct object, CS = subject complement, AJP = adjective phrase, AJPR = adjective phrase premodifier, AVP = adverb phrase, AVHD = AVP head, ADV = adverb, AJHD = AJP head, ADJ = adjective, AJPO = AJP postmodifier, TO = *to* -particle, PRTCL = particle. Features are not shown for reasons of space.

4.4.5 Bottom-up generalisation algorithms

Many linguists, particularly computational linguists, have used computation to automatically identify common (high-frequency) and statistically 'interesting' (closely associated) patterns in annotated text. These inductive algorithms abstract patterns in data to the point where we might consider whether they represent a new concept we might have missed. Here we briefly discuss some automatic methods relevant to grammarians.

Lexicons. Compiling a lexicon from a corpus requires a computer algorithm to create an index for every word, every combination of word and word class tag, then combinations with morphological stems, stress patterns, etc. This process generates a vast amount of frequency information.

Due to the computational effort involved, indexing is performed first and the results are explored with an interface. For example, ICECUP's lexicon (Nelson *et al* . 2002: 206) makes use of indexes compiled when the corpus was created. The interface lets a researcher limit the lexicon to nouns, words matching a wild card, etc. The interface selects and aggregates terms where necessary. The same approach can be extended to grammatical nodes (ICECUP has a 'grammaticon' tool).

Collocations. The idea of a collocation (see also 4.2.2) is almost as old as corpus linguistics itself, and simply means 'a sequence of words or terms which tend to co-occur more frequently than would be expected by chance' (interaction evidence). Popular tools include *AntConc* (Anthony 2005) and *WordSmith* (Scott 2012). Collocations have obvious value to language learners or researchers in semantics, and the list of examples obtained from a corpus is much broader than published idiom lists. Using business or legal English corpora, collocation can generate domain-specific lists.[14]

Pairs of words may collocate for different reasons. They may represent a specific concept whose meaning is idiomatically given by the entire string (e.g. *small businessman*). Collocation did not originally depend on grammatical distinctions: rather, the method might reveal grammatical patterning in the text (corpus-driven linguists would say they 'emerge' from the corpus). Typical two-word collocations adhere to the pattern 'adjective + noun' (e.g. *high street* , *primary school*), or 'verb + particle' (e.g. *found out* , i.e. 'phrasal verbs'), etc.

This type of algorithm may be directed by a search word and a 'slot' to be filled (a gap before or after the search term): for example, to search for collocations of the form 'X + *school* ', or '*found* + X'. A variation of this approach, sometimes termed *colligation* , is a collocation restricted by word class tag, such as 'adjective + *school* ' or '*found* + particle', and operates on a tagged corpus.

N-grams. Stubbs and Barth (2003) proposed a kind of cluster analysis to detect high-frequency sequences of words, termed *n-grams* . A variation of this approach, *phrase frames* , permits some of the word 'slots' to be filled by a particular word. The *Phrases in English* website (http://phrasesinenglish.org) allows researchers to retrieve n-grams from the *British National Corpus* (BNC). The four top trigrams found are:

I don't	36,863 instances
one of the	35,273
the end of	20,998

[14] Various statistics have been proposed for estimating the degree of interaction between words. These include *mutual information* , various statistical measures (z , t , χ^2 and log-likelihood), and *probability difference* (see Gries 2013 for a review).

part of the 16,444

These are frequent, but not very interesting![15]

Stubbs and Barth comment that these n-grams are not (grammatical) linguistic units, but may 'provide evidence which helps the analyst to identify linguistic units'. In other words, results need human interpretation. They may be useful triggers for grammatical insight, but they must be related to a theory.

Collostructions A promising approach that attempts to introduce grammatical concepts into an abstraction process is *collostructional analysis* (Stefanowitsch and Gries 2003). Collocation does not exploit grammatical information except (in the case of colligation) at a word class level. Without structural constraints the method can generate 'false positives' that skew results. In multi-word sequences there is a greater chance of 'false negatives', i.e. cases not found due to the presence of an intermediate word or phrase in sentences.

Collostructions attempt to avoid this. They extend the idea of collocations to a range of constructions,[16] such as [N *waiting to happen*], identifying 'result' arguments of *cause* , and so on. In a corpus annotated for transitivity (usually, a parsed corpus), it is possible to configure a collostruction search which ranks verbs by the extent to which they would be most typically found in the ditransitive construction.

Like collocation, collostruction uses 'attraction' measures (interaction evidence) rather than simple frequency. The authors' algorithm therefore rates potential slot-filling lexemes (e.g. N = *accident* , *disaster* , *earthquake* , etc. for [N *waiting to happen*]) by the probability that they appeared in the specified slot by chance, estimated with Fisher's exact test.[17]

A tool that brings many of these inductive algorithms together is *SketchEngine* (Kilgariff *et al* 2014). SketchEngine is the prince of bottom-up exploratory tools[18] or the ultimate dictionary creator, depending on your perspective. This tool brings many exploratory algorithms together in a single platform, focused around the idea of a 'word sketch'.

Word sketches are a 'one-page summary of a word's grammatical and collocational behaviour'. The core algorithm can be thought of as a super-lexicon entry generator, exploiting automatic lemmatisation, tagging and parsing algorithms. Since no human intervention is employed, misanalysis occurs, hence the method is exploratory.

A verb sketch might include the most frequently co-occurring subjects and objects, co-ordinated terms, phrasal prepositions, etc. Kilgariff *et al.* (2014) use the example of the verb lemma CATCH: in an illustrative corpus, the object identified as the most strongly associated collocate is *glimpse* ; the most frequent, *eye* .[19]

---

[15] This algorithm exploits frequency evidence (the string is frequent) rather than interaction evidence (the string seems to co-occur more than would be expected by chance).

[16] This sense of 'construction' spans grammar and lexicon, i.e. it is not limited to purely grammatical constructions but may include multi-word patterns. The algorithm tests for statistical association. The linguist has to decide whether this is due to shared meaning.

[17] Fisher's test is like a more accurate 2×2 χ^2 test (Wallis 2013). It can compare $p(x) \times p(y)$ and $p(x \wedge y)$ for significant difference. Ranking by error level tends to bias results towards high-frequency phenomena, which is not always desirable.

[18] The term 'sketch' is a deliberate reference to the approximate results these algorithms obtain.

[19] It is hard to think of a more poetic illustration of the difference between interaction and frequency evidence.

Of necessity this brief selection cannot do justice to the full range of computational abstraction methods. All these methods obtain results derived from interaction and frequency evidence, and these results must be interpreted theoretically.

4.5 Experimental corpus linguistics

Experimental corpus linguistics travels in the opposite direction from the inductive methods of the previous section: testing an explicit hypothesis framed by a theory against corpus data. The process includes the formal 'analysis' process (primarily working top-down, but occasionally bottom-up) in Figure 2.

Unlike data from a lab experiment, a corpus is not constructed to specifically test particular hypotheses. We cannot manipulate experimental conditions to collect new data, but must work with existing data. This is sometimes termed a 'natural experiment', or a *post hoc* analysis.[20] In the following sections we will consider briefly the methodological issues raised by two types of corpus experiment.

- **A frequency experiment**. The first type is common in corpus linguistics. It investigates whether the frequency distribution of a lexical or grammatical variable changes over a sociolinguistic contrast. For illustration, we will borrow from Aarts *et al* . (2013), where the sociolinguistic variable is the binary choice: modal *shall* vs. *will* in first person declarative contexts, as in *I will/shall go to the park* . In this diachronic study, the sociolinguistic contrast for comparing frequency distributions is time.[21]
- **An interaction experiment**. The second type concerns the interaction between two lexical or grammatical variables. Some examples are given in Nelson *et al* . (2002: 273ff), and we provide a walk-through below.

Some issues are common to both types of experiment. In this section, 4.5.1 deals with sampling and 4.5.2 the selection of a meaningful baseline. Section 4.5.3 considers interaction experiments.

4.5.1 Sampling

Corpora are sampled as whole or part-texts, according to a sampling frame that specifies numbers of words per text, and numbers of texts in each category.

Most corpora are collected according to the idea of a *balanced sample* , i.e. the distribution of material has similar proportions in each category, or is in rough proportion with the perceived availability of material. A true *random sample* , on the other hand, would populate the corpus from potential sources purely at random. The advantage of the balanced approach is that the corpus collector can aim for sufficient texts in any sub-sub-category of the corpus to permit research in this subdomain.

However, it is difficult to control this sampling process for other variables. Sociolinguists have tended to follow Labov (1966) in recommending *stratified samples* (sometimes called *quota samples*). The aim is to address two issues: including sufficient numbers of participants falling into specific subcategories, and

[20] Other sciences, such as astrophysics or evolutionary biology, work almost exclusively with observational data.

[21] Diachronic studies offer the potential for exploring evolutionary change of grammar including grammaticalization (Traugott and Heine 1991).

avoiding important variables being entangled (e.g. so that male participants do not tend to be older than female ones).

A stratified sample allows us to require that we have examples of, say, women's scientific writing in the 1840s in our corpus. Maintaining strict independent partitions also can help differentiate variables in analysis. However, the more variables we include, the greater the number of combinations to be found and sampled, so 'full' stratification is rarely attainable.

This raises an obvious question, namely: what if there is no data in a given intersection? Or what if it is extremely rare, and to include it would make the corpus unrepresentative? Principles of 'representativeness' and 'inclusion' pull in opposite directions (Wattam 2015).

All corpora embody a compromise, because different research questions impose different requirements on data collection. So a corpus neatly stratified into subcorpora of equal numbers of words may still generate a skewed sample, e.g. for expressions of obligation. We must accept that sampling is likely to be uneven, and try to address this in our analysis methods (Gries 2015).

Finally, unless we intend to investigate phenomena over the length of a text, a corpus containing many short texts is preferable to one with a few long texts. The reason is due to a problem known as *case interaction* (Wallis 2015) – the fact that datasets drawn from texts by queries are not actually a true random sample. Multiple cases of a particular phenomenon may be found in the same text – even the same sentence (Nelson et al. 2002: 272). The more independent texts and participants, therefore, the better.[22]

4.5.2 Baselines and alternation

A common methodological mistake in corpus linguistics arises when baselines are not considered (Wallis forthcoming). Indeed, researchers have traditionally 'normalised' frequencies by quoting rates per word (or per thousand or million words). There are circumstances when this may be reasonable, but it is rarely optimal.

Leech (2003) used the Brown family of corpora to investigate whether *shall* or *will* changed in frequency over time. In both US and British English, *shall* accounts for a smaller proportion of the number of words in the 1990s than in the 1960s data. In British English, modal *will* is almost constant. This method does not seem to support a claim that *shall* is being replaced by *will* , although intuitively this is what we would wish to determine.

We must distinguish two levels of variation – variation in *opportunity* to use a modal auxiliary verb, and variation in the *choice* of modal when that opportunity arises. The optimum baseline identifies these opportunities.[23]

Different baselines permit different research questions:

- If we are interested in the potential to employ a modal verb, we might choose tensed verb phrases. The research question becomes "given a tensed VP, when is a modal employed?"
- If we are concerned with the variation of one modal verb amongst others, we might choose the set of all modals: "given that we use a modal, which do we choose and when?"

[22] Case interaction affects statistical methods, but it also has implications for quasi-statistical algorithms, such as collocation tools. Several researchers have commented on this problem and proposed solutions (Nelson *et al.* 2002, Brezina and Meyerhoff 2014, Gries 2015, Wallis 2015).
[23] See Wallis (forthcoming) for a detailed discussion of how to select baselines.

- In the case of *shall* vs. *will* , the optimum baseline is simply the set of two modals {*shall* , *will* }, optionally extended to include *'ll* (= *will*) and semi-modal BE *going to* .

Defining baselines is not simply about identifying words. Ideally, we should try to limit data to where mutual replacement is possible. In the case of *shall* vs. *will* , this meant restricting data to positive, declarative first-person contexts (Aarts *et al*. 2013). To obtain this data required the use of a parsed corpus and FTFs (section 4.4.4).

In these contexts it was possible to argue that instances of *shall* were able to *alternate* (swap) with *will* (including *'ll*), without changing the surface meaning of the clause.[24] Figure 7 illustrates recent change whereby *shall* became less frequent than explicit *will* in first person declarative contexts in spoken British English at some point in the late 1970s.

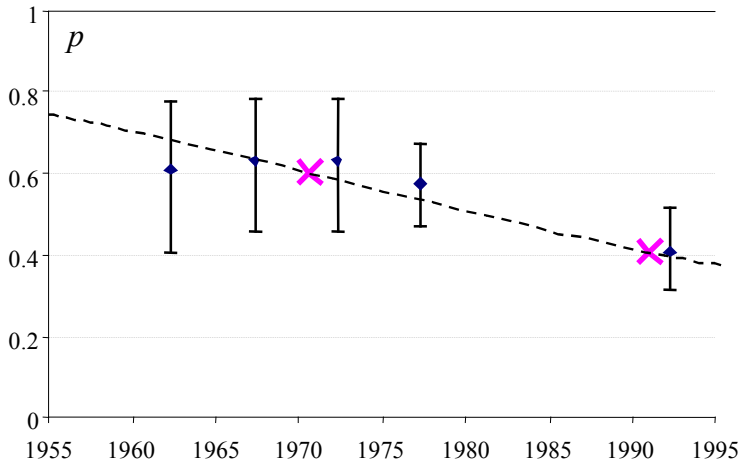

Figure 7: Comparing frequency distributions over different time periods. Declining proportion, *p* , of *shall* out of the set {*shall, will* } with first person positive declarative subjects, half-decade data ('1960' = 1958-62 inclusive, etc.) from the Diachronic Corpus of Present-day Spoken English (after Aarts *et al.* 2013). The crosses represent mid-points of two subcorpora.

Posing research questions in terms of the choices available to speakers is also a fruitful way of exploring how structural constraints in grammar interact.

4.5.3 Interaction experiments

Interaction evidence is straightforward to explore within an experimental paradigm. A simple χ^2 test can be used to explore interaction evidence in a statistically sound way (Wallis 2013). In an interaction experiment, both variables refer to different aspects of the same set of linguistic events (or aspects of two related events); see Nelson *et al.* (2002: 273ff).

Consider how we might evaluate the interaction of the polarity of a question tag ('TAGQ') with the polarity of the preceding verb phrase within a host clause using ICE-GB. Compare:

David turned up did he? <VP = positive, TAGQ = positive>

[24] Interrogative modals were excluded because they may have a different dominant pragmatic reading (cf. *Will we go?* = prediction; *Shall we go?* = suggestion). A similar issue applies to negation.

That's enough isn't it? <VP = positive, TAGQ = negative>

In the corpus, negative tag questions are identified by the presence of the feature 'neg' on the auxiliary verb or main verb. Negative verb phrases inherit the feature from the auxiliary or main verb. We create four similar FTFs using the pattern in Figure 8, allowing the upper node to match auxiliary or main verbs for robustness.

Frequency data drawn from ICE-GB is in Table 1. We can test the interaction between the two locations by applying a 2×2 χ^2 test for independence to the table. The result, as one might predict from the table, is statistically significant. Positive declarative clauses with negative question tags are the most frequent pattern, and there is a large net negative interaction (i.e. polarity consistency is rarer than inconsistency).

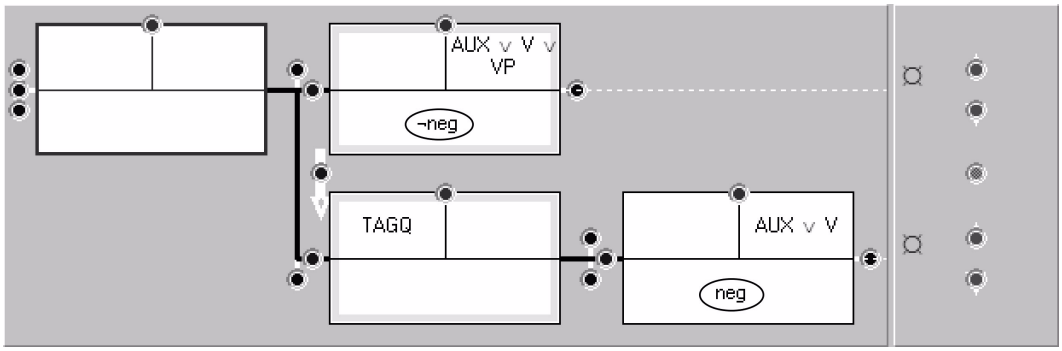

Figure 8. FTF for retrieving a positive ('¬neg' = not negative) auxiliary verb, verb or VP ('∨' = 'or'), followed by a tag question with a negative auxiliary or verb. This FTF is permuted by changing both 'neg' features (circled) to obtain all four patterns.

ICE-GB		VP		
		negative	positive	Total
TAGQ	negative	2	487	489
	positive	58	172	230
	Total	60	659	719

Table 1: Contingency table of frequencies exploring the interaction between the polarity of question tags and the polarity of preceding verb phrases. The verb phrase and tag both have a negative polarity only twice in 719 cases.[25]

Interaction experiments can be extended in a number of interesting ways. Wallis (2012) summarises an experimental design for investigating the chances of speakers/writers making serial decisions of the same type. For example, it turns out that adding an adjective in attributive position before a noun head in a noun phrase becomes progressively less likely ('more difficult') the more adjectives we add. But this observation is not true for all rules. Serially adding postmodifying clauses or adding conjoined clauses after the noun phrase head, first declines and then increases in probability ('becomes easier') with each successive addition.

[25] Jill Bowie points out that reviewing cases, the negative-negative examples are questionable. However, the fact that the cell with two cases may go to zero does not affect the argument.

4.6 Conclusions

With sufficiently flexible tools, a corpus (and a richly annotated parsed corpus in particular) has much to offer the linguist. Parsed corpora are simultaneously valuable and imperfect, so we must engage with evidence critically.

Corpus linguistics is commonly associated with evidence of frequency. Few linguists would dispute that if you wish to know how common a construction might be, you need a corpus. Such distributions can be useful in directing pedagogical material towards the constructions learners might be most exposed to. Or they can focus grammatical research by identifying the patterns that account for the lion's share of the data.

However, there is one type of evidence that corpora can provide that is particularly relevant to the study of the structure of linguistic utterances. This is interaction evidence: probabilistic evidence that two linguistic events tend to co-occur. Such evidence may be simply associative but it is often possible to show a greater size of effect in one direction than another. If we know two events tend to co-occur, we are entitled to enquire about underlying causes. These reasons may range from the trivial fact of a shared topic in e.g. a conversation through to deep cognitive processes constraining the choices that speakers and writers make.

In this perspective, grammar rules specify the choices available, whereas personal preference, language context, semantic and logical reasoning, communicative strategy and cognitive processing combine to influence the choices made.

Linguistic phenomena co-occur for a variety of reasons, not all grammatical, and so this evidence must be carefully considered and alternative explanations explored. Nonetheless, this type of evidence seems to be the most viable option for empirically evaluating grammar as expressed by human beings.

References

Aarts, B. 2001. Corpus linguistics, Chomsky and Fuzzy Tree Fragments. In Mair, C. and Hundt, M. (eds.) 2001. *Corpus linguistics and linguistic theory* . Amsterdam: Rodopi. 5-13.

Aarts, B. Close, J. and Wallis, S.A. 2013. Choices over time: methodological issues in current change, in Aarts, B., Close, J., Leech G. and Wallis S.A. (eds.) *The Verb Phrase in English* . Cambridge: CUP. 14-45.

Anthony, L. 2005. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. In: *Proceedings of IPCC 2005* , 729-737.

Biber, D. and Reppen, R. (eds.) 2015. *The Cambridge Handbook of English Corpus Linguistics* . Cambridge: CUP.

Boehm, B. 1996. A Spiral Model of Software Development and Enhancement. *ACM SIGSOFT Software Engineering Notes* **11**:4, 14-24.

Bowie, J. and Wallis, S.A. 2016. The *to* -infinitival perfect: A study of decline. In Werner, V., Seoane, E., and Suárez-Gómez, C. (eds.) *Re-assessing the Present Perfect* , Topics in English Linguistics (TiEL) 91. Berlin: De Gruyter, 43-94.

Brezina, V. and Meyerhoff, M. 2014. Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* , 19:1, 1-28.

Briscoe, E.J. 1998. Robust parsing. In Cole, R.A., Mariani J. Uszkoreit, H., Zaenen, A., Zue, V. 1998. *Survey of the State of the Art in Human Language Technology* . Cambridge: CUP. Available at: www.cslu.ogi.edu/HLTsurvey

Chomsky, N. 2002. *On Nature and Language* . Cambridge: Cambridge University Press.

Church, K. 2000. Empirical Estimates of Adaptation: The chance of Two *Noriega* 's is closer to p /2 than p^2, *Proceedings of COLING 2000* , 173-179.

Garside, R. and Leech, G. 1991. Running a grammar factory: the production of syntactically analysed corpora or 'treebanks', in Johansson, S. and Stenström A.-B. (eds.), *English Computer Corpora: Selected Papers and Research Guide* , Berlin: Mouton de Gruyter. 15-32.

Gries, S. Th. 2005. Syntactic Priming: a Corpus-based Approach. *Journal of Psycholinguistic Research* 34:4, 365-99.

Gries, S. Th. 2013. 50-something years of work on collocations: What is or should be next… *International Journal of Corpus Linguistics* 18:1, 137–165.

Gries, S. Th. 2015. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10:1. 95-125.

Jespersen, O. 1954. *A Modern English Grammar on Historical Principles* . London: George Allen & Unwin.

Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel V. 2014. *The Sketch Engine: Ten Years On* . Brighton: Lexical Computing. Available at: www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2014.pdf

Kučera, H. and Francis, W.N. 1967, *Computational Analysis of Present-Day American English* , Providence MA: Brown University Press.

Kuhn, T.S. 1977. The Historical Structure of Scientific Discovery, in Kuhn, T.S. *The Essential Tension: Selected Studies in Scientific Tradition and Change* , Chicago: University of Chicago Press. 165-177.

Labov, W. 1966. *The social stratification of English in New York City.* Washington, D.C.: Centre for Applied Linguistics.

Lakatos, I. 1970. *Criticism and the Growth of Knowledge* , New York: Cambridge University Press.

Leech, G. 2003. Modality on the move: The English modal auxiliaries 1961-1992. In Fachinetti, R., Krug, M. and Palmer, F. (eds.) *Modality in contemporary English* . Berlin: Mouton de Gruyter. 223-240.

Marcus, M., Marcinkiewicz, M.A. and Santorini, B. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19:2. 313-330.

McEnery, T. and Hardie, A. 2012. *Corpus linguistics: Method, theory and practice* . Cambridge: CUP.

Nelson, G., Wallis, S.A. and Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English* . Amsterdam: John Benjamins.

OED 1961. *The Oxford English Dictionary* (1st edition, extended). Oxford: OUP.

O'Keefe, A. and McCarthy, M. (eds.) 2012. *The Routledge Handbook of Corpus Linguistics* . London and New York: Routledge.

Putnam, H. 1974. The 'Corroboration' of Scientific Theories, republished in Hacking, I. (ed.) (1981), *Scientific Revolutions* , Oxford Readings in Philosophy, Oxford: OUP. 60-79.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language* . London and New York: Longman.
Rohdenburg, G. 2003. Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers. In Rohdenburg, G. and Mondorf, B. (eds.) *Determinants of grammatical variation in English* . Berlin: Mouton de Gruyter. 205-249.
Scott, M., 2012, *WordSmith Tools version 6* , Stroud: Lexical Analysis Software.
Sinclair, J. 1992. The automatic analysis of corpora. In Svartvik, J. (ed.) *Directions in Corpus Linguistics* , Berlin: Mouton de Gruyter. 379-397.
Sinclair, J., Hanks, P., Fox, G., Moon, R. and Stock, P. and others 1987 (eds.) *Collins Cobuild English Language Dictionary* . London: Collins.
Sinclair, J., Fox, G., Bullon, S., Krishnamurthy, R., Manning, E., Todd, J. and others 1990 (eds.) *Collins Cobuild English Grammar* . London: Collins.
Stefanowitsch, A, and Gries, S. Th. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8:2, 209-243.
Stubbs, M. and Barth, I. 2003. Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language* 10, 1. 65-108.
Svartvik, J. and Quirk, R. *A Corpus of English Conversation* , Lund Studies in English 56. Lund: Lund University Press (1980).
Tognini-Bonelli, T. 2001. *Corpus Linguistics at Work* . Amsterdam: John Benjamins.
Traugott, E.C. and Heine B. 1991 (eds.) *Approaches to grammaticalization* . Typological studies in language, 19. Amsterdam: John Benjamins.
Wallis, S.A. 2007. Annotation, Retrieval and Experimentation. In: Meurman-Solin, A. and Nurmi, A. (eds.) *Annotating Variation and Change* . Helsinki: Varieng, University of Helsinki. Available at: www.helsinki.fi/varieng/series/volumes/01/wallis
Wallis, S.A. 2012. *Capturing patterns of linguistic interaction in a parsed corpus: an insight into the empirical evaluation of grammar?* London: Survey of English Usage. Available at http://corplingstats.wordpress.com/2012/12/04/linguistic-interaction
Wallis, S.A. 2013. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* 20:3, 178-208.
Wallis, S.A. 2014. What might a corpus of parsed spoken data tell us about language? in Veselovská, L. and Janebová, M. (eds.) *Complex Visibles Out There.* Olomouc: Palacký University. 641-662.
Wallis, S.A. 2015. *Adapting random-instance sampling variance estimates and Binomial models for random-text sampling* . London: Survey of English Usage. Available at http://corplingstats.wordpress.com/2015/09/22/adapting-variance
Wallis, S.A. forthcoming. *That vexed problem of choice* . London: Survey of English Usage. Available at https://corplingstats.wordpress.com/2012/03/31/that-vexed-problem-of-choice
Wallis, S.A. and Nelson, G. 1997. Syntactic parsing as a knowledge acquisition problem. *Proceedings of 10th European Knowledge Acquisition Workshop* , Catalonia, Spain, Springer Verlag. 285-300.
Wallis, S.A. and Nelson, G. 2001. Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery* , 5: 307-340.
Wasow, T. 2001. *Postverbal Behavior* , Stanford, CA: CSLI Publications.

Wattam, S.M. 2015. *Technological Advances in Corpus Sampling Methodology* . PhD Thesis. The University of Lancaster. Available at: www.extremetomato.com/cv/papers/thesis.pdf