PART 1. MOTIVATIONS

1. What might corpora tell us about language?

1. Introduction

Corpus linguistics has become popular. Many linguists who would not otherwise consider themselves to be *corpus* linguists have started to apply corpus linguistics *methods* to their linguistic problems, partly due to an increasing availability of corpora and tools. In this chapter we will consider some kinds of research that can be done with corpora, and the types of corpora and methods that might yield useful results.[1] Corpora are also found outside of linguistics, in social sciences and digital humanities.

In this book we argue against a simplistic 'bigger is best' approach to data analysis and for the centrality of *underlying models*, theories of what might be happening linguistically 'behind the scenes', when we carry out research. More data is an advantage, but there is a trade-off between large corpora with limited annotation and small ones with rich annotation. Our perspective relates theory-rich linguistics with corpus linguistics, implying that we need corpora with rich annotation.

However, as corpus linguistics has developed as a discipline, the dominant trend has been to build ever larger lexical corpora with very limited annotation: typically structural annotation (speaker turns, overlaps, sentence breaks in spoken data, formatting in writing), *wordclass* or 'part of speech' tagging (identifying nouns, verbs, and so on) and *lemmas*. Crucially, with large 'mega' corpora, annotation must be automatically produced without human intervention. The multi-billion-word *iWeb* corpus (www.english-corpora.org/iweb), built by Mark Davies from 22 million web pages at the time of writing, is at the frontier of this trend.

Not every linguist is in favour of a methodological 'turn to corpora'. Some theoretical linguists, including Noam Chomsky, have argued that, at best, collections of language data merely provide researchers with examples of actual external linguistic performance of human beings in a given context (see, e.g. Aarts 2001). We refer to this type of evidence as 'factual evidence', see below. From this perspective, corpora do not provide insight into internal language or how it is produced in the human mind. However, Chomsky's position raises questions about *what* data, if any, could be used to evaluate 'deep' theories.[2]

Nevertheless, this contrary position represents a serious challenge to corpus researchers. Is corpus research doomed to investigate surface phenomena? At the end of this chapter, and as a motivation for what follows, we will return to the question of the potential relevance of corpus linguistics for the study of language production by reporting a recent study.

Indeed, in recent years this 'turn to corpora' has begun to influence generative linguists. Take language change: a systematic evaluation of how language has changed over time must rely on data. An old antipathy is replaced by engagement. Large historical corpora such as the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME, Kroch *et al*. 2004) are inspiring a new generation of linguistics researchers to approach corpora in new and more sophisticated ways. Similarly, it is our contention that corpora can benefit psycholinguistics, not as a substitute for laboratory experiments, but as a complementary source of evidence.

What do we mean by 'a corpus'? In the most general sense, corpora are simply collections of language data processed to make them accessible for research purposes. By contrast with experimental datasets, sampled to answer a specific research question, corpora are sampled in a manner that – as far as possible – permits many different types of research question to be posed. Datasets extracted from corpora are not obtained under controlled conditions but under 'naturalistic' or 'ecological' ones. We discuss some implications of this statement in Part 2.

Corpora also typically contain substantive passages of text, rather than, say, a series of random sentences produced by random speakers or writers.[3]

However, most corpora available today have one major drawback for the study of language production. Most data is *written*. Texts are generated by authors at keyboards, on screens or paper. Writing is rarely spontaneously produced, may be edited by others, and is often included in databases due to availability. Like this book, texts are usually written for an imagined audience, in contrast to spoken utterances that are typically produced – scripted performances and monologues aside – on-the-spot for a present and interacting audience.

In the era of the internet, written data is easy to obtain, so large corpora of writing may be rapidly compiled. But if 'language' is sampled from writing (inevitable in historical corpora), we can only draw inferences about written language. Far better to be able to test hypotheses against spontaneously-produced linguistic utterances that are *unmediated*, or, to be more precise, that are minimally affected by processes of articulation and transmission.

Not all corpora are drawn from written sources, and it is not a necessary characteristic of corpus linguistics that limits it to the study of written data. If we had no option but to use written sources, then this would still be better than relying on intuition.

But a better option is a corpus of spoken data, ideally in the form of recordings aligned with orthographic transcriptions. Transcriptions of this kind should record the output word-for-word, including false starts and self-correction, overlapping speech, speaker turns and so on. The transcription should be a coded record of the audio stream. Faithfully transcribed speech data from an uncued and unrehearsed context is arguably the closest source to genuinely 'spontaneous' naturalistic language output as it is possible to find.

A transcription can be richer than a written text. It may be time-aligned with the original audio or video recording, contain prosodic and meta-linguistic information, gestural signals, and so on. The value of these additional layers of annotation will depend on the planned research aims of users. Researchers interested in language production and syntax are less concerned whether transcriptions are time-aligned than whether they are accurate. But if pause length or words per minute is considered a proxy for mental processing then timing data is essential.

Although we refer to 'speech' here, we are really referring to *unmediated spontaneously produced language*, the majority of which will be speech. For example, we might justifiably include sign language corpora under the category of 'speech corpora'. It may be attractive to stretch this definition to include conversational text data (such as online 'chat'), but usually a user interface will allow the language producer to edit utterances as they type. If we wish to study unmediated language production, authentic data from spoken sources seems the best option.

Prioritising speech over writing in linguistics research has other justifications aside from mere spontaneity. The most obvious is historical primacy. Hunter-gatherer societies had an oral tradition long before writing was systematised. When writing developed, it was first limited to scribes, and gradually spread through social development and education. In 1820 around twelve percent of the world's population could read and write. Even today that figure is around eighty-three percent (Roser and Ortiz Espinosa 2018). So the first reason for studying speech is its near-universality. By contrast, historical corpus linguistics – which of necessity can only study written texts prior to the invention of the phonograph – is limited to the language of the literate population of the age, and their region, social class and gender distribution.

There are other important motivations. Child development sees children express themselves through the spoken word before they master putting words on a page, and many writers are aware that their writing requires a more-or-less internal speech act. Which comes first, speech or writing? The answer is speech.

Then there is the question of representativeness. A corpus of British English speech has approximately 2,000 words spoken by participants every quarter of an hour. The author Stephen King (2002) recommends aspiring writers write 1,000 words a day. Allowing for individual variation – and excepting isolated individuals or those physiologically unable to produce speech – it seems likely that human beings produce, and are exposed to, an order of magnitude more speech than writing.

Of course, not all speech data is the same. Speech data may be collected for a variety of purposes, some of which are more representative and 'natural' than others. One of the first treebanks containing spoken data, the Penn Treebank (Marcus *et al*. 1993), included parliamentary language, telephone calls and air traffic control data. Other spoken data might be captured in the laboratory: collected in controlled conditions, but unnatural, potentially psychologically stressed, and not particularly representative.

Scripting and rehearsal are a feature of many text types found in corpora. The *Survey Corpus* (Svartvik 1990) and *International Corpus of English* (ICE, Greenbaum 1996b) include 'scripted speech', i.e. transcriptions of pre-written talks. But even unscripted TV and radio broadcasts may be rehearsed, or subject to multiple 'takes'. The *Corpus of Contemporary American English* (COCA, Davies 2009)

contains transcriptions of unscripted conversations from TV and radio programmes. Not all 'spoken' data is equally spontaneous.

As we noted, historical corpora have a particular problem in this respect. The first known recording of the human voice is that of Thomas Edison on a cylinder phonograph in 1877. The *Corpus of Historical American English* (COHA, 1810-2009, Davies 2012) contains film and play scripts. The *Old Bailey Corpus* (1720-1913, Huber *et al*. 2012) captures the speech of court participants via the court stenographer.

This leads us to one final point. Some transcripts may be produced by non-linguists, such as court and parliamentary transcribers. Court transcripts are expected to be a close record for legal reasons. However, parliamentary transcripts are another matter. Cribb and Rochford (2018) illustrate how the official record is rewritten, glossed, corrected and expanded. Official transcripts should be treated with a degree of caution, and original recordings re-transcribed where possible.

For linguistic research purposes, we are primarily concerned with speech in 'ecological' contexts where speech output is spontaneous, uncued and unrehearsed. An important sub-classification concerns whether the audience is present and participating, i.e. in a monologic or dialogic setting.

The fact that a corpus ideal may be collected away from a lab should not mean that results are incommensurable with laboratory data. On the contrary, corpus data can be a useful complement to controlled 'lab' experiments. But to relate results competently requires some methodological adaptation.

The primary distinction between laboratory and corpus data is as follows. Corpus linguistics is characterised by the multiple reuse of existing data, and the *ex post facto* ('after the fact') analysis of such data. On the other hand, experimental data is obtained under laboratory conditions, where a researcher can manipulate conditions to reduce the impact of potentially confounding variables, for example by requiring each participant to perform the same task.

Corpus linguistics is thus better understood as the methodology of linguistics framed as an observational science (like astronomy, evolutionary biology or geology) rather than an experimental one. So when we refer to 'experiments with corpora' in this book we mean carrying out investigations on previously-collected data. We may select sub-categories in corpora, but if we wish to control how our data was sampled, obtain data in new contexts or cue particular responses, we must collect new data. In Chapter 2, we return to the question of what these types of natural experiment might tell us.

This 'multiple reuse' perspective shapes corpora in another way. Corpora usually contain whole passages and texts, open to multiple levels of description and evaluation. To analyse the discourse structure of a conversation, you need the entire conversation. On the other hand, if you only wish to construct a representative lexicon, random sentences will do. By contrast, laboratory research collects fresh data for every research question, and therefore tends to record data efficiently, containing relevant components of the output decided in advance.

However, the lines between the controlled experiment and the corpus are becoming blurred. Where data must be encoded with a rich annotation (see Section 4) such as a detailed prosodic transcription or parsing, data reuse maximises the benefits of costly data collection. Indeed, many sciences have begun to take data reuse seriously. Medical science has seen *meta-analysis*, where data from multiple experiments are pooled and reanalysed, become mainstream.

Let us adopt a working definition of a spoken corpus as a database of transcribed spoken data, with or without original audio files. What can such a database tell us about language?

2. What might a corpus tell us?

There are essentially three distinct classes of empirical evidence that may be obtained from any linguistic data source, whether this 'corpus' consists of plain text or is richly annotated (see Section 4).[4] These are

1. **Factual** evidence of a linguistic event, i.e. at least one event x is observed, written 'there exists' x, or, in mathematical notation, '$\exists x$'.
2. **Frequency** evidence of a linguistic event, which we can write as '$f(x)$' observed events.
3. **Interaction** evidence between two or more linguistic events, i.e. that the presence of event y in a given relationship to x affects the probability that x will occur, which we might write as '$p(x \mid y)$', 'the probability of x given y'.

Whereas much theoretical linguistic argument concerns statements that particular expressions are or are not possible, the factuality of any theory ultimately depends on real world data.

For example, dictionaries expand by observing new forms. In contemporary British English, *bare* (conventionally an adjective) has gained an informal, intensifying adverb use equivalent to 'a lot', 'very' or 'really', as in *bare money* or *bare good*. An etymological dictionary might similarly grow by finding earlier attestations of a known word meaning.

More controversially, we would argue that for a theoretical linguist to maintain that a particular construction found in a corpus is 'ungrammatical' or 'impossible' is not sufficient. The errant datum deserves explanation. Such an explanation *might* be that it represents a performance error, but this cannot be assumed *a priori*. It could be a rare but legitimate form. So factual evidence might include evidence that appears to contradict existing theories.

Occasionally, the systematic examination of corpora, which is required during annotation (see Section 3 below), uncovers a genuinely novel, previously undocumented linguistic phenomenon. Complementation patterns – patterns of possible objects and complements of verbs – are central to traditional grammars of English. Wallis (2020) relates how the annotators of the *British Component of the International Corpus of English* (ICE-GB, Nelson, Wallis and Aarts 2002) found a complementation pattern absent in their source grammar, Quirk *et al*. (1985). This 'dimonotransitive' pattern (Subject-Verb-Indirect Object, e.g. *he told her*), appears over two hundred times in the corpus. The pattern was too regular to be dismissed as an error. It had to be accounted for, either as a permutation of an existing pattern or as a distinct pattern with its own properties.

Corpora are also a rich source of *frequency evidence* for linguistic phenomena. Much corpus research reports frequencies of linguistic events. Frequency evidence is typically compiled into a *frequency distribution*, a set of related observed frequencies which can be compared.

Frequency evidence has value, even if its meaning is less easy to discern. Knowing that one construction, form or meaning is more common than another has proven beneficial for writers of dictionaries and grammar books, helping them prioritise material for learners. Frequency evidence may be counterintuitive, and theoretical linguists rarely deny corpus data this purpose. But if corpus linguistics only consists of mere counting of words or constructions, how does such evidence relate to the concerns of the theoretician?

One answer involves applying linguistic knowledge to instances (annotation). *Grammaticalization* (Traugott and Heine 1991) is a process whereby over time an erstwhile regular lexical item (or lexical-grammatical form) acquires a new grammatical function and loses lexical meaning as a consequence. Chapter 7 considers data from one such study: the growth of new uses of progressive BE *thinking*. Obviously, distinguishing these new 'grammatical' signifiers requires a careful case-by-case linguistic review.

Frequency data must also be interpreted carefully. A common confusion mixes up *exposure rates*: typically, that an event x appears f times per million words, and *choice rates*: that x is chosen with probability p when the opportunity of using x arises. See Chapter 3.

An exposure rate tells us how often an audience will be exposed to x. Such 'normalised' frequencies are vulnerable to contextual variation (produce a different text and the exposure rate may differ). There are many reasons why a speaker might utter a particular word or construction in a given text, and an elevated or reduced frequency in one context over another may be due to many factors. Exposure rates are not easily capable of comparison with the results from controlled lab experiments and (because they can arise from many factors) difficult to relate to linguistic theories.

A per word frequency measure is obtained by

$$\text{the probability of choosing } x \text{ out of the set of words, } p(x \mid \text{word}) = f(x) \,/\, f(\mathbf{words}),$$

where $f(\mathbf{words})$ is the number of words in the sample.

For example, *bare* (adjective, all meanings) appears 19 times in the (approximately) million words of ICE-GB. We can report that $p(bare{+}\langle ADJ\rangle \mid \text{word}) = 19/1{,}061{,}263 = 0.0000179032$. Since these probabilities are tiny, they tend to be quoted as a multiple of words. The per-million-word rate for *bare* (adjective) in ICE-GB is 17.9032.

A more productive way to frame frequency evidence is in terms of choice rates, i.e. the probability that speakers (or writers) will use a construction when they have the opportunity to do so. The idea is we identify a set of alternative forms, **X**, including the particular form we are interested in, x (so x is a member of **X**, $x \in \mathbf{X}$). This gives us the simple formula

the probability of choosing x out of the set **X**, $p(x \mid \mathbf{X}) = f(x) / f(\mathbf{X})$.

In a lab experiment, the choice rate method is equivalent to cueing a participant with a stimulus triggering the set of choices, **X**, and then categorizing their response (the selection, x). Since language is highly structured, an adjective such as *bare* is unlikely to appear at an arbitrary point in a sentence.

There are some rare examples (e.g. pauses, coughs, swearing) where an expression, x, could conceivably appear before any word in the corpus. In other cases, it is necessary to account for the fact that the potential for the expression is constrained by the rest of the sentence. To study *bare*, we should first identify those locations where *bare* could appear (e.g. in adjective position, possibly constrained by concrete common noun heads). This approach is mathematically more principled and experimentally more revealing. It is compatible with 'the variationist paradigm' or 'alternation studies', which are common practice in sociolinguistics, but less common in corpus linguistics. We return to this question in depth in Chapter 3.

In this book we use the term 'choice' as a shorthand for *the act of selection from a set of permissible expressions*. Considered in this way, every language production process consists of a string of choices, some of which are dependent on preceding choices, some constrain subsequent ones (see Section 7), whereas others are structurally independent.

The principal difficulty of choice-based research is a practical one. The experimenter is not present at the time of the utterance. The response is not deliberately triggered ('cued'). Inevitably the speaker (if they could be interviewed) will not recall what they were thinking at the time! Instead, the researcher must decide, retrospectively reviewing utterances, when the opportunity would have arisen.

Reliably identifying locations where choices may arise is not always straightforward. Sometimes the choice is between two words or phases, such as a choice of modal *shall* or *will*, in which case we can simply pool both options. However sometimes the choice we are interested in is one of omission, such as the choice between relative and zero-relative clauses (*he thought [that]…*). Annotation may help identify these 'counterfactual' cases. For example, the ICE-GB parsing scheme contains a 'zrel' feature identifying zero-relative clauses, without which the task would be more difficult.

A further issue concerns whether the overall meaning is allowed to change as a result of the choice being made, or whether it is sufficient to simply determine that the choice is *available*.[5]

Traditionally, corpus linguistics has tended to focus on exposure rates. Many books on corpus linguistics methodology assume that citation of frequencies per million words is the norm. It should not be surprising, therefore, that some corpus linguists have expressed unease at the choice-based paradigm, with perhaps the most common argument being that the choice appears to be arbitrary.

Intermediate positions between hearer exposure and speaker choice are also possible. It is legitimate to survey, for example, the changing distribution of modal auxiliary lemmas as a comparative exercise, i.e. whether *can* or *will* are increasing as a proportion of all modals over time, without claiming that they are mutually substitutable. A crucial skill for a corpus linguist is to recognise these different kinds of frequency or probability evidence, and to properly report their implications.

The final class of evidence that can be gained from a corpus is *interaction evidence*. This is evidence concerning the effect of choosing one word, construction or utterance on other choices in the same linguistic vicinity or given relationship. To take a trivial example, if a speaker starts by saying '*I…*' a hearer will intuit the most likely next word will be a verb. Interaction evidence is core to computer algorithms such as automatic wordclass taggers and parsers, but is often overlooked as an important method for higher-level linguistic research.

Interaction evidence is best obtained from choice rates. If we can identify the probability of a speaker employing a construction when they have that option, we can also identify the effect of a co-occurring construction on that probability. See Section 5 below.

3. The 3A cycle

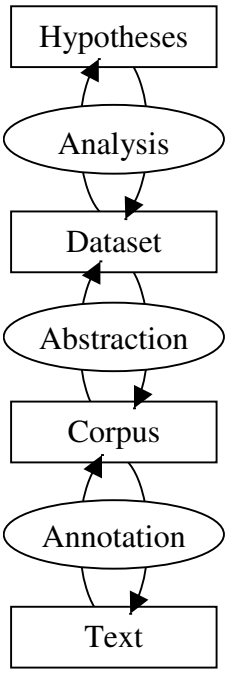

Figure 1. The 3A perspective in corpus linguistics (after Wallis and Nelson 2001)

3.1 Annotation, abstraction and analysis

Our second observation about corpus linguistics is that all traditions within corpus linguistics and related fields (such as applying corpus methods to sociolinguistic interview data) can be conceived of as consisting of three cyclic processes. These are *annotation*, *abstraction* and *analysis*. Each cycle operates between two distinct levels of knowledge, so there are four levels in all. This way of thinking about corpus linguistics, which we call the '3A perspective' (Wallis and Nelson 2001), is sketched in Figure 1.

The idea is that each process adds knowledge to one level to transform it into the next, in a cycle of extension and critical reflection. Knowledge is both *necessary* to each stage and *refutable*. It is applied at every level, from sampling decisions to research hypotheses.

For example, when we annotate a text we add information to it – such as sentence boundaries and wordclass tags – and we critically review the annotation scheme we are using. Annotation is sometimes referred to as 'Qualitative Analysis', especially at an early pilot stage, as a scheme is developed. But whereas Qualitative Analysis might be applied to a small number of texts or selected sentences, an annotation scheme should be applied systematically across an entire corpus (Wallis 2007).

Every qualitative annotation decision must be considered carefully. Is it useful to have a concept such as a 'sentence boundary' in spoken data? What set of wordclass tags should we use, and which distinctions should we capture? Does *this* word have that tag? Should the scheme be modified if it does not adequately describe a phenomenon we discover while annotating, like the previously-undescribed 'dimonotransitive' complementation pattern?

In the case of spoken data, the ultimate source is not text but an audio waveform, and so 'annotation' includes the transcription process. Annotation may be of any conceivable system of linguistic analysis: syntactic, semantic, pragmatic, prosodic or morphological.

Both annotations to the text and the annotation scheme itself may change over the course of annotating an entire corpus. The more initially tentative and experimental the annotation scheme, the more likely it will develop during an annotation phase. Obtaining complete coverage of a scheme across a corpus often raises unanticipated challenges when faced with new phenomena.

In corpus linguistics, the annotation cycle is typically, although not always, performed by corpus builders. Some research teams have added annotation to data compiled by others, or a team might extend their annotation in a series of phases, each with their own release.[6]

Traditionally corpus linguistics practice tends to place a sharp line between annotation and abstraction. Annotation conventionally ends with the distributed corpus. However, as we shall see in Section 5, during a study, researchers may perform additional annotation steps to manually classify data with new criteria.

That said, abstraction begins the process of 'research proper'. It is the process whereby a linguist takes a corpus and attempts to obtain ('abstract') examples of phenomena of interest. If a corpus is richly annotated, they may exploit the annotation.

'Abstraction' can be as simple as identifying single examples for illustrative purposes. However, for empirical research, it must be *systematic*, i.e. our task is to find every example of a phenomenon precisely and exhaustively. Abstraction can be performed top-down, driven by frameworks and hypotheses in the mind of the researcher. It can also be performed bottom-up, with what are often called 'exploration tools': *collocation* and *colligation* algorithms, *n-grams* and other tools deploying associative statistics on the plain text or existing annotation. See Wallis (2020) for a review. The combination of abstraction and annotation with research tools may be unified in an 'Exploration' cycle (Figure 2).

Exploration can, as the name suggests, be tentative and partial. It allows researchers to gain a greater understanding of viable definitions that might be worth pursuing. However, for a deeper analysis, abstraction must be systematically applied to create a sample dataset consisting of instances classified by multiple *variables*. Systematic abstraction of this kind is sometimes termed 'data transformation' or *re-representation* in the field of Knowledge Discovery, and *operationalisation* in Experimental Design and Statistics textbooks.

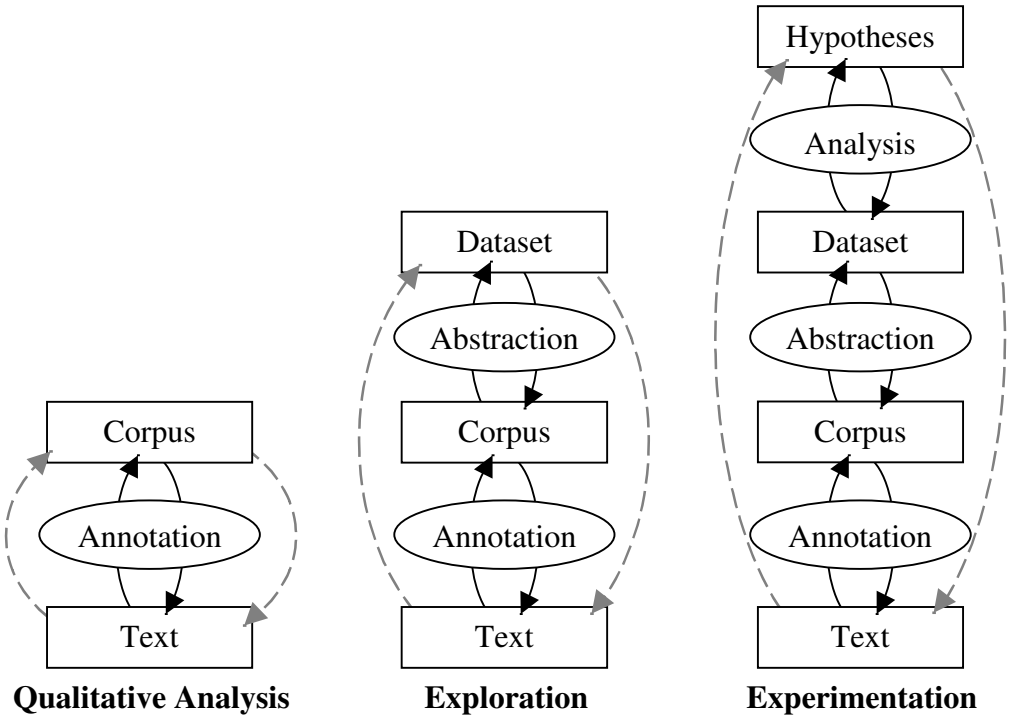

Figure 2. Qualitative Analysis, Exploration and Experimentation in the 3A model.

At the top of this 3A series of cycles is the analysis cycle. Analysis consists of working with an abstracted dataset to obtain high level generalisations, test hypotheses, etc. It is the principal focus of this book. As we shall see in Chapter 2 when we turn to discussing experimental design, we often work top-down from analysis: starting with hypotheses, we identify the concepts they rely on, and try to work out how we might instantiate them from our corpus (i.e. perform abstraction). But it is also possible to work bottom-up, drawing on observations in the corpus data to formulate new variables or reformulate existing definitions.

Taken together, all three cycles may be considered as forming a single process: the 'Experimentation' cycle, as shown in Figure 2.

Abstraction translates from one framework to another. It selects data from an annotated corpus and maps it to a regular dataset for the purposes of statistical analysis. A corpus query system is the principal tool for this process. When a query is performed, the researcher obtains a set of matching results, including the total frequency count.

Suppose we wished to obtain a set of verb phrases from a corpus. In a corpus already parsed with a phrase structure grammar the task is simple: verb phrases are part of the annotation scheme, and we perform a query, e.g. 'VP'. However, most corpora are only tagged for wordclass category (noun, verb, auxiliary verb, etc.) plus additional features.

Identifying the verb phrase itself – where it starts and where it ends – in such a corpus is more difficult, but we can obtain a count of main verbs ('V' in ICE-GB). The method is imperfect, and there are exceptions: verb phrases consisting of lone auxiliary verbs (e.g. *Yeah I <u>will</u>* (S1A-002 #137)) and interrogative verbs considered not to be part of a VP (*<u>is</u> it important?* (S1A-003 #18). However, examining the ratio V:VP across subcategories of speech and writing in ICE-GB finds that counting verbs as a proxy for verb phrases is accurate to within a percentage point.

But if we want to examine complements of the verb, an unparsed corpus creates further challenges. How may we reliably identify indirect objects, for example?

It follows that abstraction is more powerful if the 'heavy lifting' has already been performed in the annotation scheme. But this raises a sensible concern. Are we now *dependent* on the annotation? If another team created and applied one framework, am I, a researcher working in a different framework, now committed to the one embodied in the corpus?

3.2 The problem of representational plurality

A crucial problem – and a standard objection to richly annotating a corpus – concerns 'representational plurality'. Every student knows linguistics is full of competing frameworks. But to annotate a corpus systematically, it follows that the annotator must commit to one framework. The conceptual framework of the linguist researcher 'end user' may be quite distinct from that framework. Indeed, as they work with the corpus, even a researcher taking the given framework as their starting point may find refinements necessary.

Whether the original framework matters to the researcher *ultimately depends on the success of abstraction*. If they can reliably obtain examples of the phenomenon of interest, then how the corpus was originally annotated is immaterial. We refer to this perspective as treating annotation as a 'handle on the data' (see Section 4 below), whereby annotation is simply considered in terms of its value for obtaining relevant examples. Abstraction is the key step.

In any given field of research, linguists differ in their ideal representation scheme, and schemes are often in a state of development themselves. Schemes frequently differ terminologically, but much more importantly, they differ in their classification and structuring of linguistic phenomena, what they include and exclude.

In wordclass analysis, CLAWS7 differs from CLAWS5, and both differ from TOSCA/ICE. In parsing, Quirk *et al.* (1985) exclude objects from verb phrase analysis; Huddleston and Pullum (2002) include them. Constraint grammars represent verb phrases another way, and so on. After two decades of corpus parsing, we have a range of corpora attempting to capture comparable linguistic phenomena with different schemes. Anne Abeille's book *Treebanks* (2003) contains articles describing at least ten different frameworks applied to substantial corpora.

Any linguist who uses a corpus translates concepts from the annotated corpus to their preferred framework. 'Abstraction' is the elaboration of a researcher-defined set of 'translation rules' that converts terms and structures in these different schemes into concepts relevant to the researcher's framework and hypotheses.

Suppose a researcher is investigating noun phrase (NP) complexity. The definition of a more-or-less 'complex' NP will vary according to theories of complexity and processing. (Indeed, the researcher will probably wish to consider multiple definitions.) In a parsed corpus, we can identify e.g. NP postmodifying clauses, as in *the car I used to drive*.

Every instance in such a corpus contains considerable detail: the words used, their wordclasses, and their grammatical functions and structural relationships. The researcher needs to work out rules and queries that 'map' these structures to their chosen definition of complexity. A NP with no postmodifying clauses might have complexity zero, one with a single postmodifying clause might have a score of 1, and so on. But complexity may be conceptualised differently, focusing on other aspects of the NP, such as the type of head or number of adjectives. See Section 6. This does not mean that one definition is 'right' and another 'wrong'. Researchers may merely wish to investigate 'complexity' defined in different ways.

The centrality of abstraction in corpus linguistics is frequently overlooked. But it is a necessary step whereby a researcher reframes the data to their research requirements. Supporting this process is crucial to the design of software tools for working with corpora. Whenever you perform a query against a corpus using software, you need to know that the results you obtain are reliable. Finding examples may be easy. Finding *all relevant examples* can be difficult.

Consequently, the software must let you check queries to determine how it performs. It is not difficult for a query tool to list the cases it finds. The researcher can review these to check whether the cases are correctly included (termed 'true positives') and spot those that should not have been included

('false positives'). But they also need to know that *the search has not omitted cases that should have been included* ('false negatives'). There is, by definition, no simple answer to the latter problem – if we could automatically find falsely-omitted cases we would include them!

We can, however, build platforms that allow researchers to explore the corpus with many different queries, and thereby attempt to spot false negatives by other means. Normally we review examples found and then consider whether cases might have inadvertently been excluded due to variations in the annotation or for other reasons. It is often possible to selectively relax constraints in a query and err on the side of finding more false positives and previously-excluded cases, and then eliminate the false positives by review. See Chapter 16.

In summary, the more ambitious the research question, and the richer and more detailed the annotation scheme, the greater the need for researchers to revisit source sentences to ensure that their abstracted dataset is reliable. Tagged corpora offer limited search options, and may not require an extensive cyclic process of query refinement. But with richly-annotated corpora, a researcher may need to try out a range of different queries. Thanks to the diversity of frameworks, it is quite probable that a researcher will wish to use a different conceptual framework from the annotators (even if some of the labels are the same).

Annotation should never be taken on trust. It may contain errors and biases, particularly if it has been automatically applied but not checked by linguists.

3.3 ICECUP: a platform for treebank research

Corpus linguists clearly need effective software tools to engage with annotated corpora. The *International Corpus of English Corpus Utility Program* Version 3 (ICECUP III) research platform (Nelson, Wallis and Aarts 2002, see Figure 3) was designed around the abstraction cycle to support research with a parsed corpus: initially, the million-word ICE-GB, 60% of which consists of transcribed speech.

The main query system is a diagrammatic query representation that mirrors the visual appearance of parse trees in the corpus: *Fuzzy Tree Fragments* or 'FTFs'. An FTF is a tree-like query where nodes, words, and links between nodes and words may be left partially specified. At the top right of Figure 3 we have an FTF that searches for structures consisting of subject complement clauses ('CS,CL') containing a subordinator phrase ('SUB,SUBP') followed by an adverbial (adjunct) clause ('A,CL').[7] When a query is applied, the set of matching cases are immediately presented by the interface (middle right). Researchers can review how queries have been matched to the corpus and identify false positives.

One advantage of a tree-like query system is that it is easy to see how elements of the query map directly onto the tree (bottom right). We can see why and how ICECUP determined that this was a matching example. The reverse – abstracting from an example tree to a query – is also possible. A 'Wizard' tool permits a researcher to select parts of the tree annotation and convert it into an FTF.

The tools are linked by a user interface that sits on top of a specialised database system. Each window in Figure 3 depicts a different tool, and the arrows show how corpus exploration is typically carried out. Users may identify a text from the Corpus Map (top left) and, by browsing the text, an individual sentence tree (bottom left). The Wizard tool allows the researcher to select part of this tree and create an FTF query (top right). This query can then be applied to the corpus, and the matching elements in the text unit can be seen in both the query results (middle) and each tree (bottom right).

Figure 3 illustrates how tools relate to one of three levels of generalisation: level 1 are query systems or sets of queries, level 2 consists of query results ('sentences' or matching cases) and level 3 corresponds to individual instances (a sentence plus tree annotation).

The 3A perspective can be applied to many processes not immediately identified as 'corpus' linguistics. Processes of annotation, abstraction and analysis may be usefully employed in numerous automatic 'end-to-end' systems.

Consider a natural language 'understanding' application where human intervention is not possible in real time, and knowledge must be encoded in advance. Natural language processing algorithms are applied to annotate an input stream, such as speech recognition and part-of-speech tagging; particular application features, e.g. combinations of keywords and wordclasses are abstracted; and finally

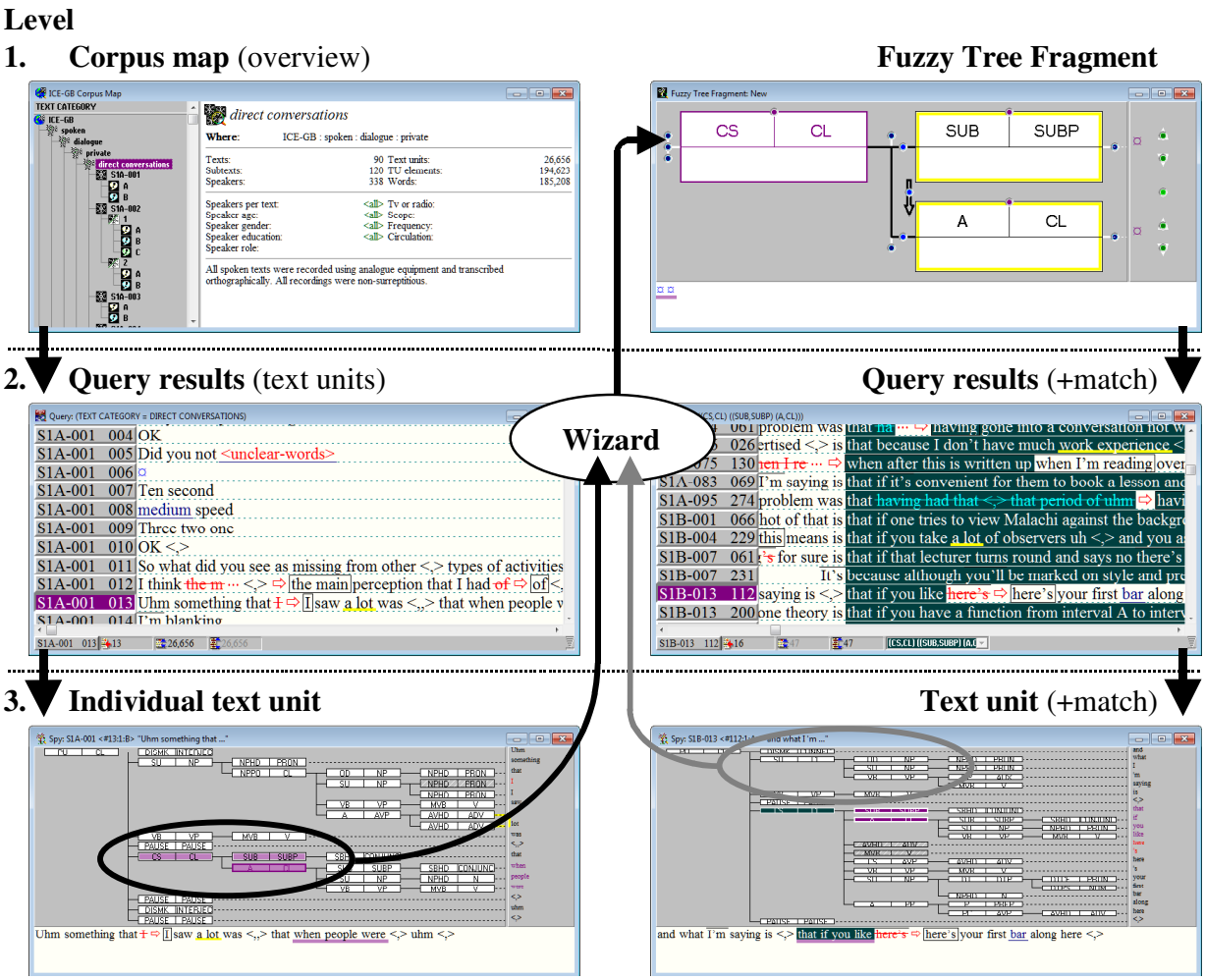

Figure 3. Exploring the ICE-GB corpus with ICECUP, after Nelson *et al.* (2002): from the top, down (left), and using the Wizard in an exploration cycle with FTFs (right). [8]

processed ('analysed') for particular actions. If Google, Langley or GCHQ are listening in, rest assured that their systems are engaged in identifiable processes of annotation, abstraction and analysis!

4. What might a richly annotated corpus tell us?

Let us briefly consider how the three types of evidence identified earlier apply to a *richly annotated* corpus. This is a corpus containing annotation that represents one or more levels of linguistic structure, such as morphological or pragmatic structure. The most common type is a *parsed corpus* (also known as a 'treebank'), i.e. a corpus like ICE-GB or its relation, the *Diachronic Corpus of Present-day Spoken English* (DCPSE).[9]

In a parsed corpus, every sentence is given a tree analysis with a chosen scheme. In the case of spoken data, where 'sentences' must be inferred, decisions to split utterances into sentences are integral to the parsing process, i.e. they are part of the annotation decisions made in applying the scheme to the data.

The notion of a 'linguistic event' identified in general terms in Section 2 is extended to

1. **any single term** in the framework, including the permutation of descriptive features,
2. **any construction** formed of multiple terms in the framework in a given relationship, and
3. **any combination** of the above with elements of the source text.

As multiple levels of annotation are added, this principle applies *between* levels. Thus a corpus consisting of parsed, phonologically and pragmatically annotated text would permit elements to be identified in combination, such as a particular clause structure in a response, a rising tone in a non-interrogative clause, etc.

All three classes of evidence discussed in Section 2, i.e. factual, frequency and interaction evidence, apply to these linguistic events, which we previously denoted by x and y. Thus, using such a corpus we can determine whether a particular construction, formed by a combination of annotated terms, is found in the corpus (x exists, i.e. factual evidence), what its distribution might be (frequency evidence, $f(x)$), and whether the presence of a term increases the likelihood that another, structurally-related term is present (interaction evidence, $p(x \mid y)$).

But if we enrich our corpora by parsing, for example, which scheme should we choose, and why? Cited criteria have included *simplicity* (Penn Treebank I, Marcus *et al*. 1993), *application potential* (e.g. predicate-argument structure for text mining, Penn Treebank II, Marcus *et al*. 1994), and *linguistic tradition* (TOSCA/ICE, based on Quirk *et al.* 1985; Prague Dependency Grammar, Böhmova *et al*. 2003, etc).

Let us consider the question from the perspective of a corpus linguist. There are at least two ways of evaluating a rich corpus annotation scheme (including but not limited to parsing).

- **Annotation facilitates abstraction** ('a handle on the data'). This is a theory-neutral position. The premise is that the annotation scheme simply makes useful distinctions between classes of linguistic event (differentiating nouns and verbs, say) and allows us to retrieve cases reliably. From this perspective, it is not necessary for a researcher to 'agree' to the framework employed. Distinctions encoded in the scheme must only be sufficient for research goals. The actual annotation scheme is irrelevant if the researcher can reliably abstract data according to their experimental paradigm.
- **Annotation facilitates analysis**. This is a theory-integral position related to the concept of *explanatory power*. Annotation should be considered according to its potential to progress research goals. For example, models of priming and spreading activation imply that decisions made by speakers and writers are influenced probabilistically by previous decisions. An annotation scheme that enables evidence of this kind to be found reliably is more powerful than one that does not. An ideal annotation scheme for psycholinguistic research could be one that reflected a credible 'trace' of the language production process undergone by the speaker.

In the first perspective, potential annotation schemes are evaluated by their ability to *reliably retrieve* linguistic events (Wallis 2008).[10] This seems intuitive. We can say that a corpus whose annotation reliably classifies nouns and verbs is better than one with an unreliable classification, and a representation that explicitly denotes subjects of clauses is preferable to one that does not.

However, this criterion is rather circular! Why should we assume, *a priori*, that reliable retrieval of subjects or nouns is important? It also admits redundancy, because any representation can improve on another by simply adding levels and becoming more complex.

The second position builds on the atomised linguistic event retrieval perspective of the first. True, it is useful for linguistic events to be reliably identified. However, for psycholinguistic research goals, it is the ability to obtain interaction evidence that has a plausible *linguistic* cause that ultimately justifies decisions regarding annotation scheme design.

If event y and event x correlate together in their co-occurrence, and we can eliminate trivial explanations of this correlation (e.g. textual topic or contextual artifacts), we are left with explanations that are more likely to be essentially psycholinguistic, such as priming or spreading activation. Of course, other research goals may prioritise other distinctions.

The argument that linguistic annotation schemes should ultimately be evaluated by their explanatory power (i.e., their ability to provide evidence for theoretically-motivated goals) is consistent with Lakatos's (1978) epistemology of *research programmes*. This philosophical perspective views science as pluralistic competition between research programmes. Successful research programmes make novel predictions that can be tested. Declining ones are unproductive: for example, they fail to explain phenomena competing programmes address.

Annotation schemes are part of the *auxiliary assumptions* of the research programme (Wallis 2020). From this perspective, the annotation scheme cannot be evaluated in the abstract, but considered in terms of whether it facilitates the end goals of the research programme – and it is the success or otherwise of the programme that ultimately determines the validity of the scheme. The key question is what linguistic

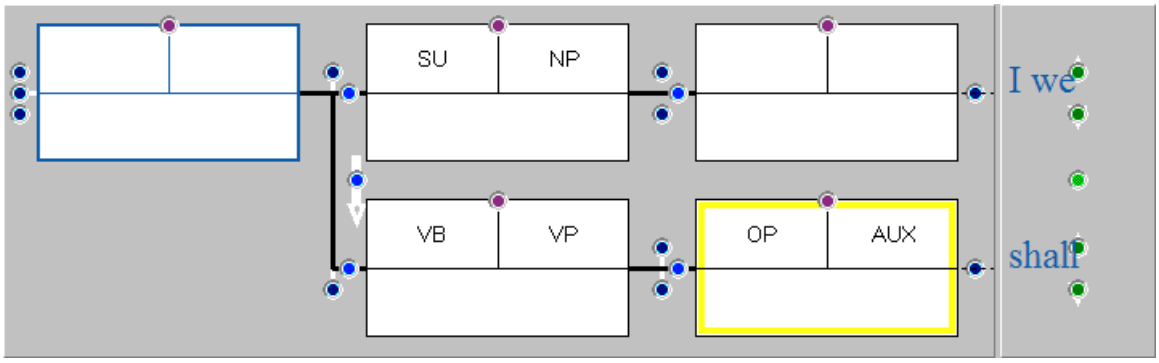

Figure 4. An FTF for a first person subject (*I* or *we*) followed by auxiliary verb *shall*, after Aarts *et al.* (2013). To search for *will* and *'ll* the lexical item *shall* is replaced.[11]

research goals could annotation schemes attempt to further? We will attempt an initial answer in Section 6, but first let us consider research of the first kind.

5. External influences: modal *shall / will* over time

Many corpus studies investigate external influences on linguistic choices. Aarts *et al*. (2013) looked at whether the alternation between the modal auxiliaries *shall* and *will* changed with time in first person positive declarative contexts.

The *shall / will* alternation is a striking example of change over two centuries. 200 years ago, writers would use *shall* to express future prediction or intention. *Will* was expressly criticised in prescriptive grammars. Yet present day British English finds *shall* rarely used (it is the archaic marked form) and *will* outnumbers *shall*.

Aarts *et al*. used an alternation methodology that employed grammatical restrictions to focus on a subset of cases. Christian Mair and Geoff Leech (2006) considered *shall* and *will* (including negative *shan't* and *won't* and cliticised *'ll* = *will*), and analysed each modal auxiliary verb in terms of exposure rates (*shall* and *will* per million words). Using their results, it is not possible to refute an alternate hypothesis that one or either trend was due to a varying potential to use either *shall* or *will*. Nor is it possible to infer the true rate over time. Yet it is a relatively simple matter of reframing their data to pose the question in terms of a basic choice rate, *shall* as a proportion of the set {*shall*, *will*}. This is exactly what Aarts *et al*. proceeded to do.

Mair and Leech had also not evaluated whether *will* was replacing *shall*, although this was an implication of their study. To do this, it is necessary to go back to source texts. Can cases of *shall* or *will* be replaced by the other modal without changing the intended meaning? If the answer is no, these 'non-alternating cases' should be removed. See Chapter 3.

These earlier studies were carried out on tagged corpora. This made it difficult to distinguish between distinct grammatical contexts where alternation is licensed to varying degrees. For this alternation, the subject matters. By the 1960s, examples of *you shall…* or *they shall…* had become rare. Alternation of *shall* and *will* rarely occur except with first person subjects. The ideal would be to identify just those cases of *shall* where the speaker has a genuine choice of using *will* instead, and vice versa.

Consider the interrogative: *Shall we go to the park?* and *Will we go to the park?* are semantically and pragmatically distinct, so do not freely alternate. Aarts and colleagues focused on first person declarative cases, and, for similar reasons, decided to eliminate negative cases.

Working with the parsed DCPSE and ICECUP, to reliably extract cases of first person declarative positive *shall* and *will* one can construct an FTF query like Figure 4. Another FTF pattern, where *shall* or *will* is followed by *not*, is applied to identify negative cases, which are then removed from the total.

This FTF works on the annotation scheme by relating individual terms and structure. It is a reliable retrieval mechanism for obtaining relevant cases – as reliable as the annotation is consistently applied, at least. The annotation is a 'handle on the data' allowing researchers to extract instances of linguistic events, in this case the use of *shall* or *will* in the context required. Graphs such as the one in Figure 5 are obtained, showing the tendency to prefer *shall* over *will*, in first person declarative positive contexts, falling over time.

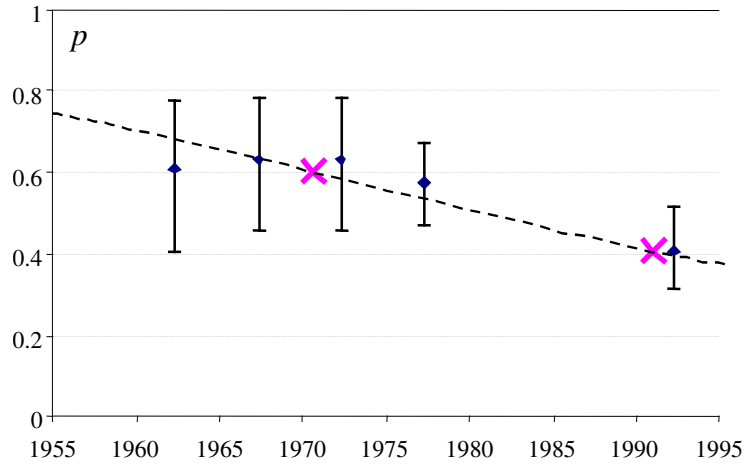

Figure 5. Declining use of *shall* as a proportion *p* of the set {*shall, will*} with first person subjects, half-decade data ('1960' = 1958-62 inclusive, etc.) in the spoken DCPSE corpus (after Aarts *et al.* 2013). The crosses are midpoints of the two DCPSE subcorpora.

Consider the steps that would be required to obtain these results were DCPSE unparsed. It would be possible to construct queries that searched for patterns of a first person pronoun followed by *shall* or *will*, but researchers would then have to manually review every instance to verify it was part of the same clause. In effect, they would be performing the necessary additional annotation stage during their research. *Annotation is unavoidable.*

In this study, one of the authors manually classified every instance of previously-identified cases of *shall* and *will* by their modal semantics (Epistemic, Root and 'other'). This allowed her to conclude that the identified fall in an overall preference for *shall* was due to a sharp decline in Epistemic uses of *shall* (i.e. those with a meaning of intention rather than prediction). This step is also a type of annotation, except one performed by a researcher for a research goal, instead of by the corpus compilers prior to distribution.

6. Interacting grammatical decisions: NP premodification

The previous study examined variation in a single linguistic choice over time. Other variables external to the text (commonly called 'sociolinguistic' variables as a shorthand), such as speaker gender, text genre, mode of delivery, monologue versus dialogue, etc. fall within this experimental paradigm.

However, if we are interested in internal influences on language choices, we must extract and attempt to interpret interaction evidence. Interaction evidence may simply consist of exploring the correlation between two grammatical variables. Examples given by Nelson *et al.* (2002: 273-283) include the interaction between transitivity and mood features of clauses, and the phrasal marking of an adverb and that applying to a following preposition within the same clause. See also Chapter 3.

Below we briefly summarise recent research that examines a more general phenomenon, i.e. serial repeated additive decisions applied in language production (Wallis 2019b). This paradigm considers the decision to add a particular construction to a base construction, and tests whether the speaker or writer is more or less likely to repeat the decision. Here we briefly introduce the idea.

This methodology may be seen as a way of examining *construction complexity* (a static interpretation) or *the interaction between language production decisions* (a dynamic one). The patterns obtained are interesting, occasionally counterintuitive, and certainly worthy of theoretical discussion. An advantage of the dynamic interpretation is that it draws our attention to the process of construction itself rather than merely the outcome of the process.

A simple illustrative example is attributive adjective phrases premodifying a NP head, thus we have *boat*, *green boat*, *tall green boat*, etc. Armed with a parsed corpus we can use FTFs to identify NPs with a head noun, a subset of these NPs with at least one attributive adjective phrase, a subset of those with at least two adjective phrases, and so on.

This model makes no assumptions about the order of decisions. Adjectives might be selected in word order (*tall*, *green*), in reverse order (*green*, *tall*) or in parallel, and then assembled according to semantic ordering preferences in a subsequent articulation process.

We obtain the data in Table 1 by applying these FTFs to ICE-GB across both speech and writing. The top line is the frequency, $f(x)$, of noun phrases with at least x attributive adjective phrases, so $f(0)$ is simply all relevant NPs. We can now derive a sequence of probabilities representing the chance that if a speaker or writer has added x–1 adjective phrases, they will add a further one:

$$\textit{additive probability } p(x) \equiv f(x) \,/\, f(x-1).$$

Thus we can see that slightly less than 20% of NPs (19.32%) contain at least one attributive adjective, but less than 8% of these contain two.

We can plot this probability over the number of adjective phrases, x, as in Figure 6. This graph includes 95% Wilson intervals (see Chapter 6) and distinguishes spoken and written performance.[12]

The first point to note about this graph is that the null hypothesis would be that decisions at each level do *not* interact. When we toss a coin repeatedly, the probability of obtaining each individual tail or head is constant. The graph should be flat.

But this is not what happens. Plotting $p(x)$ reveals that the decision to add a second attributive phrase after a first is less probable than the decision to add the first, and so on. By the fourth adjective phrase, we run out of data and obtain wide confidence intervals, but the overall trend seems reliable. Far from decisions being independent, they interact, and do so consistently in a negative feedback loop.

This is not a universal pattern. Adverb phrases premodifying a verb phrase (e.g. *quickly, intelligently, getting to the point*) finds a much weaker interaction between the probability of deciding to add one or two adverb premodifiers. The chance of adding a postmodifying clause to an NP sequentially (e.g. *a thing* [*called a carvery*] [*which had a vast menu*])) first declines and then increases. See Wallis (2019b).

There are at least three potential sources of an interaction between attributive adjectives.

- **logical-semantic constraints**, including
  - attributive ordering (cf. *long green boat* vs *green long boat*),
  - idioms and compounding (*green longboat*), and
  - avoidance of illogical descriptions (*long short boat*);
- **communicative economy**, avoiding unnecessarily long descriptions, especially on the second and third citation (i.e., on subsequent occasions referring simply to *the boat* rather than, say, *the long green boat over there*); and
- **psycholinguistic attention and memory constraints**, where speakers find it more difficult to produce longer constructions.

Tentatively, the most likely explanation seems to be the first. Communicative economy predicts a rapid drop from $p(1)$ to $p(2)$, but not a subsequent fall. Psycholinguistic constraints seem implausible, because the added constructions themselves are 'light' memory-wise. Indeed, if a speaker forgot they had said a previous adjective phrase, it seems more likely they would act in an unconstrained, rather than a constrained manner.

Figure 6 shows that speech and writing do not have the exact same distribution. For example, we can see that a greater proportion of NPs uttered by speakers have no adjective phrases. When they do employ adjective phrases, speakers tend to use fewer phrases, and so on. There may be other possible additional reasons for this, aside from the difference in mode of delivery. For instance, in a conversation, the audience is present and referents require less elaboration. Nonetheless, both subcorpora obtain a similar overall pattern: one of systematic decline.

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *f*(*x*)	193,124	37,307	2,946	155	7
Probability *p*(*x*)		0.1932	0.0789	0.0526	0.0452

Table 1. Frequency and relative additive probability of NPs with *x* attributive adjective phrases before a noun head, in ICE-GB, after Wallis (2019b).

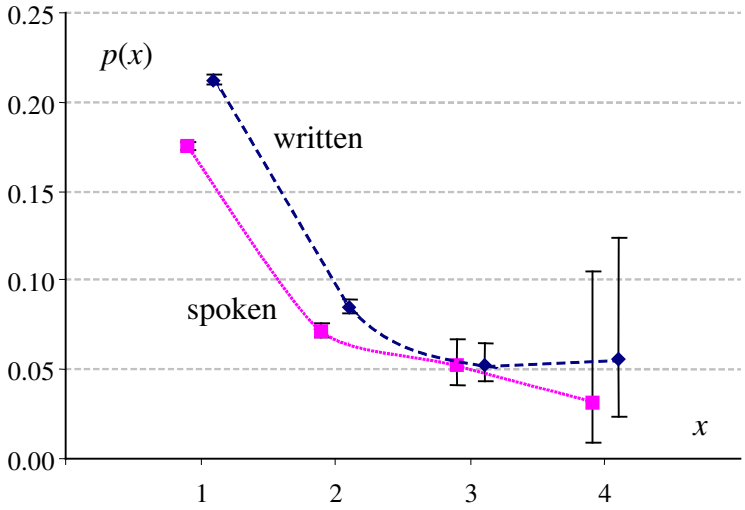

Figure 6. Declining probability of adding attributive adjective phrases to an NP noun head, data from ICE-GB, patterns for speech and writing.

This evidence is only obtainable from a corpus. One would not spot this trend by laboratory experiment: we could not acquire enough data. For NP premodification, employing a parsed corpus is not required, and simple sequences of the form <ADJ> <N> obtain similar results but with a few more errors. On the other hand, to inspect trends generated by serial embedding and postmodification, a parsed corpus is necessary. As soon as we need to examine non-adjacent terms or structure, the reliable representation of that structure is essential.

Finally, comparing spoken and written data is revealing. The majority of corpora contain little or no spoken data. Figure 6 confirms that we are observing a linguistic phenomenon not attributable to a special character of writing or speech, such as a possible tendency for writers to avoid excessive NP length by editing. The presence or otherwise of an audience may affect the rate and starting point of decline, but it does not seem to affect the overall tendency.

7. Framing constraints and interaction evidence

Our interaction experiment illustrates a principle that all such research must contend with. Whenever we carry out corpus research, we explore the effect of two different classes of restriction on the choices speakers make.

Choices affect each other in two distinct ways:

1. **framing constraints** that close off possibilities absolutely, so another choice is unavailable, and
2. **interaction evidence**, where one choice *influences* a subsequent one although the two are structurally independent, i.e. we can say these decisions 'interact'.

Framing constraints are extremely important. If a choice is unavailable, this fact must inform our experimental design. Framing constraints include basic grammar rules, such that violating a constraint can be said to be 'ungrammatical'. This is not to say that these rules are never broken in a corpus. Such violations are indeed potentially worthy of investigation. Rather it means that when they occur, they should be treated as unprincipled 'noise' *for the purposes of a particular experiment*.

In other words, framing constraints are part of our 'auxiliary assumptions' in addition to corpus annotation. They are assumed to be true before an experiment commences. They constrain the options available for the second type, which include those under investigation.[13]

7.1 Framing frequency evidence

In our experimental investigation into variation in the choice *shall* vs. *will* over time (Section 5), we justified focusing an investigation on positive, declarative first person contexts. These contexts constrain the choice of modal *shall* and *will*. Importantly, these contexts are ones where both modals share meanings. As a result, the writer or speaker's intended meaning is not affected by the choice they make.

We can also decide whether or not to include the contracted form of *will* (*'ll*), and if so, how it is incorporated into the experimental design (e.g. as a version of *will*). The process of enumerating available options and grammatical contexts is one of framing the experiment.

But the framing constraints here are ultimately static. It is extremely important that we are aware of them and maintain them across our data. They should apply to all examples in exactly the same way.

If one choice out of several was affected by a framing constraint – for example, it could only appear in a particular construction, or had a much more restricted meaning – this would bias our experiment. We must compare 'like with like'.

We evaluated the declarative alternation for first person *shall* and *will*. Consider instead the interrogative alternation. Table 2 is extracted from the same source as Figure 5 – the spoken DCPSE corpus. On the left, we simply have frequencies of *shall* and *will* tagged as 'interrogative operator'. Examples include

(1) And how long *will* you be away? [DI-A20 #50]
(2) What *shall* we do? [DL-B10 #798]

The subject of the clause turns out to be a crucial factor in the selection. We can use an FTF like the one in Figure 7 to extract interrogative examples followed by a first person pronoun subject *I* or *we*. We can see an obvious difference between how *shall* and *will* tend to be used. See Table 2, right.

Almost all of the cases of *shall* in DCPSE are followed by *I* or *we*, but very few cases of *will* take a first person pronoun as subject. The examples above could conceivably alternate – both *how long <u>shall</u> you be away?* and *what <u>will</u> we do?* are grammatical and plausible – but the meaning changes. A semantic framing constraint applies, and the result is dramatic.

interrogative	*shall*	*will*	**Total**	*shall I/we*	*will I/we*	**Total**
LLC (1957-1972)	43	71	114	41	3	44
ICE-GB (1990-1992)	39	65	104	38	3	41
Total	82	136	218	79	6	85

Table 2. Frequency data for interrogative *shall* vs. *will* in DCPSE, left: all cases, right: cases followed by a first person pronoun subject.

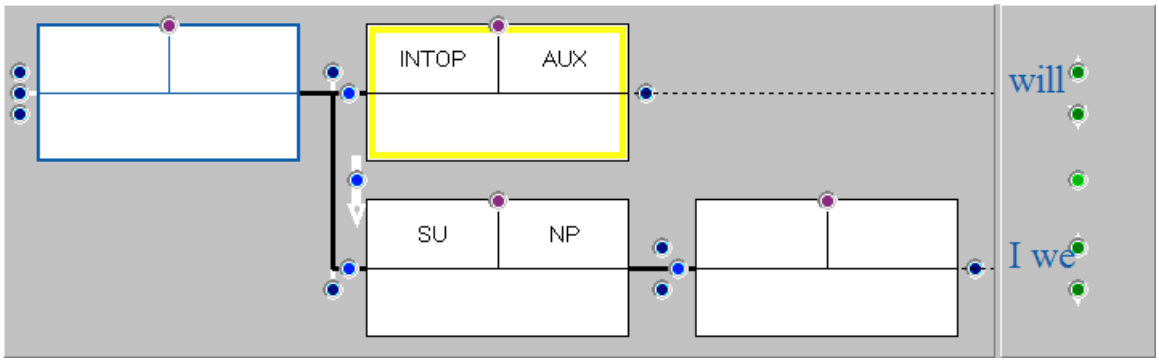

Figure 7. Fuzzy Tree Fragment for interrogative *will* followed by a first person pronoun subject. As before, we substitute *shall* for *will* to obtain the alternate pattern.

Framing constraints matter in obtaining frequency evidence for two main reasons. Firstly, they affect the population of interrogative clauses we may draw conclusions about. But they also matter because *we must ensure that constraints on selection do not differ between alternate types*. See Chapter 3.

Given that the type of subject following an interrogative modal affects the choice *shall* vs. *will*, if we wish to study variation in interrogative *shall* / *will*, we must control for the subject, i.e. we study first person cases independently from all other cases. Failing to do this would conflate two kinds of variation: variation in subject and variation in the target construction, the choice *shall* vs. *will*.

7.2 Framing interaction evidence

Framing constraints are even more important for studies of interaction evidence. This is because we are not simply concerned with the constraints on a single choice, but on multiple choices, and *the relationship between choices*. Interaction research is best considered as a process of investigating how structurally independent decisions influence each other within a set of framing constraints.

In Section 6 we discussed an experiment which measured the degree to which the presence of a number of previous adjective phrases (e.g., *tall*, *green*) before a noun head (e.g., *boat*) affected the chance that another adjective phrase would be added. This process is framed by the rules of English grammar that permit, *in theory at least*, any number of attributive adjectives to appear in this position.

The framing constraint was the availability of the option to add an adjective phrase. The interaction evidence was evaluated *in the context of that constraint*.

Clearly, every time a speaker adds an adjective phrase they change the meaning of the noun phrase – they are restricting its possible referent (expressing *green boat* eliminates boats of other colours) – but this is not to say that a meaning change caused the speaker to prioritise one choice over another.

We discuss the implications of allowing meaning to change in more detail in Chapter 3.

7.3 Framing and annotation

The requirement to consider framing constraints has one more implication. The richer the annotation scheme applied to the corpus, the greater the number of framing constraints that can be readily detected, and the larger the number of research questions that may be explored.

For instance, in a parsed corpus it is possible to study the interaction between decisions framed by grammatical constituents. Without a reliable means to classify subjects for example, it would be more difficult to distinguish cases of interrogative modals.

Similarly, grammatical annotation allows researchers to reliably identify adjectives in the same noun phrase as a noun head, because the NP brackets the query. To take a simple example, a tagged corpus does not distinguish between the following:

- an adjective before a noun, such as *they were young, people said*; and
- an attributive adjective in an NP: *they were young people*.

For obvious reasons, annotation reliability is crucial. This was, if you recall, the minimum condition in Section 4. A tagged corpus never manually corrected may contain long strings of words marked as 'adjectives', which turn out to be nothing of the kind when examined (see Wallis 2019b).

If a corpus is annotated further, e.g. morphologically, prosodically, semantically or pragmatically, then new types of research question become possible. We might characterise these as:

- **intra-level research** studying variation and interaction of decisions encapsulated within a level (e.g. within grammar or semantics);
- **inter-level research** studying the impact of decisions between different levels, e.g. the interaction between a grammatical choice and a prosodic one; and
- **integrated research** studying variation and interaction in phenomena that are each identified across multiple levels, e.g. studying variation in noun phrases that have a particular pragmatic function.

7.4 Framing and sampling

A final aspect concerns sampling. Since decisions made in the same text (particularly if they are adjacent) can interact with each other, this can also affect sampling when, as is common, multiple cases are drawn from the same text. Statistical methods assume by default that instances in a sample are randomly obtained and therefore independent from each other, but in some cases they may not be. We discuss how we may control for this problem in Chapter 17.

8. Conclusions

In this chapter we summarised two simple corpus linguistics experimental designs to show that corpus linguistics can be commensurate with other approaches to linguistics research, such as theoretical

linguistics and psycholinguistics. Corpus linguistics methods can generate linguistically interesting and novel research outcomes that require theoretical explanation and additional experiment.

Science typically proceeds by triangulation rather than refutation, not least because every field of study relies on 'auxiliary assumptions', that is, underpinning assumptions necessary for an experiment to take place. Biological research with optical microscopes relies on optics, slicing and staining techniques; early DNA research relied on electrophoresis; corpus linguistics relies on standards of linguistic representation, including transcription and annotation standards. Whereas in settled science, auxiliary assumptions infrequently change (although new techniques come to the fore), agreed linguistics frameworks are not universal. We must expect representational plurality and competing frameworks for some time to come.

We have attempted to summarise the different types of evidence that might be obtained from a corpus, and the impact of employing a particular type of rich annotation, a phrase structure parse analysis, on this evidence. We have also shown how different representations in a corpus (annotation) can be partially separable from research goals, by emphasising the need for an explicit mapping between them (abstraction). Note that they are *partially* separable: a research question that required the identification of phenomena not annotated at all would require either a fresh annotation effort or the manual extraction of examples.

The three processes of developing and applying annotation schemes, refining queries and specifying experimental datasets are knowledge-rich and cyclic. Annotation is necessarily conditional and subject to revision, either during the compilation of a corpus or in successive post-publication revision cycles.

Similarly, abstraction is cyclic, and – given the plurality of frameworks – necessarily so. We briefly noted how software like ICECUP may accommodate this. Facilitating abstraction in this way has enabled complex novel experiments.[14]

This same cyclic perspective applies to analysis. The overarching perspective of this book is that 'statistical analysis' should be considered a method for evaluating meaningful observations in data.

Analysis is cyclic when it leads to new experimental designs. Researchers should consider, not just *that* their data is distributed in a particular way, but what underlying processes might be that generated this distribution. In other words, they must consider new hypotheses and how they might be tested.

The fact that many of these results are only obtainable from volumes of linguistic data, corpora, demonstrates what corpus linguistics is capable of. Contrary to the dominant paradigm of 'big data' corpus linguistics, these studies emphasise the value of *rich* data. We need annotation to distinguish between linguistic framing constraints and the choices that speakers and writers make within those constraints.

Ignoring framing constraints in research and counting frequencies as if each word was independent from the next (reporting per million word exposure rates) is to ignore the structure of language. However, in the absence of reliable annotation, corpus linguists have tended to overlook this problem – justifying some of Chomsky's criticisms of corpus linguistics.

Corpus linguistics cannot 'prove' the correctness of one internal framework over another. In fact, due to dependence on auxiliary assumptions, no scientific research programme is ultimately capable of refutation merely by observation. Our equipment may be wrong!

Science validates and provokes theories, but theories are not disproved or proved by evidence alone. Indeed, 'evidence', its selection and interpretation, is only obtained by the application of auxiliary assumptions. Without engagement with real-world data, however, theory rests in the realm of philosophy – however sophisticated and computer literate its adherents.

This, ultimately, is the answer to Noam Chomsky's objection regarding the use of corpora. He is absolutely correct to criticise mere summaries of facts and frequencies without reference to an underlying theory. But to reject corpus evidence *per se* on the grounds that it is an external manifestation of internal processes is ultimately to reject the refutation of linguistic theory. As we have demonstrated, corpus frequency and interaction evidence may provide new evidence for theoretical linguists to engage with.

1. The research in this chapter was originally published as Wallis (2014) in L. Veselovská and M. Janebová (eds.) *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Olomouc: Palacký University, 2014. Research introduced in Section 6 was published as Wallis (2019b), Investigating the additive probability of repeated language production decisions, *International Journal of Corpus Linguistics, 24*(4), 492-525. See www.benjamins.com/catalog/ijcl.

2. Traditional discussions of corpus linguistics methodology have tended to focus on a dichotomy between top-down 'corpus-based' and bottom-up 'corpus-driven' research (Tognini-Bonelli 2001). In this book, we argue that both positions are better seen as complementary arms of cyclic research. For a detailed discussion see (Wallis 2020).

3. We return to this question in Chapter 17.

4. We could interpret the terms 'corpus' and 'linguistic event' under a still broader definition. Untranscribed tape recordings or hand-written field notes, whilst not in the digital domain, are 'corpora' for the purposes of this definition. This relaxed definition would allow us to draw parallels with non-linguistic fields such as 'digital humanities', where researchers are engaged in the digitisation and representation of cultural artifacts, from museum exhibits to architectural drawings. The same types of evidence are obtainable by the types of process that we discuss in Section 3.

5. Lavandera (1978) argued that alternation research should not involve choices that change referential meaning. See Chapter 3.

6. For example, the *University of Pennsylvania Treebank* was released in two versions, Treebank I (Marcus, Marcinkiewicz and Santorini 1993) and Treebank II (Marcus *et al*. 1994).

7. See also Nelson *et al.* (2002) and www.ucl.ac.uk/english-usage/resources/ftfs.

8. For reasons of space, ICECUP defaults to a left-right visualisation of tree structures. The top of the tree is on the left, with sub-elements to the right, and the sentence runs down the right-hand side.

9. See www.ucl.ac.uk/english-usage/projects/ice-gb and www.ucl.ac.uk/english-usage/projects/dcpse.

10. The concepts of 'decidability', syntactic gradience (Aarts *et al.* 2007) and 'retrievability' (Wallis 2007) are closely related. See also Wallis (2019b).

11. DCPSE uses the TOSCA/ICE scheme (Nelson *et al*. 2002). Gloss: SU,NP = subject noun phrase; VB,VP = verb phrase; OP,AUX = auxiliary verb acting as an operator. Some links are specified: white down arrow = node follows, but not necessarily immediately; absent up/down links below the SU,NP node insist that the NP has only one child, i.e., it consists of the single pronoun *I* or *we*. Both words are directly connected to their associated node.

12. Two points on the same line may be compared visually by checking whether an earlier point is within the interval for a later one. Such cases will be statistically significant. See Chapter 7.

13. In fact, all quantitative research with language data contains assumptions of this kind (Wallis 2020). The most basic auxiliary assumption necessary for a tagger of English to function is that space characters subdivide words. But what about hyphenated forms? Is there a difference between *larger than life* and *larger-than-life*? Is each words grammatically independent or might it be part of a compound? These questions require a decision one way or another. We might further decide to separate out genitive markers (*'s* and *'*), but not possessive markers, etc.

14. It has also permitted us to develop a range of grammar teaching resources that draw from ICE-GB but may deviate from the parsing scheme (Greenbaum 1996a, Aarts and Wallis 2011, and www.englicious.org).