What might a corpus of parsed spoken data tell us about language?

Sean Wallis
Survey of English Usage, University College London
s.wallis@ucl.ac.uk

Abstract

This paper summarises a methodological perspective towards corpus linguistics that is both unifying and critical. It emphasises that the processes involved in annotating corpora and carrying out research with corpora are fundamentally *cyclic*, i.e. involving both bottom-up and top-down processes. Knowledge is necessarily partial and refutable.

This perspective unifies 'corpus-driven' and 'theory-driven' research as two aspects of a research cycle. We identify three distinct but linked cyclical processes: annotation, abstraction and analysis. These cycles exist at different levels and perform distinct tasks, but are linked together such that the output of one feeds the input of the next.

This subdivision of research activity into integrated cycles is particularly important in the case of working with spoken data. The act of transcription is itself an annotation, and decisions to structurally identify distinct sentences are best understood as integral with parsing. Spoken data should be preferred in linguistic research, but current corpora are dominated by large amounts of written text. We point out that this is not a necessary aspect of corpus linguistics and introduce two parsed corpora containing spoken transcriptions.

We identify three types of evidence that can be obtained from a corpus: factual, frequency and interaction evidence, representing distinct logical statements about data. Each may exist at any level of the 3A hierarchy. Moreover, enriching the annotation of a corpus allows evidence to be drawn based on those richer annotations. We demonstrate this by discussing the parsing of a corpus of spoken language data and two recent pieces of research that illustrate this perspective.

Keywords: corpus linguistics, philosophy of science, epistemology, 3A cycle, parsing, speech

1. Introduction

The field of corpus linguistics has grown in popularity in recent years. Moreover, many researchers who would not otherwise consider themselves to be *corpus* linguists have begun to apply corpus linguistics *methods* to their linguistic problems, a growth that is partly attributable to an increasing availability of corpus data and tools. It therefore seems apposite to take stock, and question what kinds of research can be done with corpora and which types of corpora and methods might yield useful results.

This methodological 'turn to corpora' does not have universal support. Some theoretical linguists, including Noam Chomsky, argue that, at best, any collection of language data merely provides researchers with examples of the actual external performance of human beings in a given context (see, e.g. Aarts 2001). Corpora do not provide insight into internal language or its production processes. Such a position raises questions about *what* data, if any, might be used to evaluate 'deep' theories, as linguists' personal intuitions are no more likely to pierce the veil of consciousness. Nevertheless, this contrary position raises a serious challenge to corpus researchers. We will return to the question of the potential relevance of corpus linguistics for the study of language production by reporting on some recent research in Section 6.

What do we mean by 'a corpus'? In the most general sense, corpora are simply collections of language data that have been processed to make them accessible for research purposes. The largest current corpora contain primarily *written* texts, that is, texts generated by authors at keyboards, screens or paper. These are types of language that are rarely spontaneously produced, frequently edited by others, and often included in databases due to their ease of availability. They may also be written with an imagined audience, in contrast to spoken utterances produced for a co-present (and interacting) audience. Although written data of this kind is easy to obtain, and therefore large corpora are readily compiled, this sampling methodology places significant limitations on the types of inference that might be safely

drawn. The ability to test hypotheses against unmediated, spontaneously produced linguistic utterances seems paramount.

However, not all corpora are collected from written sources. In this paper, we are particularly interested in what corpora of *spoken* data, ideally in the form of recordings aligned with an orthographic transcription, might tell us about language. Transcriptions of this kind should record the actual lexical output, e.g. including false starts, examples of self-correction and overlapping speech, unedited by the speaker. In an uncued, unrehearsed context, this kind of speech data is arguably the closest to genuinely 'spontaneous' naturalistic language output as is achievable. The lexical record can be aligned with an audio and video recording, contain meta-linguistic information, gestural signals, and so on.

Prioritising speech over writing in linguistics research has other justifications aside from mere spontaneity, which might otherwise be achieved by simply recording every keystroke. Speech predates writing historically, both generally and in relation to literacy spread. Child development sees children express themselves through speech earlier than they write, and many writers are aware that their writing requires a more-or-less internal speech act. Our corpus data has approximately 2,000 words spoken by participants every quarter of an hour. By contrast, the author Stephen King (2002) recommends that authors try to write 1,000 words a day. Allowing for individual variation, and with the exception of isolated individuals or those unable physiologically to produce speech, it seems likely that human beings produce, and are exposed to, much more speech than writing.

Axiomatically, different sampling frames obtain different kinds of corpora. Spoken data may be collected for a variety of purposes, some more representative and 'natural' than others, such as telephone calls or air traffic control data. Some spoken data might be captured in the laboratory: collected in controlled conditions, but unnatural, potentially psychologically stressed, and not particularly representative. So when we refer to 'spoken corpora', we are fundamentally concerned with naturally-occurring 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 away from a lab does not mean that results should not be commensurable with laboratory data. On the contrary, corpus data can be a useful complement to lab experiments. 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* analysis of such data, rather than a controlled data collection exercise under laboratory conditions. 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.

As a result of this perspective, corpora usually contain whole passages and texts, in order to be open to multiple levels of description and evaluation. Laboratory research collects fresh data for each research question, and therefore may record data efficiently, containing relevant components of the output determined *a priori*.

However, the lines between the lab 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, data reuse maximises the benefits of a costly research effort. Other sciences have also begun to take data reuse seriously. Medical science has seen computer-assisted meta-analysis, where data from multiple experiments are combined and reanalysed, become increasingly standard.

Given that we have 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? Traditional discussions of corpus linguistics methodology have tended to focus on a dichotomy between top-down 'corpus-based' and bottom-up 'corpus-driven' research. We will argue that both positions are one-sided and are usefully subsumed into an exploratory cyclic approach to research.

2. What can 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).[1] These are

1. **Factual** evidence of a linguistic event, i.e. at least one event x is observed.
2. **Frequency** evidence of a linguistic event, i.e. $F(x)$ events are observed.
3. **Interaction** evidence between two or more linguistic events, i.e. that the presence of a different event y in a given relationship to x affects the likelihood that x will occur, which we might write as $p(x \mid y)$.

Whereas much theoretical linguistic argument is given over to stating that particular expressions are or are not possible, the **factuality** of any theory is ultimately only testable against real world data. Dictionaries expand by observing new forms and earlier attestations. More controversially perhaps, we would argue that for a theoretical linguist to maintain that a particular construction found in a corpus is 'bad' or 'impossible' constitutes an insufficient argument and the errant datum deserves explanation. Such an explanation might be that it represents a performance error, but this cannot be assumed *a priori*. So factual evidence might present evidence that appears to contradict or challenge existing theories.

Perhaps the least controversial statement above is that corpora are a rich source of **frequency** evidence for linguistic phenomena. Most existing corpus research concentrates on frequencies of linguistic events.

Frequency evidence has value, even if its meaning is less easy to discern. Knowing that one construction, form or meaning is more frequent than another has proven beneficial for writers of dictionaries and grammar books, helping them prioritise pedagogically. Frequency evidence may be counterintuitive, and it is harder for the intuition-driven linguist to deny corpus data this purpose. On the other hand, the most common criticism of corpus linguistics is that it consists of mere counting of words or constructions. How does such evidence relate to the concerns of the theoretician?

Frequency data must be interpreted carefully. A common confusion mixes up **exposure rates**, typically, that x appears n times per million words, and **choice rates**, that x is chosen with probability p when the choice of using x arises.

An exposure rate tells us how frequently an audience for a set of utterances 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 thus an elevated or reduced frequency in one context over another may be due to many factors. Most importantly, however, exposure rates are not easily commensurable with linguistic theory.

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 given the opportunity to do so. If we identify a superset of alternative forms including x, which we might denote as $\mathbf{X}$, we simply obtain $p(x \mid \mathbf{X}) = F(x)/F(\mathbf{X})$. In a lab experiment, this is equivalent to cueing a participant with an input and observing their response. Employing choice rates (also known as 'the variationist paradigm') is common practice in sociolinguistics but less common in corpus linguistics more generally.

The principal difficulty is practical. Particularly with lexical corpora, reliably identifying all possible choices at any given point is difficult. Many corpus linguists have

[1] We could interpret the terms 'corpus' and 'linguistic event' under an even broader definition. Untranscribed tape recordings or hand written field notes, whilst not in the digital domain, are still 'corpora' for the purposes of this definition. Such a generous definition would allow us to draw parallels with non-linguistic fields such as 'digital humanities', where researchers are engaged in the digitization and representation of cultural artefacts, such as museum exhibits and architecture. The same types of evidence are obtainable by the types of process that we will discuss in the following section.

expressed unease as the choice appears arbitrary, and a number of objections have been raised by corpus linguists to this approach. See (Wallis forthcoming a) for a thorough discussion.

Intermediate positions between hearer exposure and speaker choice are also possible.[2] For example, it is legitimate to survey the behaviour of modal auxiliary lemmas as a comparative exercise, i.e. whether *can* or *will* are increasing as a proportion of all modals, without claiming that they are mutually substitutable, so that the speaker can simply choose between them. A crucial skill for a corpus linguist is to recognise these different kinds of frequency evidence and to properly report their implications.

Finally, **interaction** evidence concerns the effect of one word, construction or utterance on others. To take a trivial example, if a speaker begins an utterance with a personal pronoun the hearer will intuit that the most likely next word will be a verb. Interaction evidence is employed in computer algorithms, such as part-of-speech (POS) taggers and parsers, but it may appear at multiple levels. An important class of interaction evidence is obtained from choice rates. If we can identify the probability of a speaker employing a construction when they have the option, we can also identify the effect of a co-occurring construction on that probability.

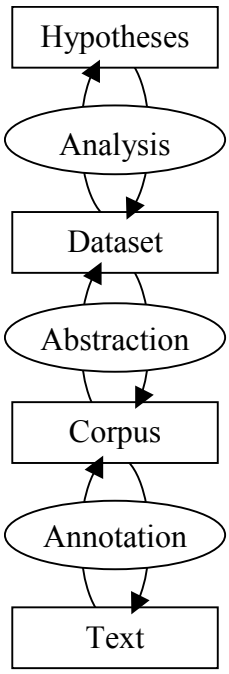

Figure 1: The 3A perspective in corpus linguistics (after Wallis and Nelson 2001)

Note that thus far we have been discussing corpora in general without considering the classes of linguistic event that might be reliably obtained from them. If a corpus consists of plain text, then the events identified above are lexical, and this evidence can only really inform lexical studies. However, the pre-computer era corpora (Brown, Survey) may not have always been digitized, but they have always relied on annotation.

3. The 3A cycle

Our second epistemological 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 – **annotation, abstraction,** and **analysis** – bridging four distinct levels of knowledge. This approach, which we call the '3A perspective' (Wallis and Nelson 2001), is sketched in Figure 1.

Each process adds knowledge in a cycle of addition and critical reflection. Knowledge, necessary and refutable, is applied at every level, from sampling decisions to hypotheses. When we annotate a text we both add information to it – e.g. sentence boundaries, POS tags – and, simultaneously, critically reflect on our frameworks. Is it useful to have a concept such as a 'sentence boundary' in spoken data? Does this word have this part-of-speech tag? Should the scheme be modified?

In the case of spoken data, the source is not text but an audio signal, and 'annotation' is properly conceived of as including the process of transcription. Whatever the source data, both the annotated text and the annotation scheme are subject to change over the course of annotating an entire corpus. The more experimental the annotation scheme, the more likely that it will be subject to co-evolution while the corpus is annotated. Obtaining complete coverage of a scheme across a corpus will inevitably throw up unanticipated challenges in dealing with the new 'facts' we observed in the previous section (hence factual evidence is sometimes referred to as *coverage* evidence).

In corpus linguistics, the annotation cycle is typically, although not exclusively, performed by the collectors of the corpus prior to distribution. However, such a practice is

[2] See also e.g. http://corplingstats.wordpress.com/2013/04/02/a-methodological-progression/ and http://corplingstats.wordpress.com/2013/03/06/choice-vs-use/

clearly not a *defining* characteristic of corpus linguistics. Indeed, one team might add annotation to data obtained by another, or the same team might add or modify annotation in a series of phases, each with their own release.

Usually, however, corpus linguistics practice places a sharp line between annotation and abstraction. Annotation commonly ends with the distributed corpus, although as we see in Section 5, sometimes researchers have to perform additional annotation steps to manually classify data according to unencoded criteria.

Abstraction begins the process of 'research proper', when linguists *with a particular research goal in mind* attempt to obtain data from an annotated corpus and transform it into a dataset that can be analysed using conventional data analysis methods. 'Abstraction' is sometimes termed 'data transformation' or 're-representation' in the field of Knowledge Discovery, or 'operationalisation' in Experimental Design and Statistics textbooks.

Abstraction 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. How does she know that her results are correct, i.e. that they reliably identify the examples that she wants? Like annotation, abstraction must be cyclic, including the reverse process ('concretisation'). In other words, it is necessary for the researcher to see how her query matched cases in the corpus, try other queries, etc.

Even experienced researchers have to learn an annotation scheme to formulate meaningful queries on a given corpus. To do this they must be able to perform abstraction by approximation, evaluation and refinement.

In the case of lexical queries, the opportunity for testing and revision may appear unimportant. Provided that searches are not case sensitive, lemmas are identified appropriately, etc., it may be assumed that a trained researcher will obtain an accurate query first time. However, the more complex the annotation scheme, the greater the need for the researcher to review and revise her queries in the light of their application. Indeed, it is difficult to see how a researcher can ever be said to 'know' a parsing scheme, for example, sufficient to obtain data for her research, unless she is able to see how it has been applied to the relevant data in a corpus.

A crucial problem, and a standard objection to richly annotating a corpus, concerns **representational plurality**. Assume that in any given field of research, linguists differ in their ideal representation scheme, and schemes are often in a state of development themselves. Schemes may differ terminologically, but far more importantly, they may differ in their classification and structuring of linguistic phenomena.

Thus Quirk *et al* (1985) exclude objects from the VP analysis, Huddleston and Pullum (2002) include objects, dependency grammars represent Quirk's VPs another way, and so on. After some 20 years of corpus parsing, we have a wide range of corpora attempting to capture an overlapping set of comparable linguistic performances with very different schemes. Leaving aside the fact of divergent corpus annotation schemes, any linguist who uses a corpus must abstract from the annotated corpus to concepts that are commensurable with their preferred framework.

The necessity and importance of abstraction as a process has been frequently overlooked. However, it is a central issue in the design of software tools for working with richly annotated corpora. As we noted, lexical corpora with simple POS tagging may not require an extensive cyclic process of query refinement. The more extensive the annotation, however, the more frequently a researcher will need to try out different queries.

The *ICECUP* software (Nelson, Wallis and Aarts 2002) was designed around the abstraction cycle to support research with a parsed corpus: initially, the 1 million word *British Component of the International Corpus of English* (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 kind of

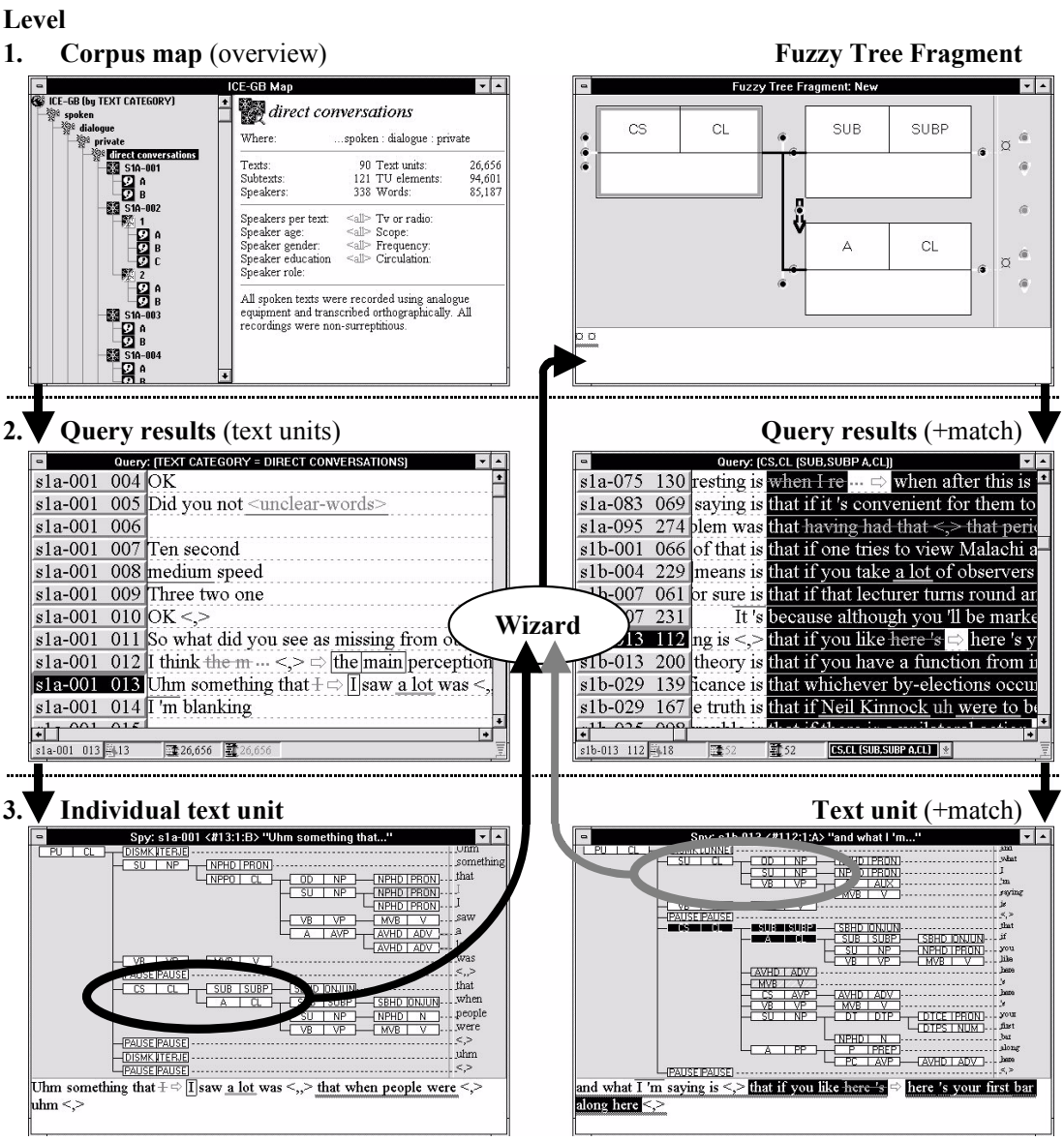

Figure 2: Exploring the corpus, after Nelson *et al*(2002): from the top, down (left), and, using the Wizard in an exploration cycle with FTFs (right).

"abstract tree" where both nodes and links between nodes may be incompletely specified, similar to a lexical wild card. At the top right of Figure 2 we have an FTF that searches for structures consisting of subject complement clauses (CS,CL) containing a subordinate phrase (SUB,SUBP) followed by an adverbial clause (A,CL).[3] When a query is applied, the set of matching cases are immediately presented by the interface (middle right). Researchers can review how their queries have been matched to the corpus and identify false positive cases. A 'Wizard' tool permits a researcher to select parts of the tree annotation and convert it into a FTF.

The tools are linked together in a forgiving user interface on top of a specialised database system. Each window in Figure 2 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).

[3] See also Nelson *et al* (2002) and www.ucl.ac.uk/english-usage/resources/ftfs.
[4] Note that, for reasons of space, ICECUP defaults to a left-right visualisation of tree structures. The top of the tree is on the left, and the sentence runs down the page.

Figure 2 also shows how ICECUP tools are considered to relate to one of three levels of generalisation: level 1 consists of sets of queries, level 2 consists of query results (sentences/matching cases) and level 3 corresponds to individual instances (sentence + tree annotation).

Finally, 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 therefore knowledge must be encoded in advance. Suppose 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 POS tags are abstracted; and finally processed for particular actions. If Langley or GCHQ are listening in, rest assured that their systems are engaged in identifiable processes of annotation, abstraction and analysis![5]

4. What can a richly annotated corpus tell us?

Let us now briefly consider how the three types of evidence identified in Section 2 apply to a richly annotated corpus. A good example is a **parsed corpus**, i.e. a corpus like ICE-GB, or its relation, the *Diachronic Corpus of Present-day Spoken English* (DCPSE),[6] consisting wholly of spoken transcriptions. The same principles apply, however to any corpus containing annotation that represents one or more levels of linguistic structure, such as morphological or pragmatic structure.

In this paper, we focus on parsed corpora, where every sentence has been given a tree analysis according to a chosen scheme. In the case of spoken data, where 'sentences' may not exist or must be inferred, decisions to split utterances into sentences will be integral to the parsing process, i.e. they are part of the analytical decisions that must be made in applying the scheme to the data. The notion of a 'linguistic event' identified in general terms in Section 2 may now be 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 (such as two terms bridged by a relationship link or a particular clause structure) and
3. **any combination**of the above with elements of the source text.

As multiple levels of annotation are added, it is additionally possible to identify co-occurrences *between* levels. Thus a corpus consisting of parsed and pragmatically annotated text would permit grammatical and pragmatic elements to be identified in combination, such as a particular opening 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, $p(x \mid y)$).

If we must enrich our corpora with annotation, how do we choose between potential schemes? In parsing, for example, schemes applied to corpora in the past were chosen according to a range of criteria. These included simplicity and minimalism (Penn Treebank I,

[5] Similarly, statistical methods and machine learning algorithms can be applied to each cycle. The most common application is in POS-tagging and parsing, which may be seen as sub-processes within the annotation cycle. In principle, knowledge at any level may be enhanced by statistical generalisation from the level below. 'Skipping' levels, however, risks superficial generalisation from surface features.

[6] See www.ucl.ac.uk/english-usage/projects/ice-gb and www.ucl.ac.uk/english-usage/projects/dcpse.

Marcus *et al*. 1993), text mining applications (Treebank II, Marcus *et al*. 1994), and linguistic tradition (ICE, based on Quirk *et al.* 1985; Prague Dependency Grammar, Böhmova *et al*. 2003, etc).

In this paper we take a different approach. Let us consider the question from the perspective of a corpus researcher. There are at least two different ways of evaluating an annotation scheme, such as parsing or any other level of rich annotation, applied to corpora.

- **Annotation facilitates abstraction** ('a handle on the data'). In this theory-neutral position, 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, provided that distinctions embodied in the scheme are sufficient for research goals. In other words, provided that a researcher may reliably abstract from annotated corpus to their experimental paradigm, the actual annotation encoding is irrelevant.
- **Annotation facilitates theoretical goals** (potentially, the identification of linguistic processes). Models of priming and spreading activation imply that decisions made by speakers and writers are influenced probabilistically by previous decisions, as we see in Section 6. An annotation scheme that enables evidence of this kind to be found reliably would thus be better justified than one that does not. Design of such a scheme is less theory-neutral than in the first position, and the ideal annotation would be one that reflected a credible trace of the language production process undergone by the speaker, what we have elsewhere referred to as 'speaker parsing' as distinct from 'hearer parsing'.

In the first perspective, annotation schemes may be compared in relation to their ability to reliably retrieve linguistic events (Wallis 2008), a criterion sometimes termed *decidability*. We can say that a corpus whose annotation reliably classifies nouns and verbs is better than an unreliable classification, and a representation that explicitly denotes subjects of clauses is preferable to one that does not. However, as these examples imply, this criterion is circular. Why should we assume, *a priori*, that reliable retrieval of subjects or nouns is important? Moreover, as Wallis (forthcoming b) observes, such criteria admit redundancy, because any representation can improve on another by simply gaining 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. But 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 non-trivial causes 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.

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

Annotation schemes are the 'auxiliary assumptions' of the research programme. From this perspective, the annotation scheme cannot be evaluated in the abstract, but should be 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, then, is what linguistic research goals could annotation schemes attempt to further? We will attempt an initial answer in Section 6, but first let us look at research of the first kind.

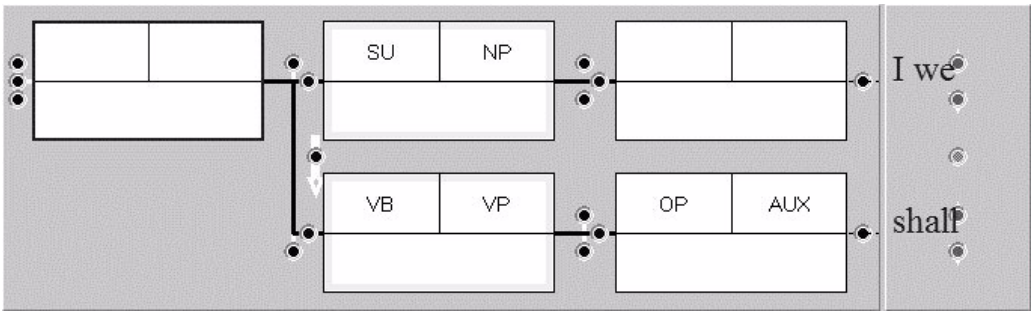

Figure 3: An FTF for a first person subject (*I* or *we*) followed by auxiliary verb *shall*, after Aarts *et al* (2013). To search for *will* and *'ll* the lexical item *shall* is replaced.[7]

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

Much research is typical of the 'annotation drives abstraction' perspective. Aarts *et al*. (2013) looked at the alternation between the modal auxiliaries *shall* and *will* over time, in first person declarative contexts.

Whereas previous studies (e.g. Mair and Leech 2006) had considered *shall* and *will* (including negative *shan't* and *won't* and cliticised *'ll* = *will*), these studies had a number of drawbacks. First, they tended to analyse these modals in terms of exposure rates (*shall* and *will* per million words). This meant that it was not possible to factor out sampling variation due to varying potential to use either *shall* or *will* (e.g. in past-oriented texts either would be less frequent than in present-oriented ones). Despite this, it is a relatively simple matter of reanalysis to pose the question in terms of a basic choice rate (*shall* as a proportion of the set {*shall*, *will*}).

Second, these studies were carried out on part-of-speech tagged corpora which were not parsed. However, alternation of *shall* and *will* rarely exists except with first person subjects. The ideal is to identify just those cases of *shall* where the speaker has a genuine choice of using *will* instead, and vice versa. Consider the interrogative case: *Shall we go to the park?* and *Will we go to the park?* are semantically and pragmatically distinct, and therefore do not freely alternate. We therefore focused on first person declarative cases, and for similar reasons we also decided to eliminate negative cases.

This was made much easier by the fact that we were working with the parsed corpus,

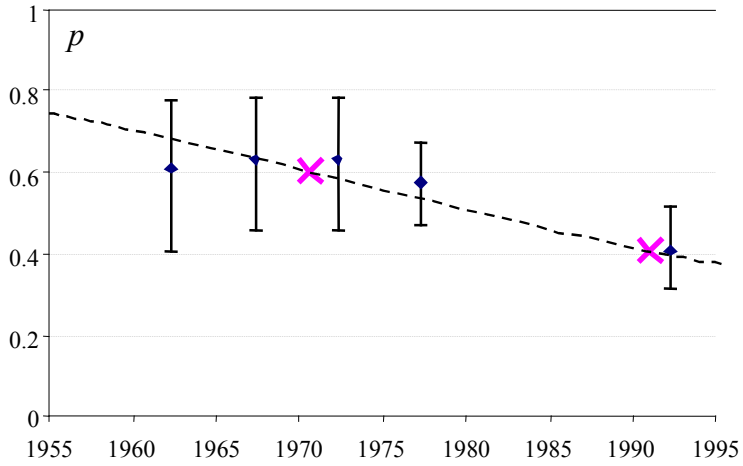

Figure 4: Declining use of *shall* as a proportion *p* of the set {*shall, will*} with first person subjects, half-decade data ('1960' = 1958-62 inclusive, etc.) (after Aarts *et al* 2013).

[7] 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 SU,NP node insists that the NP has only one child, i.e., it consists of the single pronoun *I* or *we*. Finally, and possibly the most subtle point, both words are directly connected to their associated node.

DCPSE, and ICECUP. In order to reliably extract cases of first person declarative positive uses of *shall* and *will*, we were able to use FTF queries like Figure 3.

The FTF works on the annotation scheme by relating individual terms and structure, and the result is a reliable retrieval mechanism for obtaining relevant cases. The annotation is a 'handle on the data' allowing us to pull out instances of linguistic events, in this case the use of *shall* or *will* in the particular context required. We obtained graphs such as the one in Figure 4, showing the tendency to prefer *shall* over *will* falling over the course of time.

Consider the steps that would be required to obtain these results were DCPSE only analysed using part-of-speech tagging. It would be possible to construct queries that searched for patterns of a first person pronoun followed by *shall* or *will*, but we would then have to manually review each pattern to verify that it was part of the same clause. In effect, we would be performing the necessary additional annotation stage (cf. Figure 1) at research time. Annotation is unavoidable.

Similarly, in this study, Joanne Close manually classified each instance of previously identified *shall* and *will* by their modal semantics (Epistemic, Root and 'other'), allowing her to conclude that the identified fall in an overall preference for *shall* was actually due to a sharp decline in Epistemic uses of *shall*. Again, this step is a type of annotation, except that it is being performed by researchers using the corpus for a particular research goal instead of being performed by the publishers of the corpus.

6. Interacting grammatical decisions: NP premodification

The previous illustrative study examined variation in a linguistic choice over time. Other sociolinguistic variables, such as speaker gender, text genre, mode, contrasting monologue and dialogue, etc. are within the same experimental paradigm.

On the other hand, if we are interested in linguistic, rather than sociolinguistic, influences on language choices, we need to extract and attempt to interpret interaction evidence. Interaction evidence may simply consist of exploring two closely related grammatical variables (see Chapter 9.7 in Nelson *et al.* 2002). Examples given 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.

Recent research (Wallis forthcoming b) examines a different and more general phenomenon, i.e. serial repeated additive decisions applied in language production. This research paradigm evaluates decisions to add or not to add a particular construction to a superordinate one, and tests whether the speaker or writer is more or less likely to make the decision on subsequent occasions.

This methodology can be seen as a way of examining construction complexity (a static interpretation) or as a way of examining the interaction between language production decisions (a dynamic one). Either way, the patterns we obtain are highly interesting, occasionally counter-intuitive, and worthy of theoretical discussion.

A simple illustrative example is attrributive adjective phrases premodifying a noun phrase head, thus we have *ship*, *green ship*, *tall green ship*, etc. We can use FTFs to identify NPs with a common noun head, NPs with at least one attributive adjective phrase, NPs with at least two adjective phrases and so on. We obtain the data in Table 1 by applying these FTFs to ICE-GB across both speech and writing.

From the frequency of at least a attributive phrases, $F(a)$, (top line) we derive a set of

***a* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *a*' *F*(*a*)	193,135	37,305	2,944	155	7
Probability *p*(*a*)		0.1932	0.0789	0.0526	0.0452

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

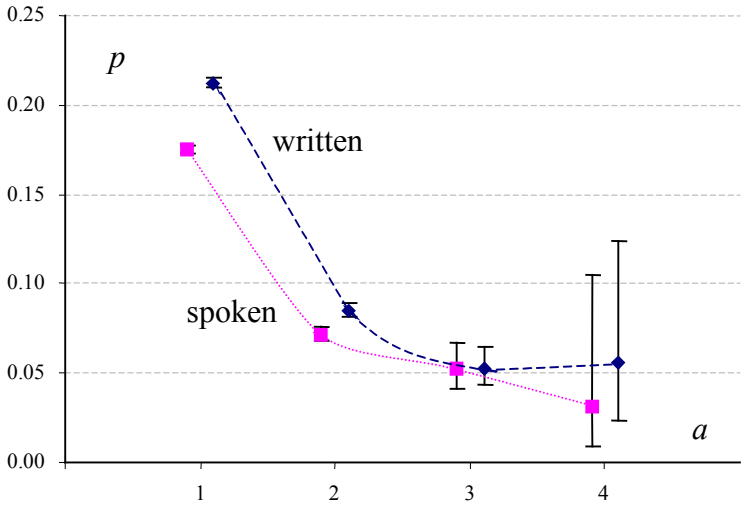

Figure 5: Declining probability of adding attributive adjective phrases to an NP noun head, data from ICE-GB, patterns for speech and writing.

probabilities, $p(a) \equiv F(a)/F(a–1)$. These probabilities are the observed likelihood that, given $a–1$ attributive phrases, the speaker/writer will add a further adjective phrase. 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, a, as in Figure 5. This graph includes 95% Wilson score confidence intervals and distinguishes spoken and written performance.[8]

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.

Plotting $p(a)$ reveals that the decision to add a second attributive phrase after a first (the particular order of decisions is irrelevant, and they may be made in parallel) 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 within grammar. Wallis (forthcoming b) considers adverbial phrases premodifying a VP (e.g. *quickly*, *intelligently*, *getting to the point*) and finds no interaction between the decision to add one or two adverbial premodifiers. It is necessary to consider possible explanations for this phenomenon.

There are at least three potential sources of this interaction.

- **semantic and logical constraints**, which would include the well-researched English phenomena of attributive ordering (cf. *tall green ship* vs *green tall ship*) and avoidance of illogical descriptions (*tall short ship*);
- **communicative economy**, avoiding unnecessarily long descriptions, especially on the second and third citation (on, subsequent occasions referring simply to *the ship* rather than *the tall green ship*), and
- **psycholinguistic attention and memory constraints**, so that speakers found it more difficult to produce longer constructions.

In the case of NP premodification, the most likely explanation is the first. Communicative economy would predict a rapid drop from $p(1)$ to $p(2)$ but no subsequent fall. Psycholinguistic constraints are implausible because the added constructions themselves are relatively 'light' memory-wise. Indeed, if a speaker forgot that they had said a previous adjective phrase, it

[8] 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.

seems more likely they would act in an unconstrained, rather than a constrained manner. However, the impact of psycholinguistic constraints are much more plausible explanations for patterns observed with multiple postmodification of NPs (e.g. *the ship [by the harbour] [which we sailed on]*) and embedding (*the ship [by the harbour [in the town]]*), which the author also examines.

Figure 5 also shows that the speech and writing data does not have the exact same distribution, so we can see that a greater proportion of NPs uttered by speakers have no adjective phrases. When they do employ adjective phrases, they tend to use fewer phrases, and so on. There may be a number of possible reasons for this, e.g. the fact that in a conversation the audience is present and referents require less elaboration. Nonetheless, both datasets obtain a similar overall pattern.

Note that the evidence in this experiment is only obtainable from a corpus. One would not spot this trend by laboratory experiment: we simply do not have enough data. For NP premodification, employing a parsed corpus is not required, and simple sequences of the form <ADJ> <N> obtain similar results (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 want to look at non-adjacent terms or structure, the reliable representation of that structure is essential.

Finally, the fact that we can compare spoken and written data is also important. As we noted, the vast majority of corpora exclusively or overwhelmingly contain written data. But we find essentially the same pattern in speech and writing. Figure 5 confirms that we are observing a linguistic phenomenon that is not attributable to a special character of writing or speech: for example, a possible tendency for writers to avoid excessive NP length by editing. The presence or otherwise of an audience may affect the rate of decline but not the overall tendency.

7. Conclusions

We have attempted to summarise the state-of-art in corpus linguistics to show that it does not embody a competing methodology with other approaches to linguistics research, such as theoretical linguistics and psycholinguistics. On the contrary, corpus linguistics can obtain linguistically interesting and novel research outcomes which require theoretical explanation and additional psycholinguistic experiment. Science typically proceeds by triangulation rather than refutation, not least because every field of study relies on 'auxiliary assumptions', underpinning assumptions that are necessary for an experiment to take place. Biological research with optical microscopes relies on optical physics, early DNA research relied on electrophoresis, and corpus linguistics relies on standards of linguistic representation, including transcription/annotation.

Whereas in settled science, auxiliary assumptions infrequently change (although new techniques come to the fore), linguistics frameworks are not universally agreed. Consequently we must expect representational plurality and competing frameworks in our corpora for some time to come. In this paper 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 analysis, a phrase structure parse analysis, on this evidence. We have also shown how different representations in a corpus (annotation) are partially separable from research goals, by emphasising the need for an explicit mapping between them (abstraction).

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

Abstraction is also cyclic, and, given the plurality of frameworks, necessarily so. We briefly noted how software may be developed from the ground up to accommodate this. Facilitating abstraction in this way has enabled complex novel experiments. It has also permitted us to develop a range of grammar teaching resources that draw from ICE-GB but

deviate from the parsing scheme (Greenbaum 1996, Aarts and Wallis 2011, and www.englicious.org).

Finally, we attempted to illustrate our argument with two recent studies, a relatively conventional sociolinguistic predictor of diachronic language change, and a more unusual experiment which examined interaction between grammatical structures, which we might term 'intra-structural priming'. The fact that both sets of results are only obtainable from volumes of linguistic data, i.e. corpora, demonstrates what corpus linguistics is capable of achieving. Contrary to the dominant paradigm of "big data" corpus linguistics, these studies emphasise the value of *rich* data.

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 capable of refutation of deductive internal proof by inductive observation. Our equipment may be wrong! Rather, recent research of the kind we describe in Section 6 may provide evidence that can *validate* possible frameworks, just as physical experiments validate, but do not 'prove', theories of gravity.

This, ultimately, is the answer to Chomsky's objection regarding the use of corpora. It rests on a misconception about science and philosophy. Science validates and provokes theories, but theories are not disproved or proved by evidence alone. Without such engagement with real-world data, however, theory rests in the realm of philosophy – however sophisticated and computer literate its adherents.

Acknowledgments

The author would like to thank Bas Aarts, Joanne Close and Gerald Nelson for their support and critical engagement with the 3A perspective over the years. As a methodological review, this paper cannot do justice to the research they contribute here, and colleagues are encouraged to read the original books and papers cited.

Building a large parsed corpus such as ICE-GB and DCPSE is necessarily a large team effort. This paper is dedicated to all our transcribers and annotators. I hopefully convey the point that the linguistic knowledge they painstakingly applied to the corpus was worthwhile! Sixteen years after ICE-GB was first published we are still scratching the surface of what we might be able to do with it.

References

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

Aarts, B. & S.A. Wallis 2011. *The interactive Grammar of English* (app). London: Survey of English Usage: UCL Business. http://www.ucl.ac.uk/english-usage/apps/ige

Aarts, B., J. Close & S.A. Wallis 2013. Choices over time: methodological issues in current change. In B. Aarts, J. Close, G. Leech and S.A. Wallis (eds.) *The Verb Phrase in English*. Cambridge: Cambridge University Press.

Böhmová, A., J. Hajič, E. Hajičová & B. Hladká. 2003. The PDT: A Three-Level Annotation Scenario, in A. Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer. 103-127.

Greenbaum, S. 1996. *The Oxford English Grammar*, Oxford: Oxford University Press.

Huddleston, R. & G. K. Pullum 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

King, S. 2002. *On Writing*. New York: Pocket Books.

Lakatos, I. 1978, *Mathematics, science and epistemology*. Cambridge: Cambridge University Press.

Mair, C. & G. Leech 2006. Current changes in English syntax. In: B. Aarts & A. McMahon (eds.) *The Handbook of English Linguistics*. Malden MA: Blackwell Publishers. 318-342.

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

Marcus, M., G. Kim, M.A. Marcinkiewicz, R. MacIntyre, M. Bies, M. Ferguson, K. Katz, & B. Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. *Proceedings of the Human Language Technology Workshop*. San Francisco: Morgan Kaufmann.

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

Quirk, R., Greenbaum, S., Leech, G. & Svartvik J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Wallis, S.A. & G. Nelson 2001. Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, **5**: 307-340.

Wallis, S.A. 2007. Annotation, Retrieval and Experimentation. In A. Meurman-Solin & A.A. Nurmi (eds.) *Annotating Variation and Change*. Helsinki: Varieng, UoH. http://www.helsinki.fi/varieng/series/volumes/01/wallis

Wallis, S.A. 2008. Searching treebanks and other structured corpora. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics: An International Handbook. Handbücher zur Sprache und Kommunikationswissenschaft series*. Berlin: Mouton de Gruyter. 738-759.

Wallis, S.A. forthcoming a. *That vexed problem of choice*. prepublished: London: Survey of English Usage, UCL. http://corplingstats.wordpress.com/2012/03/31/that-vexed-problem-of-choice

Wallis, S.A. forthcoming b. *Capturing patterns of linguistic interaction in a parsed corpus: an insight into the empirical evaluation of grammar?* prepublished: London: Survey of English Usage. http://corplingstats.wordpress.com/2012/12/04/linguistic-interaction