Capturing patterns of linguistic interaction in a parsed corpus: an insight into the empirical evaluation of grammar?

Sean Wallis, Survey of English Usage, University College London
December 2012

Numerous competing grammatical frameworks exist on paper, as algorithms and embodied in parsed corpora. However, not only is there little agreement about grammars among linguists, but there is no agreed methodology for demonstrating the benefits of one grammar over another. Consequently the status of parsed corpora or 'treebanks' is suspect.

The most common approach to empirically comparing frameworks is based on the reliable retrieval of individual linguistic events from an annotated corpus. However this method risks circularity, permits redundant terms to be added as a 'solution' and fails to reflect the broader structural decisions embodied in the grammar. In this paper we introduce a new methodology based on the ability of a grammar to reliably capture patterns of linguistic interaction along grammatical axes. Retrieving such patterns of interaction does not rely on atomic retrieval alone, does not risk redundancy and is no more circular than a conventional scientific reliance on auxiliary assumptions. It is also a valid experimental perspective in its own right.

We demonstrate our approach with a series of natural experiments. We find an interaction captured by a phrase structure analysis, between attributive adjective phrases under a noun phrase with a noun head, such that the probability of adding successive adjective phrases falls. We note that a similar interaction (between adjectives preceding a noun) can also be found with a simple part-of-speech analysis alone. On the other hand, preverbal adverb phrases do not exhibit this interaction, a result anticipated in the literature, confirming our method.

Turning to cases of embedded postmodifying clauses, we find a similar fall in the additive probability of both successive clauses modifying the same NP and embedding clauses where the NP head is the most recent one. Sequential postmodification of the same head reveals a fall and then a rise in this additive probability. Reviewing cases, we argue that this result can only be explained as a natural phenomenon acting on language production which is expressed by the distribution of cases on an embedding axis, and that this is in fact empirical evidence for a grammatical structure embodying a series of speaker choices.

We conclude with a discussion of the implications of this methodology for a series of applications, including optimising and evaluating grammars, modelling case interaction, contrasting the grammar of multiple languages and language periods, and investigating the impact of psycholinguistic constraints on language production.

Keywords: grammar, linguistic interaction, evaluation, language production, corpora, treebanks

1. Introduction

Parsed corpora of English, where every sentence is fully grammatically analysed in the form of a tree, have been available to linguists for nearly two decades, from the publication of the University of Pennsylvania Treebank (Marcus *et al*. 1993) onwards. Such corpora have a number of applications including training automatic parsers, acting as a test set for text mining, or as a source for exemplification and teaching purposes.

A range of grammatical frameworks have been exhaustively applied to corpora. Penn Treebank notation (Marcus *et al*. 1993) is a skeleton phrase structure grammar that has been applied to numerous corpora, including the *University of Pennsylvania Treebank* and the Spanish *Syntactically Annotated Corpus* (Moreno *et al*. 2003). Other phrase structure grammars include the Quirk-based TOSCA/ICE, used for the *British Component of the International Corpus of English* (ICE-GB, Nelson, Wallis and Aarts 2002) and the *Diachronic Corpus of Present-day Spoken English*. Dependency grammars include the Helsinki Constraint Grammar (Karlsson *et al*. 1995), which has been applied to (among others) English, German and numerous Scandinavian language corpora. Other dependency corpora include the *Prague Dependency Treebank* (Böhmová *et al*. 2003) and the *Turkish Treebank* (Oflazer *et al*. 2003).

1.1 The problem of grammatical epistemology

Naturally this brief list understates the range of frameworks that have been applied to corpora, and concentrates on those applied to the largest amount of data.

The status of knowledge embedded in a corpus grammar raises some problematic questions. Given the range of frameworks adopted by linguists, how should annotators choose between them? The choice of grammar risks a circular justification – one can train a parser based on one framework on a corpus analysed by the same framework (Fang 1996), but this does not tell us anything about whether the framework is *correct*. To put it another way, what general *extra-grammatical* principles may be identified that might be informed by corpus data? In this paper we argue that parsed corpora can help us find evidence of psycholinguistic processing constraints in language production that might allow us to re-examine this question from a perspective of cognitive plausibility.

Cited motivations for annotator's choice of scheme range from commensurability with a traditional grammar such as Quirk *et al*. (1985) (Greenbaum and Ni 1996), reliability of automatic processing against a minimum framework, and maximising the opportunities for information extraction (Marcus *et al*. 1994).

A related question concerns *commensurability*. If we choose one scheme out of many, are results obtained from our corpus commensurable with results from data analysed by a different scheme, or have we become lead up the garden path by a particular framework? Indeed a standard criticism of the treebank linguistics community is that since theorists' knowledge of grammar is contested and imperfect, corpus annotation is likely to be wrong. John Sinclair (1987) argued that linguistic insight should be driven by word patterns rather than subsumed under a given grammar. Many linguists do not use parsed corpora. Part of the reason may be misgivings about the annotations of others.

Nelson *et al*. (2002) argue that this problem is primarily one of research stance. The research paradigm should consider the limitations of corpus annotation and the potential for results to be an artefact of the chosen framework. They propose a cyclic exploratory methodology where reference back to original sentences, and 'playing devil's advocate', is constantly encouraged.

The gulf between theoretical linguists such as Chomsky, and lexical corpus linguists like Sinclair, is wide. This does not mean, however that this gulf cannot be bridged, as Aarts (2001) points out. Corpus linguists need not eschew theory and theoreticians should consider how their frameworks may be evidenced.

A parsed corpus is a source of three principal types of evidence. First, applying an algorithm to a broad range of text samples provides **frequency** evidence of known phenomena found in the parser rulebase. The manual correction and completion of the parser output both improves this frequency evidence and supplements it with a second type of evidence: enhanced **coverage** with previously unknown rules.

Third, a parsed corpus is a rich source of evidence of lexical and grammatical **interaction**. As speakers form utterances they make a series of conscious and unconscious decisions: to use one word, phrase, etc., rather than an alternative. These decisions are often not independent from each other (i.e., they interact). In this paper we will consider whether evidence of one type of interaction is relevant to the psycholinguistic evaluation of grammatical frameworks.

1.2 Deciding between frameworks

In corpus linguistics the first evaluation (and evolution) of a grammar takes place during annotation. The task of ensuring that every utterance is correctly and consistently described by a grammar presents a series of methodological challenges (Wallis and Nelson 1997; Wallis 2003). The descriptive task typically leads to minor modifications of the grammar. However, such on-the-fly adaptation risks being *ad hoc*, local, and unprincipled. This paper concerns a second process: the review of completed parsed corpora. In order to do this we must first agree evaluative criteria.

By far the most common criterion for arguing that one representation is 'better' than another is *decidability*, or the **retrievability of linguistic events** (Wallis 2008). If a concept – the subject of a clause, a particular type of direct object, etc. – can be reliably retrieved from one corpus representation but cannot be as reliably retrieved with another, then the first representation can be said to be 'better' than the second. This is another way of saying that the event can be distinguished from other similar events.

For example, the scope of attributive adjectives over co-ordinated noun heads varies. The following are not grammatically distinguished in the ICE-GB corpus (see Section 2) and therefore one cannot be retrieved without the others.

fried aubergines and yoghurt [S1A-063 #19] (only aubergines are fried)

late teens and twenties [S1A-013 #107] (ambiguous)

recent article and correspondence [S1B-060 #42] (both are recent)

Retrievability of events is a useful criterion, but it has three problems. These are **circularity** (the value of a concept in question, such as attribute scope, must be assumed), **redundancy** (a representation can 'improve' by simply adding distinctions like those above to the framework), and **atomisation** (single events within a grammatical structure are evaluated, rather than the structure itself).

1.3 Interaction along grammatical axes of addition

In this paper we propose and explore a complementary 'structural' approach to empirically evaluating grammar by examining patterns of interaction between concepts along grammatical axes. We believe that our approach has cognitive plausibility, and that the parsed corpus may reveal some novel non-obvious consequences of language production. The method builds on the 'linguistic event retrieval' principle above by exploring the impact of one linguistic event on another event of the same type.

We will study patterns of repeated decisions of the following form:

$$base \rightarrow +term_1 \rightarrow +term_2 \ldots +term_n,$$

where arrows indicate separate decisions to add a further term and plus signs indicate the application of an operator that adds terms in a specific way (i.e., governed by a particular structural relationship) along a particular grammatical axis. Grammatical axes must be defined within the framework of a given grammar and operators must, in principle at least, be repeatable.

In summary, our proposal is to analyse a fully parsed corpus of natural language of speech and writing to evaluate evidence that related series of speaker (or author) choices in the

production of language are partially mutually constrained rather than independent, and investigate how these effects may be evidenced along multiple grammatical axes.

Our position is consistent with a view that grammar partly encapsulates the 'trace' of speaker choices. It is not necessary to make claims about a particular *mechanism* by which speakers make these choices or, indeed, as we shall see below, the *order* in which they do so.

Our proposed methodology has psycholinguistic implications. Anderson (1983) refers to the actual results of a psychological process as the 'signature' of the phenomenon, and points out that computer simulations may replicate that signature without exposing the underlying process of cognition. A computer system for generating 'natural language' does not necessarily provide understanding regarding how humans produce language, nor parsers, how we interpret sentences. At best they may help identify parameters of the human analogue. Our proposition is that natural experiments[1] on the results of parsing may help identify parameters of corpus contributors' processes of language production.

2. A worked example: Grammatical interaction between prenominal AJPs

The definition of 'grammatical axes' provided above is rather abstract. Let us consider a simple example. English noun phrases can (in principle) take any number of adjectives in an attributive position before the noun: *the old ship*, *the old blue ship*, etc. (Huddleston and Pullum 2002: 57).[2] By way of exemplifying our method, we will investigate the general proposition that the introduction of one adjective constrains the addition of another.

We wish to refute a null hypothesis of the independence of each decision to add an attributive adjective, i.e., that the chance of preposing a second adjective is the same as the chance of preposing the first, and so on. The assumption that adjectives are not independent, but instead semantically restrict a head noun (and thus each other), is fundamental to any investigation of, e.g., constraints on adjective word order.

Note that identifying that an interaction is taking place between the decisions to insert two

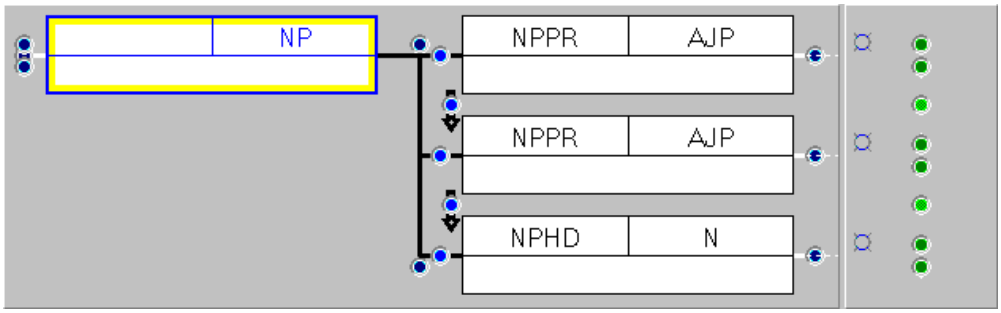

Figure 1: Fuzzy Tree Fragment for *x*=2, retrieving NPs with *at least* two adjective phrases (AJP) preceding the noun head (NPHD, N). For reasons of space the tree is drawn from left to right, with the sentence on the right hand side.

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*	193,135	37,305	2,944	155	7
probability *p*		0.1932	0.0789	0.0526	0.0452
upper bound w^+			0.0817	0.0613	0.0903
lower bound w^-			0.0762	0.0451	0.0220
significance ($p > p^+$)			s-	s-	ns

Table 1: Frequency and relative additive probability of NPs with *x* attributive adjective phrases before a noun head, in ICE-GB.

adjectives *A* and *B* does not in itself demonstrate the order of decision making by a speaker. The decision to insert *A* could be made prior to the decision to insert *B*, vice versa, the decisions may co-occur, or (in the case of writing) lead to later decisions to revise previously-inserted adjectives.

The methodology we describe is, as we shall see, a valid investigatory approach in its own right, and could be a precursor to further individual choice experiments concentrating on, e.g., whether particular classes of adjective in particular positions limit the introduction of other adjectives. However in this paper we will concentrate on what the methodology can tell us about the grammar in the corpus *per se*.

2.1 AJPs with noun heads

Our first task is to collect data. In a part of speech (POS)-tagged corpus we can obtain frequencies of cases of single, double, etc., adjectives followed by a noun, which we do below. In a parsed corpus we can be more precise, limiting our query by the noun phrase (NP) and by counting attributive adjective phrases rather than adjectives alone. This permits us to count cases such as *the old [pale blue] ship* correctly (cf. the retrievability criterion outlined above).

The *British Component of the International Corpus of English* (ICE-GB, Nelson, Wallis and Aarts 2002) is a fully-parsed million-word corpus of 1990s British English, 40% of which is written and 60% spoken. In this paper our results come from ICE-GB Release 2. ICE-GB is supplied with an exploration tool, ICECUP, which has a grammatical query system that uses idealised grammatical patterns termed *Fuzzy Tree Fragments* (FTFs, Wallis and Nelson 2000) to search the trees.

We construct a series of FTFs of the form in Figure 1, i.e., a noun phrase containing a noun head and x adjective phrases before the head. This FTF will match cases where **at least** x AJPs precede the noun (the FTF does not exclude cases with terms prior to the first AJP). This obtains the raw frequency row F in Table 1. Using this information alone we can compute the *additive probability* of adding the x-th adjective phrase to the NP, $p(x)$, as

$$p(x) \equiv F(x) \,/\, F(x\text{-}1).$$

So if there are in total 193,035 NPs in the corpus with a noun head, and 37,305 have at least one attributive adjective phrase, then the probability of adding the first adjective phrase is 37,305/193,035 = 0.1932. We can then compute $p(x)$ for all values of $x > 0$.

Table 1 reveals that the additive probability, p, falls numerically as x increases. We plot probability p with Wilson score confidence intervals in Figure 2. Appendix 1 explains the Wilson interval and the overall approach to analysing sequential decisions step-by-step.

You can 'read' this figure from left-to-right: if a point is outside the upper or lower interval of the next point, the probability has significantly changed with increasing x at that subsequent point. Figures are given in Table 1.

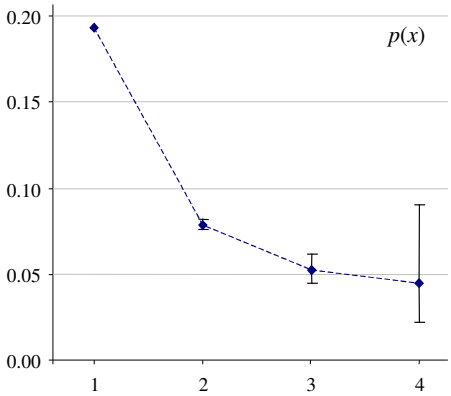

Figure 2: Additive probability $p(x)$ for a 'run' of x AJPs in an NP with a noun head in ICE-GB, with Wilson 'score' confidence intervals

Adding a second adjective phrase to an NP occurs in 1 in 12 (0.0789) cases where a first adjective phrase has been introduced. This contrasts with the introduction of an initial attributive AJP, which occurs in approximately 1 in 5 cases of NPs (0.1932). The difference is comfortably statistically significant (0.0817 < 0.1932).

Adding a third adjective phrase to an NP occurs in 155 out of 2,944 cases where two AJPs had been introduced, i.e., 1 in 19 (0.0526). This fall is also statistically significant (0.0613 < 0.0789). In the final case, adding a fourth AJP to an NP occurs 7 times out of 155 (1 in 22). This has an upper interval of 0.0903, which is greater than 0.0526, and therefore does not represent a significant fall.

Note that the results might also allow the conclusion that the fall in probability over multiple steps is significant, e.g., that the probability of adding a fourth AJP is greater than that of adding the first (0.0903 < 0.1932). However, in this paper we will restrict ourselves to conclusions concerning 'strong' (i.e., successive and unbroken) trends.

The data demonstrates that decisions to add successive attributive adjective phrases in noun phrases in ICE-GB are *not* independent from previous decisions to add AJPs. Indeed, our results here indicate a **negative feedback loop**, such that the presence of each AJP dissuades the speaker from adding another. We reject the null hypothesis that the relative probability is constant for the addition of the second and third successive attributive AJPs.

There are a number of potential explanations for these results, including

- **adjective ordering rules**: adjectives of size before colour, etc.
- **logical semantic coherence**: it is possible to say *the tall green ship* but *the tall short ship* is highly implausible, i.e. adjectives logically exclude antonyms.
- **communicative economy**: once a speaker has said *the tall green ship* they tend to say *the ship* or *it* in referring to the same object.
- **psycholinguistics**: cognitive constraints, such as processing and memory.

Without examining the dataset more closely it would be difficult to distinguish between these potential sources, and indeed multiple sources of interaction may apply at once. Nonetheless, as we shall show, the interaction between adjective phrases modifying the same noun head is a substantial effect which can be identified without recourse to parsing.

In the case of adjective ordering rules, comparative studies with other languages where no strong ordering rules apply might allow us to eliminate this as an explanation. Interestingly we find evidence of a different feedback effect in the case of multiple postmodification of the same head (see Section 4).

It is possible to 'fit' the curve to a power function of the form $f = m.x^k$. Allowing for increasing variance as F falls, the data obtains the function $f \approx 0.1931x^{-1.2793}$ with a correlation coefficient R^2 of 0.9996.[3] This model suggests that probability falls (k is negative) according to a power law. We discuss the implications of modelling with a power law in the conclusion.

This result upholds general linguistic expectations that the use of successive multiple adjectives are avoided due to linguistic or contextual constraints. It does not inform us what these constraints might be, although we may hypothesise about these.[4] Nor does the evidence inform us to the *order* of insertion. These questions are not our principal interest

in this paper. Rather, we have demonstrated evidence of a general trend along the axis of the grammatical analysis of NP constituents.

2.2 AJPs with proper and common noun heads

Readers may object that not all NPs are equally likely to take attributive adjectives. NPs with proper noun heads are less likely to do so than those with common noun heads, as Tables 2 and 3 demonstrate. However, this reinforces, rather than undermines, our observation that the probability of adding an AJP will fall as NPs become longer. A similar argument would apply to other variations in the experimental design, such as removing the restriction that the head be a noun. The method is extremely robust.

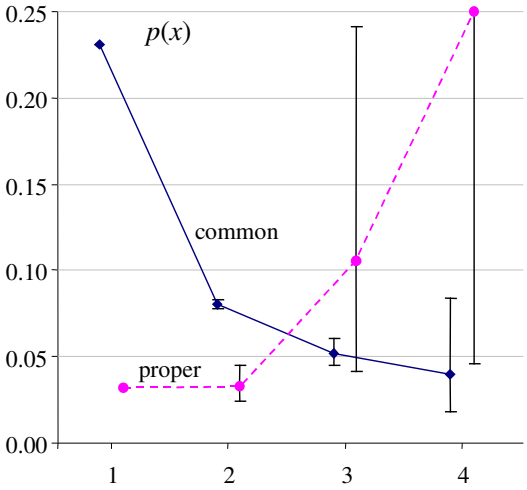

Figure 3: Additive probability *p*(*x*) for adding an AJP to a NP with a common or proper noun head.

The argument goes like this. NPs which cannot take a pre-head adjective phrase (or would rarely do so, e.g., *Long Tall Sally*) are eliminated first. Therefore, were we to focus on common nouns alone (as in Table 2), the proportion of NPs with one AJP or more would increase and the relative decline from this point would become greater.

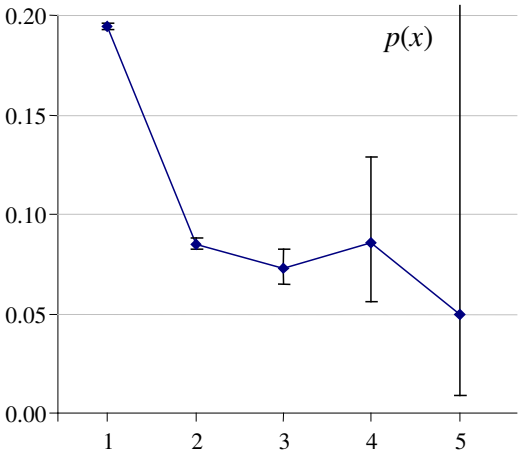

Figure 4: Probability of a 'run' of *x* adjectives before a noun.

The results appear to bear this out. NPs with common and proper noun heads behave differently (Figure 3). Nearly 1 in 4 common noun NPs contain at least one adjective phrase and the probability of subsequent AJPs falls. Against this, almost 1 in 32 proper noun NPs in ICE-GB are analysed as containing one or more AJPs. Indeed, once the decision to add an AJP is made, the proportion appears to *increase*. The result is significant, but with one substantive caveat (see below).

Were this to be substantiated it might represent a different type of phenomenon, e.g., that the use of adjectives are marked (i.e., the speaker is making a deliberate point in using adjectives). Like repetition and *horror aequi*, we cannot assume that all interactions suppress a trend (i.e., involve negative feedback).

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*	155,961	35,986	2,892	151	6
probability *p*		0.2307	0.0804	0.0522	0.0397
significance			s-	s-	ns

Table 2: A significant fall in additive probability for NPs with common noun heads.

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*	36,172	1,143	38	4	1
probability *p*		0.0316	0.0332	0.1053	0.2500
significance			ns	s+	ns

Table 3: Evaluating the *rise* in probability for NPs with proper noun heads only.

***x* adjectives**	**0**	**1**	**2**	**3**	**4**	**5**
'at least *x*' *F*	193,191	37,548	3,187	233	20	1
probability *p*		0.1944	0.0849	0.0731	0.0858	0.0500
significance			s-	s-	ns	ns

Table 4: Variation in the probability of adding adjectives before a noun.

This graph shows that negative feedback identified by a successive falling additive probability is not a universal phenonemon. On the contrary, the proper noun curve in Figure 3 shows that this probability may be constant over successive steps or increase.

A more serious caveat with this data requires further investigation however. This is the practice of ICE annotators to employ a compound analysis of many proper nouns. In just one text, W1A-001, we find a variety of treatments. Compounds include border-line cases such as *Northern England* (#61) and *Roman Britain* (#62) – where *England* and *Britain* could be treated as the head – plus more clearly 'multi-headed' titular compounds such as *the Llandaff Charters* (#80), as well as those analysed adjectivally, such as *the lower Loire* and *a British Bishop* (#83). Reliably counting adjectives requires us to agree what is and is not an adjective. The presence of this observed 'noise' should prompt a review of this aspect of the grammar.

2.3 Grammatical interaction between prenominal adjectives

So far we have used the ICE-GB parse analysis to extract adjective *phrases* within NPs. Let us put the parse analysis aside and replace the FTF queries with a series of simple sequential part of speech queries: '<N>' (single noun), '<ADJ> <N>' (adjective preceding a noun), etc. This is, of course, precisely the type of search possible with a POS-tagged corpus. The result (Figure 4) shows similar evidence of an initial decline (the apparent rise is not statistically significant, see Table 4). How do these results compare with those of our first experiment?

Inspecting the corpus reveals a certain amount of noise. We find 19 cases of a 4-adjective string but only 7 cases with four attributive adjective phrases. There are no cases of five attributive AJPs. The single 'five attributive adjectives' case is *pale yellow to bright orange contents* [W2A-028 #72] where *pale yellow to bright orange* is analysed as a single compound adjective under an AJP. Lexically, it is marked as five adjectives in a compound (including *to*), but only one AJP. Many of the 4-adjective strings are also somewhat unreliable, including:

specious ex post facto justification [S1B-060 #8]	(2 AJPs)
mauvey blue beautiful little thing [S1B-025 #28]	(3 AJPs)
long long long straight roads [S2A-016 #29]	(4 AJPs)
glacial, aeolian, fluvial, marine conditions [W1A-020 #84]	(conjoined)

Of nineteen 4-adjective strings, 3 consist of a single AJP, 4 of two AJPs, 4 of three AJPs and 7 of four AJPs. The two remaining conjoined cases register as a single premodifying AJP containing the conjoin. The variation between the number of lexical adjectives and the number of adjective phrases constitutes a very large amount of classification noise in the data – nearly two thirds of the cases have less than four AJPs! The overcounting of adjective (phrases) explains the result.

This affects the strength of the results. Both AJP and adjective experiments find evidence of a successive significant fall in probability from x=1 to x=3. We can say that the results

are 'cleaner' when parsing is employed, but also that the effect of the interaction between adjective decisions is so strong that it comes through irrespective of whether we employ parsing (section 2.1) or indeed limit the analysis to proper noun heads (2.2).

Thus without parsing, the results are significant for x=3 only at the 0.05 error level, whereas in the first experiment both observed differences were significant at the 0.01 level. Fitting the data to a power law obtains $f \approx 0.1942x^{-1.1594}$ with a correlation coefficient R^2 of 0.9949. The fall is less steep, and the model a little less reliable, than that obtained from the first adjective phrase experiment.

3. Grammatical interaction between preverbal adverb phrases

We identified a general feedback process that appears to act on the selection of attributive adjective phrases prior to a noun, and speculated on the potential causes. Let us briefly investigate whether the same type of effect can be found in adverb phrases (AVPs) prior to a verb. The following are examples of double AVP cases.

rather$_{\text{AVP}}$ *just*$_{\text{AVP}}$ *sing*$_{\text{V}}$ [S1A-083 # 105]

only$_{\text{AVP}}$ *sort of*$_{\text{AVP}}$ *work*$_{\text{V}}$ [S1A-002 #109]

always$_{\text{AVP}}$ [*terribly easily*]$_{\text{AVP}}$ *hurt*$_{\text{V}}$ [S1A-031 #108]

In the second example, *sort of* is treated as a compound intensifier. In the third, *terribly* modifies *easily* rather than *hurt*. Employing the adverb phrase analysis (counting *terribly easily* as a single AVP) should focus our results.

Table 5 summarises the results from ICE-GB. The probability of adding the first and second AVP are almost identical (about 1 in 19). However, at the third AVP the probability falls significantly. The pattern is shown in Figure 5.

Overall, however, this is a null result (power law R^2 = 0.2369), i.e., we cannot reject the null hypothesis that the decision to premodify a verb with an adverb (or adverb phrase) occurs independently from any previous decision.

This underlines the point that the type of feedback we are discussing is not an abstract phenomenon, but arises from **specific** identifiable distributions captured by a sentence grammar. Different repeating decisions may be subject to different sources of interaction. Our observation echoes that of Church (2000), who noted variation in lexical interaction between the reoccurrence probability of 'content' and 'function' words in a text.[5]

Linguists have assumed that preverbal adverb phrases do not interact semantically in a comparable manner to attributive adjective phrases. Taken with the previous experiments, these results appear to imply that semantic interaction effects are likely to be stronger than those attributable to communicative economy (otherwise the adverb phrase additive probability would have fallen from the start). However more research is clearly needed.

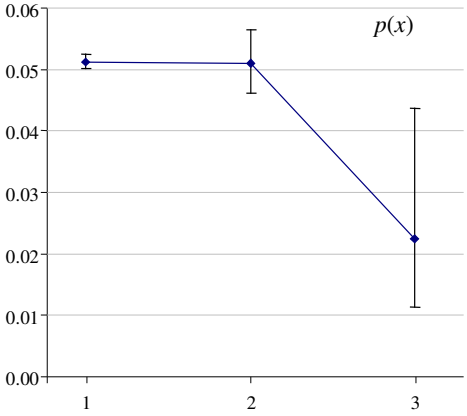

Figure 5: Variation in the probability of adding a preverbal AVP in ICE-GB.

x adverb phrases	0	1	2	3
'at least *x*' *F*	136,554	7,004	357	8
probability *p*		0.0513	0.0510	0.0224
significance			ns	s-

Table 5: Variation in probability of adding AVPs before the verb (same VP).

4. Grammatical interaction between postmodifying clauses

4.1 Embedded vs. sequential postmodification

In addition to adjective premodification of a noun phrase head, the ICE grammar allows for **postmodifying clauses** to further specify it. Such clauses are similar to adjectives in that they semantically specify the noun head, but consist of entire clauses, e.g.,

the Byzantine emperor [*whose face has been somewhat rubbed*] [S1A-064 #83]

In the example above, *whose face* is analysed as the subject NP of the postmodifying clause *whose face has been somewhat rubbed*. In this experiment we will investigate the impact of multiple postmodifying clauses. We consider two phenomena in parallel:

1) **sequential** postmodification where clauses modify the *same* NP head, and
2) **embedded** postmodification, where clauses modify the *prior* NP head.

These two types of multiple postmodification are summarised by the pair of FTFs in Figure 6 below. Since they operate on the same head, we might expect that sequential postmodifying clauses will behave similarly to sequential adjective phrases in an attributive position. In ICE-GB we find cases up to three levels of embedding of postmodifying clauses containing NPs. An example of two-level embedding is below (see Figure 7):

a shop [*where I was picking up some things* [*that were due to be framed*]] [S1A-064 #132]

With multiple postmodification, we have constructions such as

a woman [*called Caitlin Matthews*] [*who is a pagan priestess…*] [S1A-096 #98]

ICE-GB also contains examples analysed as co-ordinated postmodifying clauses, such as

A very very major job [[*relaying floors*] [*redoing all the beams*]] [S1A-009 #86]

Depending on interpretation, cases of this third type may be counted as examples of single or multiple postmodification. We will consider both interpretations in our analysis.

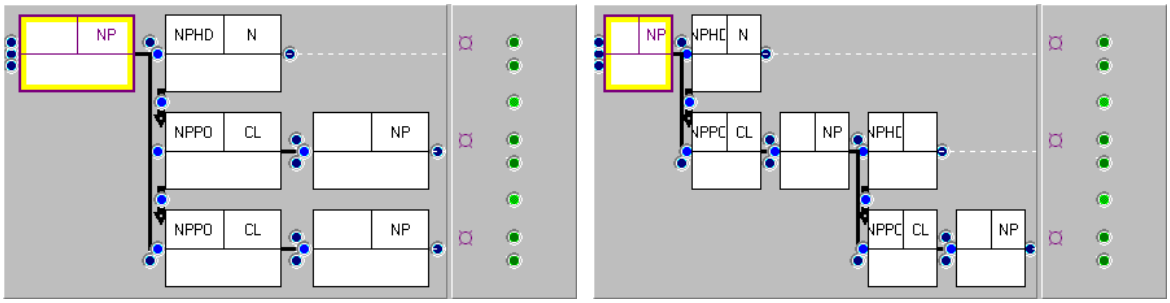

Figure 6: Basic FTFs for finding NPs with 2 postmodifying clauses containing NPs, sequential (left), embedded (right).

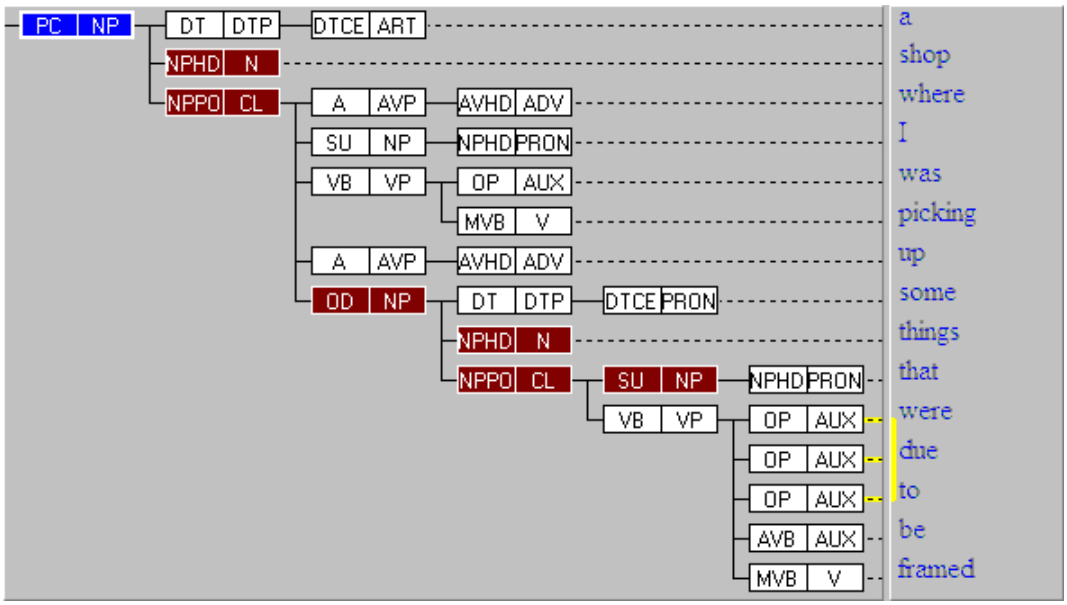

Figure 7: An example of two levels of embedding in ICE-GB (S1A-064 #132).

4.2 Data gathering

Gathering data requires some care. There are three problems.

1) **Matching the same case more than once**. Using a proof method, ICECUP matches every structural permutation, and thus counts each one. Matching the same upper NP, any of the lower NP nodes in the FTF in Figure 6 (right) could match a subject, direct object, etc. However we must only count each case once. The 'next generation' ICECUP IV software[6] allows us to formalise the extraction of samples of data and unifies every FTF to the same point in the corpus. This is less labour intensive than the alternative, which is to review cases and eliminate duplicates manually.

2) **Double-counting subordinate FTFs**. For every tree matching the two-level embedded FTF, the equivalent one-level FTF matches the same tree structure twice – upper and lower. We should not count this second case because it is dependent on the first. To eliminate secondary embedded cases, we simply calculate the *unique* frequency F' by subtraction, i.e., $F'(x) \equiv F(x) - F(x+1)$ (Table 7). Relative additive probability is then recalibrated as $p(x) \equiv F'(x)/F'(x\text{-}1)$ and the standard deviation is adjusted accordingly.

3) **Matching across co-ordination.** Conjoins can interrupt the matching of FTFs to trees in both cases. Any intermediate clause or NP in the structure may (in principle) be co-ordinated. We address the problem of co-ordination within an FTF by matching FTF patterns across co-ordination nodes. In addition to unifying every case to a single point in the corpus tree, ICECUP IV permits a categorical variable, x, to be elaborated with each term {0, 1, 2, 3…} defined by a logical combination of queries. In this case, this is a disjunctive set of possible FTFs with conjoins in each possible position.

4.3 Results and discussion

We obtain the results in Tables 6 and 7. These demonstrate an initial fall in additive probability in both sequential and embedding cases. The probability of adding a following clause falls initially from approximately 1 in 19 to 1 in 61. In the embedding case, the fall from one to two levels of embedding is about 1 in 18 to 1 in 44. However the sequential case also sees a subsequent rise in probability. Applying power law fitting to the embedded pattern obtains $f_{embed} \approx 0.0539x^{-1.2206}$ ($R^2 > 0.99$).

***x* NPPO sequential clauses**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*	193,135	10,093	166	9	
probability *p*		0.0523	0.0164	0.0542	
significance			s-	s+	
'at least *x*' *F* (+coord'n)	193,135	10,093	195	16	2
probability *p*		0.0523	0.0193	0.0821	0.1250
significance			s-	s+	ns

Table 6: A fall and rise over successive sequential postmodifying clauses.

***x* NPPO embedded clauses**	**0**	**1**	**2**	**3**
'at least *x*' *F*	193,135	10,093	231	4
'at least *x*' unique *F'*	183,042	9,862	227	4
probability *p*		0.0539	0.0230	0.0176
significance			s-	ns

Table 7: A fall in probability over two levels of embedded postmodifying clauses.

Results are summarised in Figure 8. The graph also indicates that the probability of embedding a second clause is significantly higher than that of applying a second (sequential) postmodifying clause to the same head.

What is the source of the secondary 'tail' effect of a **rise** in probability for sequential postmodification at x=3?

As we observed, there is a potential ambiguity in the corpus analysis. Double postmodification may be analysed as two independent postmodifying clauses or as a single co-ordinated pair of postmodifying clauses, as in the following example.

…the process [[*how you turn off*]$_{CJ}$ [*how you turn back on*]$_{CJ}$]$_{NPPO}$ [S1-050 #109]

In the upper part of Table 6 we count this as a single case. If, instead, we count each conjoin as if it were an independent postmodifier, this secondary rise is accentuated, indicated by the upper line in Figure 8. This rise does not appear to be a 'blip'. So what is going on?

We hypothesised that sequential postmodifying clauses, operating on the same head, could semantically constrain each other in a manner comparable to attributive adjectives. However, on reflection, the productivity of clauses is such that semantic exclusion is unlikely to play a role.

The initial fall may be due to either a difficulty in 'backtracking' to the same head or communicative economy.

The subsequent rise is not easily explained until we consider the pattern of results obtained by counting each conjoin separately.

This suggests that, far from one clause limiting the next, a clause may actually provide a **template** for the construction of the next. In the previous example, *how you*

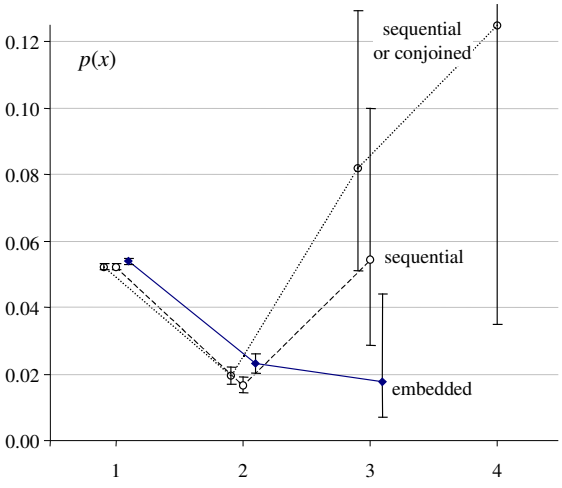

Figure 8: Contrastive patterns for sequential and embedded clauses that postmodify a noun in ICE-GB.

turn off provides an easily modified template for *how you turn back on*.

Evidence of templating is clearly identifiable in cases analysed as co-ordinated (as above), but may also apply in other sequential cases. The first of the following examples are analysed as postmodified by a co-ordinated set of postmodifiers, the second as sequentially postmodified.

> *…his consistent bids* [[*to puncture pomposity*] [*to deflate self-importance*] [*to undermine tyranny*] *and* [*to expose artifice*]] [S2B-026 #81]
>
> *…one path* [*which was marked… on a …map*] [*which is no longer marked*] [S1B-037 #85]

Even where the same construction is not used, we find evidence of lexical repetition in subsequent clauses.

Finally, alternative psychological explanations may be pre-linguistic and not require evidence of explicit templating. It may simply be that the mental referent of the NP head is foregrounded cognitively in the speaker's consciousness, and speakers are simply articulating characteristics of this concept "in the mind's eye". This would seem to imply that switching focus has a cognitive cost and tends to be avoided.

In the case of embedding, clauses apply to different heads, so semantic exclusion, templating or foregrounding do not apply. Instead we see a systematic fall, which may be explained by communicative economy or the cognitive cost of refocusing attention. Compared to sequential postmodification, however, the *initial* decline is not as great. This suggests that initially speakers (or writers) find it easier to postmodify the most recent head than to 'backtrack' to the previous head, which may be due to language production (attention and memory) constraints.

Nonetheless, the embedding case also demonstrates a significant fall in probability, and hence, an interaction. Adding a postmodifying clause containing a further NP comes with a linguistic cost. Furthermore, if a speaker's decision to postmodify an NP takes place in sentence order (i.e., with limited 'look-ahead'), then the first embedding decision may be said to constrain the second. In cases of embedding at least, this ordering assumption seems probable, in part due to psycholinguistic attention and memory constraints on the part of both speaker and audience.

Applying separability meta-tests on pairs of goodness of fit tests (Wallis 2011, see Appendix 1) allows us to conclude that the patterns of embedding and sequential postmodification are significantly separable at both stages (from $x = 1$ to 2 and from 2 to 3). In other words, first, the two curves can be statistically separated over the initial decline seen in Figure 8. Second, the subsequent rise of the sequential feedback effect is also significantly distinct from the corresponding embedded pattern.

4.4 What might these results imply?

The next question is this: what do these results imply for the status of grammar? Clearly, it would be difficult indeed to identify and count cases such as these in a POS-tagged corpus. In order to argue that these results can only be explained by the existence of a grammar framing linguistic choices made by speakers, we need to eliminate alternative hypotheses.

1. **Lexical interaction** (cf. section 2.3). Figure 7 illustrates that embedded terms need not be lexically adjacent. In cases of multiple levels of embedding this is common. Reviewing the results obtained from the embedding experiment finds over 85% of two-

level cases contain at least an intermediate VP, NP or clause between the upper and lower NP heads. We may therefore reject this hypothesis.

2. **Misclassification of embedding**. Could embedded clauses be incorrectly attached to trees? Perhaps additional embedded cases were misclassified as sequential? Reviewing the set of two-level sequential cases reveals that parse analysis errors do not explain the results.

As we have seen, multiple postmodification in sequence obtains a steeper probabilistic fall than multiple embedding from x=1 to 2. 166 cases of double postmodification were found in the one-million word ICE-GB. If these cases were incorrectly attached, then this might increase the total, reducing the fall in probability. Reviewing these 166 cases we identify a maximum of 9 potentially embedded ambiguous cases (we find one incorrectly attached example). 95% are identifiably correct, or simply cannot be embedded.[7]

We may also review the 227 embedded cases. Note however, that were any of these to be misclassified, this would increase the fall due to embedding, rather than undermine it. Many, like the example in Figure 7, are unambiguous. *Things*, rather than *a shop*, must be postmodified by *were… framed*, due to, e.g., number agreement. Even if a small number of these cases were incorrectly attached, this would not refute the claim that the observed interaction on the embedded axis was grammatical in nature.

3. **The effect of segmentation**. A final alternative explanation might be that results are due to segmentation. Writers may edit material into sentences and annotators frequently break transcriptions into putative sentences. If this broke a sequence this would tend to reduce the additive probability at each stage, causing a successive fall.

This hypothesis is perhaps the most difficult to exhaustively disprove by automatic methods or by browsing the corpus. Here we are primarily concerned with annotator decisions in the spoken subcorpus, and with embedding rather than sequential postmodification.

Sequential postmodification or coordination may be underestimated. The following utterances are spoken by the same speaker:

> *Uh my conversation was going in on Saturday morning into a shop where I was picking up some things that were due to be framed | And I went in | And I had a very very busy Saturday* [S1A-064 #132-134]

However, this does not appear to be particularly problematic for our central thesis. Simply adding to a narrative has negligible implications for language processing and as we note, would boost the observed additive probability. If the decisions led to a bias, we would expect that bias to impact on longer constructions to a greater extent than shorter, but in fact we see a rise in additive probability with longer constructions.

To address the problem of embedding we decided to review all cases where the final noun preceded a text unit break. We could also discount cases where the following text unit was spoken by another speaker,

In conclusion, rejecting the null hypothesis that each decision to postmodify the noun phrase is independent from previous decisions leaves us with evidence for our claim that the grammar reflects the language production process – in this respect at least.

Another way of expressing this is the following. *Our study provides empirical evidence for grammatical recursive tree structure as a framework of language production decisions.*

Since the probability of a speaker choosing to embed a postmodifying clause falls with every subsequent decision, we have discovered statistical evidence of an interaction along this embedding axis. This does not mean that this particular grammar is 'correct', rather, that a grammar that represents such structures is required to account for this phenomenon. Indeed, the distinction between cases of multiple postmodification and asyndetic coordination is far from clear, unlike the distinction between multiple postmodification and embedding. However, grammatical models of recursively embedded postmodifying clauses may be said to be empirically supported by evidence rather than axiomatically assumed.

5. Conclusions

At first glance, our initial experiments offer little more than empirical support for general linguistic observations, namely that constraints apply between adjectives, phrases and clauses, but apply weakly, if at all, between preverbal adverb phrases. These constraints are likely to include **semantic coherence** (possibly revealed by clusters of co-occurring adjectives) and **communicative economy**, where (e.g.) adjectives are used sparingly to distinguish elements in a shared context. Semantic coherence may also include a certain amount of idiomatic 'boilerplate' phrasing illustrated by the problem of compound nouns mentioned in section 2.3. Our primary aim in presenting these studies was to exemplify our methodology. We were then able to employ the same approach in the evaluation of sequential decisions to apply embedded and sequential postmodification to noun heads.

In the case of complex constructions such as embedded postnominal clauses, each NP has a different head. Cumulative semantic constraints of the type operating on adjectives are therefore unlikely to explain this interaction. In addition to communicative constraints we might expect to find evidence of psycholinguistic constraints.

In order for linguistics to move forward as an observational science, theoretical insights must be evaluable against data. Chomsky's (1986) distinction between I(nternal)- and E(xternal)-Language allowed linguists to theorise about general principles of grammar. Natural language data, including corpora, are limited to E-Language 'epiphenomena'. However if I- and E-Language are related through a production process, the structure of I-Language may be testable.

Even if 'language data' corpora merely collect external 'performances' this does not belie the implication that *the internal representation must perform*. In other words, in order to evaluate theories of grammar, the underlying grammar (and constraints placed on it by speakers) should be perceivable in uncued language data.

Selecting between and optimising grammatical frameworks on intrinsic or deductive grounds appears to us to be necessary, but one-sided. In the natural sciences such practice would be termed 'philosophy' or 'modelling', rather than empirical science. We believe, therefore, that

i) there is an imperative to develop methodologies for evaluating grammatical frameworks against natural language corpora *extrinsically*, avoiding circularity; and, furthermore,

ii) sizeable parsed corpora provide some of the necessary resources to initiate this process.

5.1 Implications for corpus linguistics

From the perspective of corpus linguistics the methodology proposed is novel. These experiments concern general 'structural' trends between grammatical elements, and we have not here attempted to investigate specific explanations of why speakers might make particular individual choices. Such an 'individual choice experiment' may start with a broad research question such as the following.

> *Given that a speaker may choose between two or more alternative ways of expressing a particular concept, what factors influence that choice?*

Examples of this type of research abound in, e.g., Nelson *et al.* (2002). Individual choice research often requires a narrow focus, because the influencing factors may be highly specific in order to identify a genuine choice (true alternates). These experiments may also require additional annotation. For example, to investigate the factors influencing the choice of an adjective expressing height or age, it may be necessary to first classify adjectives and nouns into broad semantic classes.

Lexical collocation approaches essentially the same problem in a data-driven way. Identifying a common set of lexical patterns does not demonstrate a semantic constraint but might allow semantic constraints, *pace* Sinclair (1987), to 'emerge'.

We might term these experiments Level I research, addressing low-level and specific questions. Our experiments here are at a higher level of abstraction, summarising the total pattern of interaction at the level below. They might 'frame' or parameterise Level I research, in a similar way as they triggered our review of sentences. Discovering that terms *A* and *B* interact does not tell us why – this is a Level I research question.

Level II observations of the type described in this paper concern patterns of interaction over multiple choices. Experiments 1 and 2 demonstrate that the decision to include a further adjective phrase within an NP is influenced by previous decisions to do so, and this trend is consistent across any number of adjective phrases. This is empirical evidence of a general phenomenon or feedback loop in language production. Experiment 2 suggests that this feedback differs for NPs with common and proper noun heads.

Experiment 4, on preverbal adverb phrases, demonstrates that this observed interaction is not simply a consequence of the contingent nature of language, but arises from *particular* processes. Some speaker choices are less restricted than others.

We believe that we have also shown that Level II observations have a further value, namely that they can be used as a building block towards the evaluation of grammar.

What we might ambitiously call Level III research concerns the validation of grammar itself. Experiments 1 and 3 appear to indicate that the chosen grammar has a benefit in more reliably capturing a general coherent trend partially obscured in the lexical sequence. Section 4 identifies a result over lexically separate units that can only be explained by the 'deep' grammatical property of embedding.

5.2 Towards the evaluation of grammar

Comparing grammars requires us to compare patterns of variation in the **relative probability** of linguistic phenomena captured across different grammatical representations.

Note that our evaluations throughout this paper have been both analytical – evaluating patterns in abstracted data and concrete – carefully reviewing clauses and context that support the results. It would be a mistake to infer that grammar may be evaluated by statistical generalisation alone. We must be capable of attesting to the consistency of the annotation from which we abstracted our data, and the 'reality' of distinctions made.

Figure 8 contains results of multiple experiments. Since probabilities are by definition normalised, they may be visualised on the same vertical axis and compared. Data are obtained from the same corpus. The graph summarises the effect of three trends: (a) embedded postmodification, (b) sequential postmodification where coordinated clauses count as one clause, and (c) sequential postmodification or coordinated postmodifiers. We may inspect this graph to make further conclusions, employing an appropriate separability meta-test (see Appendix 1) to compare each slope.

Employing this test we find that (a) the embedding pattern is statistically separable from (b) but that (b) is not statistically separable from (c). Whether or not we consider coordinated postmodification as consisting of single or multiple serial postmodifiers does not lead to a significantly different result.

In Section 2 we saw how employing the parse analysis could be said to bring results into clearer focus. However the successive decline of adjectives is such that we would expect to find comparable results without a parsed corpus. Lexical proximity and counting words are sufficiently reliable in the case of attributive adjectives.

By contrast, the same cannot be said for Section 4. We detect distinct patterns of interaction between axes of embedding and multiple postmodification. The simplest explanation is that the grammatical structure in the model reflects an aspect of the production process.

Co-ordination may be an important grammatical axis in its own right. Intuitively, co-ordination should exhibit two types of interaction: a strong interaction between conjoined terms (based on semantic coherence), and an interaction between each term and other components of the host clause or phrase. These interactions are implied by the principle of replacement. Thus in the following, *learn* could replace *learn and develop*, so *learn* (and indeed, *develop*) should interact with both *opportunity* and *physical skills*.

> …the opportunity *to learn and develop* physical skills [S1A-001 #34]

However, some phrase structure representations of co-ordination (including ICE) do not represent this principle well. As we have already seen in Section 4, the ICE grammar introduces additional nodes bracketing conjoined terms. The analysis of *to learn and develop* differs from that of *to learn* by the presence of a superordinate co-ordinating node grouping the infinitive verbs. Moreover, *to learn and to develop* is analysed differently (two conjoined *to*+infinitive patterns) from *to learn and develop* (*to* + a conjoined predicate group). There are arguments for both representations. However these rationales are based on general theoretical principles, not empirical evidence.

Recall our opening remarks. A simple 'retrievability of event' test predicts that the optimal representation is simply that which is the most consistent while distinguishing distinct phenomena. But which distinctions are important?

Our solution is to argue that grammar should aim to consistently capture patterns of interaction. Different current representations may be capturing genuinely different

phenomena, i.e., those with distinct patterns of interaction. Grammatical interaction research allows us to evaluate this question.

In conclusion, the evaluation of interaction along grammatical axes is an important methodology for the empirical verification of grammar. We supplement a methodology of contrasting grammars based on the retrievability of linguistic **events** with one based on the retrievability of linguistic **interactions**. From this point of view, a scientifically 'better' grammar is a better theory: one whose *explanatory power* can be shown to be empirically greater – reliably explaining more results, or enabling us to capture more phenomena – than another.

5.3 Have we arrived at a non-circular 'proof of grammar'?

There is a relationship between the retrieval of single lexico-grammatical events and the retrieval of patterns of interaction. Ultimately, the second is dependent on the first, although the investigation of general trends may be more 'forgiving' of occasional analytical error than that into specific classes of event.

Contrasting linguistic interaction patterns in a grammar does not suffer from quite the same methodological weaknesses (circularity, redundancy and atomisation, see Section 1) as event retrievability.

- **Atomisation.** A pattern of linguistic interaction reflects the structure in addition to individual terms in the analysis. In addition to relying on event retrieval, the identification of interaction depends on cohesive 'system' principles of the grammar. Moreover, multiple interactions may impact on the same class of event.

- **Redundancy**. Redundant terms in grammatical structure will tend to mislead measures of grammatical distance. This 'system' requirement makes it difficult to introduce redundant terms (cf. co-ordination, above) or relations without impacting on the retrievability of multiple trends. The principle of Occam's razor applies (only add elements that explain a detected interaction otherwise not explicable). This may potentially be operationalised by entropy (Shannon 1948).

- **Circularity**. It is true that there is no escape from a hermeneutic cycle of annotation (one must define an NP, adjective, etc., prior to retrieval). This is not a critical weakness, however. The requirement to *a priori* definition of 'auxiliary assumptions' is common to scientific method.

Puttnam (1974) points out that all sciences are based on auxiliary assumptions, just as all experiments are based on instruments. The key to avoiding circularity is to ensure that assumptions are of a lower order than observations and that assumptions can be tested (instruments calibrated) in turn. Hypotheses should ultimately be 'triangulated': supported by multiple independent sources of evidence.

Treating a grammar as a structural theory evidenced by linguistic interaction implies that bracketing terms (ensembles of named relations, clauses, phrases, etc.) which fail to show evidence of linguistic interaction within them would be open to challenge. This is particularly the case, for example, if there is stronger evidence of linguistic interaction across a bracket boundary than between terms enclosed within.

Our approach is similar to other 'secondary observation' approaches in other fields. For example, physicists cannot 'see' sub-atomic particles directly. Particles travel at extremely

high speeds and many are smaller than photons. However, numerous particles have been identified by experiments that render a trace of the path of the particle visible. (Each particle makes a line of bubbles in a tank of a transparent gel as it travels. The tank is placed within a strong magnetic field to separate particles by mass, velocity and charge.) Likewise, Mendelian genetics predated the discovery of mechanisms of evolutionary inheritance. In our case, we cannot 'see' language production constraints directly, and lab experiments designed to isolate them may be over-constrained. However, by adopting a particular standpoint – that of observing grammatical interaction – we may perhaps observe their imprint on our utterances.

The approach outlined here is also related to bottom-up parsing and automatic part of speech tagging. These algorithms exploit the fact that language data reflects the contingent nature of language production (the fact that one decision leads to another). Consequently adjacent terms often statistically cluster together (they interact). Algorithms store banks of terms and probabilities captured from corpus data and attempt to apply these to new text.

Note, however, a crucial distinction. Rather than identify phrases from patterns of tags, or select a POS-tag by exploiting patterns of co-occurring lexical items, we took the noun phrase structure (for example) as given and then explored patterns of interaction within the NP.

This demonstration of the explanatory power of a particular grammar can be seen as positive evidence for the existence of grammatical structure, i.e., a pattern of speaker decisions to construct a sentence possessing a recursive tree structure. The failure of statistical NLP parsers to completely analyse natural language might be seen as negative evidence of the same phenomenon – negative, naturally, because it is potentially refutable by the development of an improved algorithm.

5.4 Further applications

In addition to optimising and improving grammars by using treebank data, a number of further applications for our methodology have been identified.

Case interaction (Wallis 2008) is the problem, simply stated, that cases obtained from continuous text are not strictly independent. Suppose that cases of adjective phrases are sampled exhaustively from the corpus. We saw in Experiment 1 that AJPs interact with each other. The problem of modelling case interaction optimally is to estimate the degree to which each case interacts with the others in the same sample, and scale each datapoint appropriately. Usually corpus linguists skip over this problem or subsample to reduce the standard error estimate. However, more efficient solutions are possible through modelling.

Random sub-sampling is sub-optimal because information is lost. Note also that not all adjacent terms interact, or indeed interact to the same extent, as Experiment 4 demonstrates. The level of interaction is term-dependent (cf. Church 2000). The optimal solution therefore involves measuring the extent of interaction between cases and weighting each datapoint by its relative independence, before investigating interaction between variables.

Our research into grammatical interaction reveals patterns of co-dependence between generalised terms in the grammar, measurable along grammatical axes. The robustness and generality of the method are such that it should be possible to **compare corpora of different languages** for evidence of the same or comparable constraints (for example, some languages, such as modern French, accept attributive adjectives in the postnominal

position). It should also be possible to measure diachronic grammatical change, e.g., comparing early modern English to present day English.

A major future challenge is to **investigate the language production process** more closely. The idea is to contrast collections of parsed orthographically transcribed spontaneous speech with other forms of speech data (e.g., rehearsed speech, public readings) and writing. Currently, we have over half a million words of fully-parsed spontaneous speech of educated adults' English, in both ICE-GB and a companion corpus, the *Diachronic Corpus of Present Day Spoken English* (DCPSE). If suitable data sources were compiled and parsed comparably to ICE-GB, applications in, e.g., speech therapy research and childhood language acquisition suggest themselves.

Acknowledgments

Thanks are due to Bas Aarts, Joanne Close, Stefan Th. Gries, Kalvis Jansons, Evelien Keizer, Gerry Nelson and Gabriel Ozón for comments on versions of this paper. Although the methodology is relatively novel, the perspective is based on discussions about empirical linguistics with linguists over several years, and there are ultimately too many people to thank for their contributions. The development of ICECUP IV was funded by ESRC grant R000231286. Research into case interaction was carried out within this project and attempts to measure interaction led to the experiments here.

Ultimately the ability of researchers to generalise from analysed corpora is based on the quality of that analysis. It is therefore to the body of hard-working linguists who construct and correct our corpora that this paper is dedicated.

Appendix 1: Analysing sequential decisions

In this paper we perform a relatively unusual analysis problem. We investigate phenomena consisting of sequential decisions, where a speaker/writer is free to choose to add a term to a construction (and be free to add another, recursively) – or stop. As there is no off-the-shelf method that can be simply applied to this problem, and as it is central to the argument in the paper, it seems necessary to explain the solution from first principles.

The observed relative probability of adding the x-th term to a construction is simply

$$p(x) \equiv F(x) \,/\, F(x\text{-}1),$$

where $F(x)$ represents the frequency distribution of all constructions of at least x terms. To investigate what is happening to $p(x)$ we need to ask what is the expected distribution of $p(x)$? Second, how might we test for significant difference from this expected distribution?

A1.1 The expected distribution of $p(x)$

A 'run' of independent chance events follows a **geometric distribution**.[8] Perhaps the simplest example of a geometric distribution is as follows. If we toss a fair coin 0, 1, 2, 3,… times and obtain only heads, the probability of this event will be distributed (1, 0.5, 0.25, 0.125,…). This distribution follows $P = pr(x \text{ heads}) = p^x$ with relative probability $p = 0.5$.[9]

The key property is the following. If each event is independent from the next, then the probability of the event occurring will be **constant** ($p = pr(1 \text{ head}) = 0.5$). Conversely, if the probability varies across successive events, we may refute this null hypothesis and conclude that events are not independent.

The raw frequency distribution F in Table 1 is plotted in Figure 9 (upper). The distribution appears to correspond to a geometric distribution. The frequency appears to decay exponentially.

However appearances can be deceptive. When we plot the relative probability of adding each AJP, p (Figure 9, lower), we find that p falls as successive AJPs are added.

In every step $p(x)$ represents the proportion of 'survivors', $F(x)$, out of those in the previous round, $F(x\text{-}1)$. The expected distribution is that $p(x)$ is constant, so each decision is at least uncorrelated with the next (ideally, independent[10]).

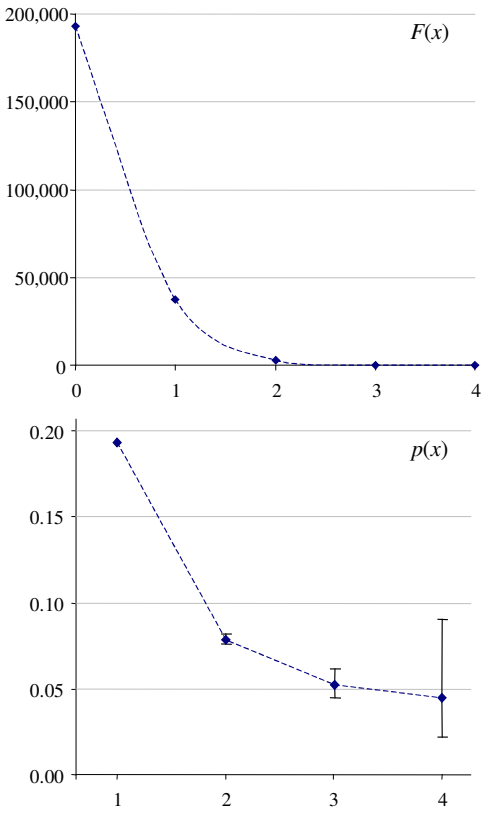

Figure 9: Frequency $F(x)$ and relative probability $p(x)$ for a 'run' of x AJPs in an NP with a noun head in ICE-GB

A1.2 Confidence intervals on $p(x)$

Our observations $p(x)$ are each supported by very different amounts of data. The first observation, $p(1)$, is supported by nearly 200,000 NPs. The last, $p(4)$ is based on a mere 155.

The more data we have the more confident we can be in the observation. The standard way of plotting this varying degree of certainty is to plot *confidence intervals*. A large interval means a greater degree of uncertainty regarding the observation.

By far the most common method of estimating error margins involves a further assumption, namely that the distribution about the mean, $p(x)$, is approximately Normal. This holds for large samples where p is not close to 0 or 1. The formula for the Normal ('Wald' or Gaussian) interval approximation is

$$\text{Wald}\ (e^-, e^+) \equiv z_{\alpha/2}\sqrt{p(1-p)/n}\ ,$$

where $p = p(x)$, $n = F(x)$, and the constant $z_{\alpha/2}$ represents the critical value of the standardised Normal distribution at appropriate two-tailed percentage points (for $\alpha = 0.05$, $z \approx 1.95996$; for $\alpha = 0.01$, $z \approx 2.57583$). The idea is that an observed probability should fall outside the range (e^-, e^+) no more than 1 in 20 occasions by chance.[11]

However, in our data $p(x)$ can indeed be very close to zero. As we have already seen, the sample size, $F(x)$, falls rapidly. As p approaches 0, error bars become skewed, as shown in Figure 9. A more correct confidence interval estimate is known as the *Wilson score interval* (Wilson 1927), and may be written as

$$\text{Wilson}\ (w^-, w^+) \equiv \left(p + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \right) \Bigg/ \left(1 + \frac{z^2_{\alpha/2}}{n} \right).$$

This statistic deserves to be much better known, especially among linguists used to plotting skewed probabilities. Newcombe (1998a) shows that the Wilson interval is a much more

precise interval than the 'Wald' approximation, which, he says, should simply be 'retired'. See also Wallis (forthcoming).

A1.3 Significance testing changes in $p(x)$

The null hypothesis is that the probability of the speaker/writer choosing to include an additional adjective phrase is independent of the number of adjective phrases previously included, i.e. p is constant between decision x and x+1, $p(x) \approx p(x+1)$.

The simplest approach if we are plotting Wilson intervals is to test if the expected value, $p(x)$ is outside the confidence interval for x+1. This test, sometimes referred to as 'the z test for a population proportion' (Sheskin, 1997: 118), can be readily employed with any interval, including the Wilson score interval above. A second method, which obtain the same result (see Wallis forthcoming), is to employ a 2 × 1 'goodness of fit' χ^2 test, where the total frequency is $F(x)$. This can be laid out as follows.

	Expected **E**	Observed **O**
	x	x+1
true	$p(x)$ $F(x)$	$p(x+1)$ $F(x)$
false	$(1 - p(x))$ $F(x)$	$(1 - p(x+1))$ $F(x)$
TOTAL	$F(x)$	$F(x)$

Both methods test variation found in the more specific observation, that is, while we also know that the probability $p(x)$ is subject to variation, we are interested in if $p(x+1)$ differs from $p(x)$. We simply want to know whether the probability of further addition has changed.

A1.4 Identifying significant difference between patterns of change in $p(x)$

One further useful analytic step should be discussed. Wallis (2011) describes a series of 'separability tests' which allow researchers to compare the results of two structurally similar experiments. One reason for doing this may be to investigate whether one set of results is significantly different to those found with a different dataset or a different permutation of the same basic experimental design. The separability test used depends on the original test that was employed.

The optimum method for our purposes employs the Wilson score intervals (at an error level $\alpha = 0.05$) already obtained for the sub-observations, which we might write as $p_i(x+1)$, where $i \in \{1, 2\}$ represents the set of test results.

The method consists first of calculating the difference between the two observations $d_i(x+1) = p_i(x+1) - p_i(x)$ for each test, and the Wilson score intervals for $p_i(x+1)$ centred on zero (i.e. we rewrite $e_i^+ = w_i^+ - p_i$, etc.). If the difference exceeds this interval then it is significant. In Section 4 we compare goodness of fit test results (see above) from Tables 6 and 7. The data is summarised below.

	sequential[†]			**embedded**				
***x* NPPO clauses**	**1**	**2**	**3**	**1**	**2**	**3**	**1-2**	**2-3**
F	10,093	166	9	9,862	227	4		
probability *p*	0.0523	0.0164	0.0542	0.0539	0.0230	0.0176		
difference *d*		-0.0358	0.0378		-0.0309	-0.0054	-0.0050	0.0432
upper interval width e^+		0.0027	0.0456		0.0031	0.0265	0.0039	0.0367
lower interval width e^-		0.0023	0.0254		0.0027	0.0107	-0.0038	-0.0468

[†]**Note:** In the table, sequential cases count coordinated terms as a single postmodifier.

In this form it is simple to convert these formulae to a meta-test to compare these results (final two columns). We take the difference of differences, $d_1 - d_2$, and calculate a new combined interval using a Pythagorean formula combining the upper interval width of test 1 with the lower interval width of test 2, and vice-versa.

If the following inequality holds the difference $d_1 - d_2$ is within expected bounds and the test is non-significant.

$$-\sqrt{(e_1^+)^2 + (e_2^-)^2} < d_1 - d_2 < \sqrt{(e_1^-)^2 + (e_2^+)^2} .$$

In this case this obtains a significant difference of differences for both first and second stages (-0.0050 < -0.0038; 0.0432 > 0.0367). On the other hand different methods for counting sequential coordination do not obtain significantly distinct results. Finally, it is also possible to compare pairs of successive falls or rises in the same pattern with this method. In this case we find what we might expect: that the pattern changes over subsequent changes in $p(x)$.

A spreadsheet for computing this test is also available at www.ucl.ac.uk/english-usage/statspapers/2x2-x2-separability.xls.

Notes

1. Experiments on corpus data are not 'true experiments' in the sense that conditions are not manipulated by researchers and therefore results are likely to be indicative. Sheskin (1997: 18) calls this a 'natural experiment'. Corpus linguistics inevitably consists of *ex post facto* studies of previously collected data, where manipulation of experimental conditions is viable only at the point of data collection. The benefits of a corpus approach include ecological soundness or 'naturalism'. This does not rule out conclusions from observations, but it may limit the refutation of specific explanations. We would hope that such corpus linguistics 'experiments' prove to be complementary to true laboratory experiments.

2. Some linguists, including Huddleston and Pullum (2002: 555), have distinguished between restrictive uses of adjectives, where the adjective defines a **subset**, and non-restrictive uses, where it defines a **characteristic** of the set. This distinction is not particularly relevant here: in either case, semantic and communicative constraints between adjectives will likely equally apply, regardless of their relationship with the noun head.

3. We use the method of least squares **over the variance**, i.e. optimising m and k by successive iteration, minimising $SS_{err} = \Sigma((f_i - y_i)^2 / s_i^2)$ for $f = mx^k$.

4. The action of semantic constraints would predict a consistent effect over x, whereas the communicative economy hypothesis might predict a sharp initial drop from x=0 to x=1 and a distinct curve for x>1. If the power law observation is found elsewhere this tends to support the semantic hypothesis. However, it is only through careful examination of texts for topic noun productivity that this question might be definitively answered.

5. Church (2000) showed that **lexical** items interact to varying degrees by a different method: split texts into two parts (**history** and **test**) and compare the probability that a word will occur in the second part if it occurred in the first ($pr(w \in \textbf{test} \mid w \in \textbf{history})$, termed 'positive adaptation') with the baseline probability of the word occurring in the second ($pr(w \in \textbf{test})$). He finds that not all words behave in the same way, with semantically unambiguous 'content words' such as *Noriega* (potentially a topic noun) increasing the probability of their reoccurrence the most, and 'function words' the least.

6. ICECUP IV is available from www.ucl.ac.uk/english-usage/projects/next-gen.

7. A close examination of the 166 cases of double postmodification finds 9 'problematic' cases which could be embedded. Others may be disambiguated by semantic or syntactic constraints, e.g., the first clause of 'declarative' cases name an individual or thing. These tests are not necessarily exclusive. Only 7 (4%) rely on context beyond the current clause for disambiguation.

declarative	25 (16%)	*this girl [called Kate][who's on my course]* [S1A-038 #20]
is/exists, etc.	16 (10%)	*the other thing [that's marvellous][I started doing for singing]* [S1A-045 #28]
anthropomorphic	25 (16%)	*people [who are already studying dance][that... found contact work]* [S1A-002 #150]
abstract vs. concrete	16 (10%)	*Master's courses [that are... a year...][that uh ... related to them]* [S1A-035 #131]
abstract vs. process	3 (2%)	*the necessity [it seems to have][to tell these stories]* [S1B-045 #81]
pronoun head	27 (17%)	*the work [that I'm now doing][which involves disabled people]* [S1A-004 #85]
relative pronoun	10 (6%)	*so much work [that's got to be done][that we won't have time...]* [S1A-053 #12]
number agreement	11 (7%)	*other things [that came out of the design][that were also important...]* [S1B-020 #15]
repetition	4 (3%)	*horrid dresses [that they had][which they had like a...shawl...]* [S1A-042 #320]
explicit reference	2 (1%)	*there was a second accident [involving the rear of the vehicle...][which was described as a much less violent accident]* [S1B-068 #87]
punctuation	1 (1%)	*governments [(that of Britain prominent among them)][which have forces ranged against Saddam]* [W2E-009 #52]

context beyond clause problematic	7 (4%)	*immensely Christian gentleman [as ever chiselled anybody out of five cents][who taught his Sunday school class in Cleveland]* [S1B-005 #176]
	9 (6%)	*my colleague [who's a pensioners' worker][who isn't doing this…]* [S1A-082 #102]

8. Sometimes the geometric distribution is referred to as the 'exponential distribution'. While these distributions are mathematically closely related, the term 'exponential distribution' is properly reserved for a continuous distribution (i.e., a smooth line rather than a series of events).

9. This probability distribution matches an idealised frequency distribution. Were we to sample 'coin toss runs' n times and count the frequency of sequences, we would obtain a frequency distribution tending towards nP, where P is the cumulative probability of at least x heads being thrown. See also Wallis (forthcoming).

10. Strictly, each individual NP in a corpus may not be independent from one another and may interact, see Wallis (2008) and Section 5.4. We demonstrate this property for embedded NPs in Section 4. We assume here that the degree to which NPs interact overall is small, which is a reasonable assumption except in cases of embedding, where we eliminate duplicates.

11. This is a bi-directional or 'two-tailed' confidence level, based on $z_{\alpha/2}$, which is a little conservative. If we had made an *a priori* prediction of falling p we might have used a one-tailed test with a slightly lower value of z_{α}. However, as we find elsewhere, it is conceivable that successive probability could rise.

References

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

ABEILLÉ, A. (ed.) 2003. *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer.

ANDERSON, J.R. 1983. *The Architecture of Cognition*, Cambridge, MA: Harvard University Press.

BÖHMOVÁ, A., HAJIČ, J., HAJIČOVÁ, E., & HLADKÁ, B. 2003. The Prague Dependency Treebank: A Three-Level Annotation Scenario, in Abeillé, A. (ed.) 2003. 103-127.

CHURCH, K.A. 2000. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2. *Proceedings of Coling-2000*. 180-186.

CHOMSKY, N. 1986. *Knowledge of language: Its nature, origin and use*. New York: Praeger.

FANG, A. 1996. The Survey Parser, Design and Development. In Greenbaum, S. (ed.) 1996. 142-160.

GREENBAUM, S., & NI, Y. 1996. About the ICE Tagset. In Greenbaum, S. (ed.) 1996. 92-109.

GREENBAUM, S. (ed.) 1996 *Comparing English Worldwide*. Oxford: Clarendon.

HUDDLESTON, R. & PULLUM, G.K. (eds.) 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

KARLSSON, F., VOUTILAINEN, A., HEIKKILÄ, J., & ANTILLA, A., (eds.) 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text. Natural Language Processing* vol. 4. Berlin: Mouton de Gruyter.

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

MARCUS, M., KIM, G., MARCINKIEWICZ, M.A., MACINTYRE, R., BIES, M., FERGUSON, M., KATZ, K., & SCHASBERGER, B. 1994. The Penn Treebank: Annotating Predicate Argument Structure. *Proceedings of the Human Language Technology Workshop*. San Francisco: Morgan Kaufmann.

MORENO, A., LÓPEZ, S., SÁNCHEZ, F., & GRISHMAN, R. 2003. Developing a Spanish Treebank, in Abeillé, A. (ed.) 2003. 149-163.

NELSON, G., WALLIS, S.A. & AARTS, B. 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.

NEWCOMBE, R.G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**: 857-872.

NEWCOMBE, R.G. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**: 873-890.

OFLAZER, K., SAY, B., HAKKANI-TÜR, D.Z. & TÜR, G. 2003, Building a Turkish Treebank, in Abeillé, A. (ed.) 2003. 261-277.

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

QUIRK, R., GREENBAUM, S., LEECH, G., & SVARTVIK J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

SHANNON, C.E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**: 379-423.

SHESKIN, D.J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Fl: CRC Press.

SINCLAIR, J.M. 1987. Grammar in the Dictionary. In Sinclair, J.M. (ed.) 1987. *Looking Up: an account of the COBUILD Project in lexical computing*. London: Collins.

WALLIS, S.A. & NELSON, G. 1997. Syntactic parsing as a knowledge acquisition problem. *Proceedings of 10th European Knowledge Acquisition Workshop*, Catalonia, Spain, Springer Verlag. 285-300.

WALLIS, S.A. & NELSON, G. 2000. Exploiting fuzzy tree fragments in the investigation of parsed corpora. *Literary and Linguistic Computing* **15**:3, 339-361.

WALLIS, S.A. 2003. Completing parsed corpora: from correction to evolution. In Abeille, A. (ed.). 61-71.

WALLIS, S.A. 2008. Searching treebanks and other structured corpora. Chapter 34 in Lüdeling, A. & Kytö, M. (ed.) 2008. *Corpus Linguistics: An International Handbook*. Handbücher zur Sprache und Kommunikationswissenschaft series. Berlin: Mouton de Gruyter. 738-759.

WALLIS, S.A. 2011. *Comparing χ^2 tests for separability*. London: Survey of English Usage. www.ucl.ac.uk/english-usage/statspapers/comparing-x2-tests.pdf

WALLIS, S.A. forthcoming. *z*-squared: the origin and use of χ^2. *Journal of Quantitative Linguistics*.

WILSON, E.B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.