Investigating the additive probability of repeated language production decisions

Sean Wallis

Survey of English Usage, University College London[†]

In this paper we introduce an experimental paradigm based on probabilistic evidence of the interaction between construction decisions in a parsed corpus. We use a one million-word corpus of English annotated with a phrase structure analysis, ICE-GB. We find an interaction between attributive adjective phrases in noun phrases with a noun head, such that the probability of adding adjective phrases falls successively. Preverbal adverb phrases do not exhibit the same interaction. Systemic decline in additive probability is not a universal trait, but characteristic of particular production processes.

Examining noun phrase postmodifying clauses, we find a similar initial fall in the probability of successive clauses modifying the same head, and embedded clauses modifying new heads. Successive postmodification reveals a secondary phenomenon of an increase in additive probability in longer sequences. We argue these results can only be explained as cognitive and communicative natural phenomena acting on and within recursive grammar rules.

Keywords: grammar, additive probability, language production, parsed corpora, grammar evaluation

1. Introduction

Parsed corpora of English have been available to linguists for three decades, from the publication of the *Lancaster-Leeds Treebank* (Garside, Leech and Sampson, 1987) and the rather larger *University of Pennsylvania Treebank* (Marcus *et al*., 1993) onwards. Parsed corpora in numerous languages have been available for well over a decade (see, e.g. Abeillé, 2003). A parsed corpus, or 'treebank', is a corpus where every sentence is fully grammatically analysed in the form of a tree according to a given framework. Such corpora

[†] The development of ICECUP IV was funded by ESRC grant R000231286, which made the research in this paper possible. An earlier version of this paper was published online at https://corplingstats.wordpress.com/ 2012/12/04/linguistic-interaction.

have a number of applications including training automatic parsers, acting as a test set for text mining, or providing a source for teaching and exemplification.

Within corpus linguistics, the epistemological status of a corpus grammar is, however, more uncertain, evidenced by the plurality of grammatical frameworks adopted by linguists. Which grammar should one choose, and what are the consequences of a decision? Motivations cited by corpus builders include simplicity and ease of annotation (Garside and Leech, 1991); grammatical tradition and knowledge (Greenbaum and Ni, 1996); computational purposes, such as information extraction (Marcus *et al*., 1994) or parser evaluation (Carroll *et al*., 2003); or consistency with a previously-adopted standard.

These justifications often lead to a hermeneutic cycle, such as when a probabilistic parser is trained on the framework employed in the training set (Garside, Leech and Sampson, 1987; Fang, 1996). The experimental analogue of this – defending a framework on the basis that it permits us to retrieve phenomena captured by the framework – is also extremely common.

But if we do not know that any given grammatical framework is 'correct', why should linguists commit to it to parse a corpus and carry out research? Sinclair's (1987) response was simple: we should not. However, all scientific research inevitably requires 'auxiliary assumptions', i.e. assumptions which facilitate scientific practice, such as the accuracy of measuring equipment, but are not evaluated simultaneously with research hypotheses (Wallis, forthcoming). From this perspective, a parsed corpus can be thought of as containing a system of auxiliary assumptions in the form of a grammatical framework applied to sentences.

A parsed corpus is a source of three principal types of evidence (Wallis, 2014). First, applying an algorithm to a broad range of text samples provides *frequency* evidence of known phenomena described by the parser knowledge base. Manual correction and completion of parser output improves the accuracy of frequency evidence and supplements it with a second type of evidence: enhanced *coverage* ('factual' or 'discovery') evidence, such as identifying previously unknown rules.

The third type of evidence is central to this paper. A parsed corpus is a rich source of evidence of lexical and grammatical *interaction* (also referred to as statistical association or co-occurrence). At the risk of stating the obvious, as humans form utterances they make a series of conscious and unconscious decisions: to use one word, phrase, etc., rather than alternatives. These decisions are rarely wholly independent from each other (i.e., they 'interact'). In this paper we will demonstrate an experimental paradigm for studying

repeating construction decisions, and then consider the implication of this class of evidence for the evaluation of grammars, i.e. consider the effect of this evidence on our auxiliary assumptions.

This paper is organised as follows. The remainder of Section 1 discusses the divergence of syntactic frameworks, and what independent meta-criteria might be used to decide between them. We propose a criterion that goes beyond the retrieval of given terms within a framework. This is exemplified by the method demonstrated in this paper, namely examining the distribution of the probability of making a decision to add a construction over successive applications of the same addition rule. In Section 2 we demonstrate our method with a simple example, namely adjectives in attributive position in a noun phrase, and compare the effect of different variants of the same experiment. In Section 3 we take a different construction – preverbal adverbs – and find that the trend we saw in Section 2 is not replicated.

Section 4 extends the method into the clausal postmodification of noun heads by contrasting the addition of the same structure in two different ways: serial postmodification of the same noun phrase head by clauses and embedded postmodification of embedded heads by clauses. Each obtains distinct distribution patterns, and these distributions differ between speech and writing. Section 5 concludes by locating the methodology within corpus linguistics and reviewing its potential for evaluating grammatical frameworks.

1.1 The divergence of frameworks

A number of distinct parsing schemes have been exhaustively applied to corpora. Penn Treebank notation (Marcus *et al*., 1993) is a skeleton phrase structure grammar applied to a number of 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, see Section 2) 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).

A standard criticism of the treebank linguistics community is that since theorists' knowledge of grammar is contested, any selected framework is likely to be 'wrong'. Sinclair (1987) influentially argued that corpus linguistic insight should be driven by word patterns rather than subsumed under a given grammar. Many corpus linguists do not use parsed corpora. A key reason is the perception that the linguist is inevitably trapped in the framework decided by annotators. Addressing this concern is central to the design of the ICECUP parsed corpus exploration software (Nelson *et al*., 2002: 86; Wallis, forthcoming). This relies on a theoretical decomposition principle: that a particular aspect of a framework – the analysis of co-ordination, say – may be dealt with independently from other aspects.

Finally, whereas the gulf between theoretical linguists such as Chomsky and lexical corpus linguists like Sinclair is wide, the bridge will most likely be via engagement with parsed corpora. We take the view that the bridge between rationalism and empiricism, as in any scientific paradigm, is in the application of theory to data: in other words, through the parsing of corpora and the evaluation of theoretical claims using such corpora.

1.2 Criteria for selecting and evaluating frameworks

What are the benefits of parsing a corpus with a particular framework, and can they be empirically demonstrated? Ideally, we would wish to compare different parsing schemes applied to the same corpus data. However, a linguist-validated parallel, multi-parsed corpus of sufficient size does not exist (van Zaanen, *et al.*, 2004), so we could only compare frameworks applied to different corpora. This means we cannot rule out an alternate hypothesis that an observed difference is due to the data rather than the framework. So, over the course of this paper we consider two more modest propositions: empirical benefits of parsing compared to a part of speech tagged analysis, and step-wise refinement of a single aspect of a parsing scheme. The latter can be achieved by changing the definition of queries (the 'abstraction layer') rather than the annotation itself.

It is worth noting that all frameworks are evaluated empirically by corpus parsing for decidability and coverage (Wallis and Nelson, 1997; Wallis, 2003). Annotators classify existing terms in the framework and decide how to incorporate unanticipated phenomena. Of necessity, this process typically leads to minor modifications of the grammar rather than a critical review of an entire aspect or even the whole framework. The input text, particularly in the spoken domain, may be 'noisy'. A key aspect of the parsing task is deciding how self-correction, incompleteness and other grammatical infelicities are dealt

with. Should they be annotated as performance errors and the 'correct' sentence parsed? Where do we draw the line?[1]

This paper concerns a second evaluation process: the review of completed parsed corpora and the syntactic frameworks they incorporate. First we must agree evaluative criteria.

1.3 Retrievability of linguistic events

The most common criterion used for differentiating grammatical frameworks is often referred to as 'distinguishability' or 'decidability', i.e. that one linguistic concept can be distinguished from other similar concepts. This is intuitive and has deductive appeal. Once categories are applied to the corpus, conceptual distinguishability becomes empirical 'retrievability', i.e. 'the reliable retrieval of linguistic events' (Wallis, 2008). This implies that if a concept – the subject of a clause, a particular type of direct object, etc. – can be reliably retrieved from corpus A but not from B, then the representation used in A can be said to be 'better' than the one annotating B.

For example, the scope of attributive adjectives over co-ordinated noun heads can vary. The following are not grammatically distinguished in the ICE-GB corpus (Section 2). Scope is not encoded. We cannot retrieve Example (1) without obtaining Examples (2) and (3).

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

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

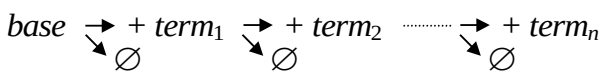
(3) *late teens and twenties* [S1A-013 #107] (ambiguous)

Reliable retrieval of instances of linguistic concepts ('linguistic events' for shorthand) is a necessary criterion, but it is insufficient. It has three fundamental disadvantages. First, it is *circular*. The value of the concept in question, such as attribute scope, is assumed in advance. Another linguist may simply consider it to be theoretically irrelevant. Second, it admits *redundancy*: a multi-parsed scheme C containing all the terms and relations of parsing schemes A and B will always be 'better' than both. However, theory complexity should be considered against Occam's razor (the principle that theories should be as simple as possible, but no simpler). Third, it is fundamentally an *atomic* evaluation concerned with evaluating single events within a grammatical structure, rather than the structure itself.

1.4 Retrievability of patterns of interaction

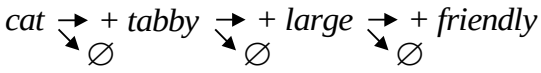
In this paper we tentatively propose a new criterion based on the evidence of interaction between instances of linguistic concepts. This is an issue of explanatory and predictive power: a 'better' framework allows us to make novel theoretical predictions or identify novel phenomena that others cannot. Our proposal is to examine evidence of patterns of interaction between construction decisions expressed along grammatical axes. Our paradigm builds on the 'linguistic event retrieval' principle above by exploring the impact of one linguistic event on another.

We study patterns of repeatable construction decisions of the following form:



Starting with a base construction, arrows indicate distinct autonomous *decisions* to add a further term (or not, hence '∅'), and plus signs indicate the application of an *operator* that adds terms in a specific way (i.e., governed by a particular structural relationship). The 'axis' is defined by this addition operator, and each addition is constrained by grammar rules. The method simply requires that an operation is defined by a grammar rule and is repeatable.

For example, we might infer that a speaker communicating a description of the family pet to a would-be house-sitter might make mental decisions to express adjectives in the following order.

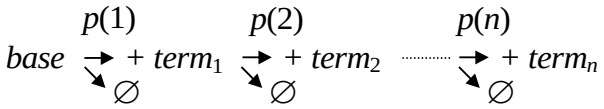


Such construction decisions are not necessarily reflected in the word-order articulated, indeed we expect the head *cat* to be decided upon before adjectives are considered. The speaker might then say *my cat, she's a tabby, large and friendly*, or *my large friendly tabby cat*. The former has no premodifying adjectives, the latter, three.

In this experimental paradigm we study the relative probability $p(x)$ of adding each successive term x in a given position, up to an observed maximum n, obtained from corpus data. We will refer to $p(x)$ as 'the additive probability' for term x, i.e. the chance of adding a term to a construction already containing x-1 terms. The grammatical framework is

necessary to extract data along the axis under consideration; the corpus evidence gives us the chance of each term being added to the construction.

The scheme therefore looks something like this:



The probability of obtaining a sequence of at least x terms is then the product $p(1) \times p(2) \times \ldots p(x)$. By studying variation in $p(x)$ over successive addition operators, we can explore the impact of one language production decision on the next. Since each addition makes the construction syntactically more complex, our method may also be considered as selectively visualising the rate of generation of a single dimension of syntactic complexity (cf. Beaman 1984: 45).

Caution is required in interpreting results. An observed pattern could arise for multiple reasons. Whereas it may be tempting, for instance, to see distributions as evidence of underlying psycholinguistic processes, other explanations may be valid.

Anderson (1983) makes this argument in a different way. He distinguishes between evidence of performance and the underlying cognitive processes that bring this evidence about. Empirical evidence of a psychological process are the shadow or 'signature' of the phenomenon. Since computer simulations might replicate that signature by different methods, we should not claim (as do some 'strong AI' proponents) that a closely-matching simulation is an accurate model of human cognition. Similarly, a computer system for generating 'natural language' does not provide understanding of how humans produce language; nor parsers, how humans interpret sentences. Rather, simulations are useful because they help identify *parameters* of the human cognitive process. Our proposition is that this type of natural experiment[2] on parsed corpora may help identify some parameters of corpus contributors' processes of language production.

In sum, our proposal is to employ a criterion for grammatical evaluation based on the reliable retrieval of patterns of interaction between language construction decisions. This proposal is consistent with a perspective of scientific theories (grammars) being stronger where they have greater explanatory power.

In this paper we examine patterns obtained with a particular grammar in a parsed corpus. We compare results with those obtained from part of speech annotation, and we consider a permutation of a single aspect of the grammatical scheme.

2. Three experiments with attributive adjectives in noun phrases

Let us consider a simple example to illustrate our method. 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).[3] We will investigate the proposition that the introduction of each adjective constrains the addition of the next.

The null hypothesis is that each decision to add an attributive adjective is independent from the next, i.e., that the probability of adding a second adjective is the same as the probability of adding the first, and so on. In short, the null hypothesis is that $p(x)$ is constant.

As we saw in the *tabby cat* example, identifying that an interaction is taking place between decisions to add two adjectives *A* and *B* does not establish that speakers made decisions in this order. The decision to insert *A* could be made prior to the decision to insert *B*, or vice versa; made in parallel; or (in writing), be subsequently revised. Our method does not rely on the order of decisions being known.

2.1 Experiment 1: Attributive adjective phrases

Our first task is to collect data. In a part-of-speech tagged corpus we can obtain frequencies of cases of single, double, etc., adjectives followed by a noun (see Section 2.3 below). In a parsed corpus we can be more precise, limiting our query by the noun phrase (NP) and counting attributive adjective phrases (AJPs) rather than adjectives. This permits us to count cases such as *the old [pale blue] ship* correctly (cf. 'retrievability', Section 1.3).

We use a parsed corpus as our source. 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. ICE-GB is supplied with an exploration tool, ICECUP, which has a grammatical query system using idealised grammatical patterns termed *Fuzzy Tree Fragments* (FTFs, Wallis and Nelson, 2000) to search corpus trees (for an example tree, see Figure 5 below).

We perform a series of queries to obtain the raw frequency, $F(x)$, of each set of constructions consisting of at least x terms, i.e. all constructions where the decision to add x terms to the base (in this case, the noun) were made. We construct a series of FTFs of the form in Figure 1, i.e., an NP containing a noun head and x adjective phrases before the head. These FTFs find cases where at least x AJPs precede the noun.

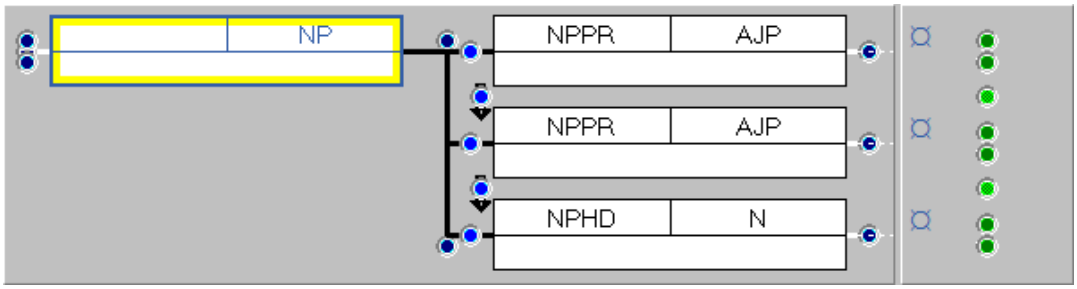


Figure 1. Fuzzy Tree Fragment for x=2, retrieving NPs with *at least* two premodifying adjective phrases (NPPR, AJP) preceding the noun head (NPHD, N). For reasons of space, the tree is drawn from left to right, with the sentence on the right hand side.

Table 1. Frequency and relative probability of NPs with x attributive adjective phrases before a noun head, in ICE-GB. A significant decline is found where $w^+(x) > p(x\text{-}1)$ (highlighted for x=2).

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*(*x*)	193,135	37,305	2,944	155	7
probability *p*(*x*)		**0.1932**	0.0789	0.0526	0.0452
Wilson upper bound $w^+(x)$		0.1949	0.0817	0.0613	0.0903
significance ($p > w^+$)			s-	s-	ns

By applying a series of FTFs we obtain the raw frequency F in Table 1. The additive probability for the x-th addition operation, $p(x)$, is obtained simply from

$$p(x) \equiv F(x) / F(x-1).$$

In this experiment, $p(1)$ represents the chance that a noun ('<N>') is preceded by an adjective phrase; $p(2)$ the chance that the sequence '<AJP> <N>' is preceded by an adjective phrase; and so on.

As we noted, the null hypothesis is that $p(x)$ is constant for all x, like the probability of a coin toss being a 'head' or of throwing a six with a die. However many times we toss a coin, the probability of a 'head' is always the same.

The raw frequency distribution $F(x)$ in Table 1 is plotted in Figure 2a. The data appears to decay exponentially – like a coin toss series distribution. However, when we examine the probability of adding each AJP, $p(x)$ (Figure 2b), we find that $p(x)$ falls as successive AJPs are added.

This observation is statistically robust. The graph includes 95% Wilson score confidence intervals (Newcombe, 1998; Wallis, 2013) visualised as 'I'-shaped error bars.[4] Since the data supporting an observation is a subset of the previous set, we employ a 'goodness of fit' test to identify significant difference. This test can be performed by visual

inspection. Where an earlier point, $p(x\text{-}1)$, lies outside the interval for a point, $p(x)$, the difference between points is statistically significant. See Figure 2b.

We conclude that $p(3) < p(2) < p(1)$, i.e. we see a significant decline in the probability of adding each successive adjective phrase. See also Table 1. The results also reveal that the fall in probability over multiple steps is significant, so we could also claim that $p(4) < p(1)$. However, in this paper we will mainly restrict ourselves to conclusions concerning successive trends.

We conclude that, at least for $x<4$, decisions to add successive attributive adjective phrases in noun phrases in ICE-GB are not independent from previous decisions to add AJPs. On the contrary, our results reveal a negative feedback loop, such that the presence of each AJP reduces the chance the speaker will add another.

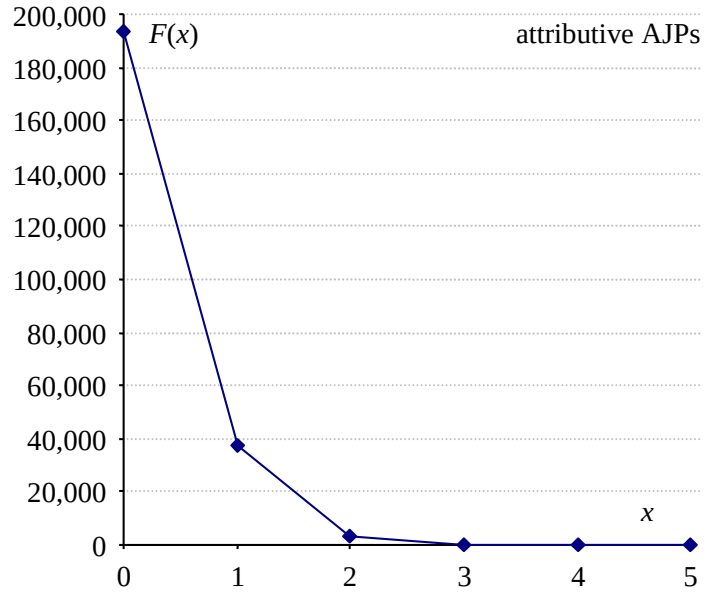


Figure 2a. Frequency $F(x)$.

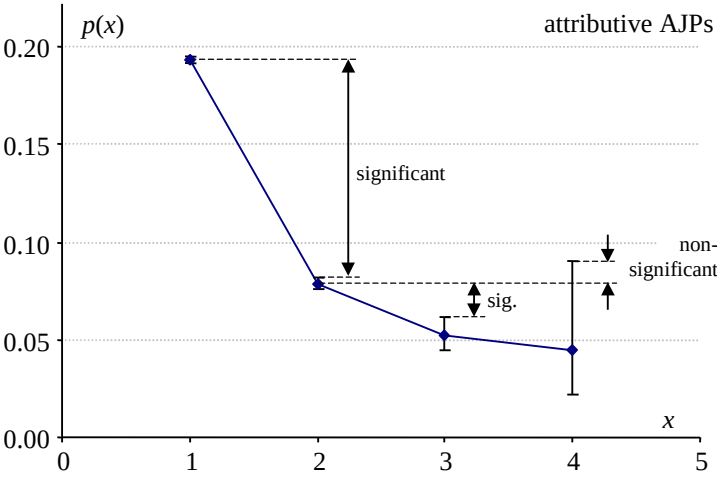


Figure 2b. Additive probability $p(x)$, with 95% Wilson intervals indicating statistically significant falls in $p(x)$.

Figure 2. Plotting frequency and probability of a 'run' of x attributive AJPs in a noun phrase in ICE-GB.

This result appears to confirm evidence that the use of successive multiple adjectives is avoided due to linguistic or contextual constraints (Feist, 2011). It does not tell us what these constraints might be.

Possible hypotheses for explaining these results include the following.

Table 2. The additive probability of attributive adjectives in NPs with common noun heads – we see a significant serial decline.

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

Table 3. The same probability for NPs with proper noun heads – we find no significant change over each addition step.

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

i. Logical-semantic constraints. Speakers tend to say *the tall blue ship* rather than *the blue tall ship* or *the tall short ship*. This is the most likely explanation: each adjective reduces the set of semantically coherent adjectives that may be added at the next stage. It is consistent with Feist's concept of 'zones' (Feist, 2011: 8).

ii. Communicative economy. Corpus examples include multiple references to the same entity in a conversation. It seems probable that on subsequent references, speakers tend to shorten to *the ship* (or use pronouns) for reasons of communicative efficiency. However this hypothesis, sometimes called the 'principle of linguistic economy' (Zipf, 1949), predicts the almost complete avoidance of adjectives on subsequent reference. It does not seem to explain successive reductions in additive probability.[5]

iii. Cognitive memory/processing constraints. Limits on short-term memory processing have been discussed since Miller (1956), who observed that many processing tasks seem to falter once more than five to nine items were held in short-term memory. A similar effect may apply in this case, although we are seeing a gradual reduction, not a steady state and then a drop-off at $x>5$.

The observed curve may be due to more than one of these potential causes.[6] However, the curve's shape (cf. Anderson's 'signature') could provide clues to the most likely source of the interaction.

We have demonstrated evidence of a general trend along the axis of the grammatical analysis of NP constituents. In the remainder of this section we will consider permutations

of this experiment. In sections 3 and 4 we apply the method to other repeating linguistic phenomena.

2.2 Experiment 2: Attributive adjective phrases with proper and common noun heads

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

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 we do in Table 2), the proportion of NPs with one AJP or more would increase and the relative decline from this point would become greater.

Tables 2 and 3 bear this out. NPs with common and proper noun heads behave differently. 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 is analysed as containing one or more AJPs, but we cannot identify statistically significant variation over addition steps.[7]

A caveat applies to Table 3. ICE annotators used a compound analysis for 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, as well as those analysed adjectivally, such as *the lower Loire* and *a British Bishop* (#83).

This returns us to the point we made in the introduction. 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 before further research is undertaken. However, this observation should not detract from our observed systematic decline, reinforced by the distribution for NPs with common noun heads in Table 2.

2.3 Experiment 3: Attributive adjectives, without parsing

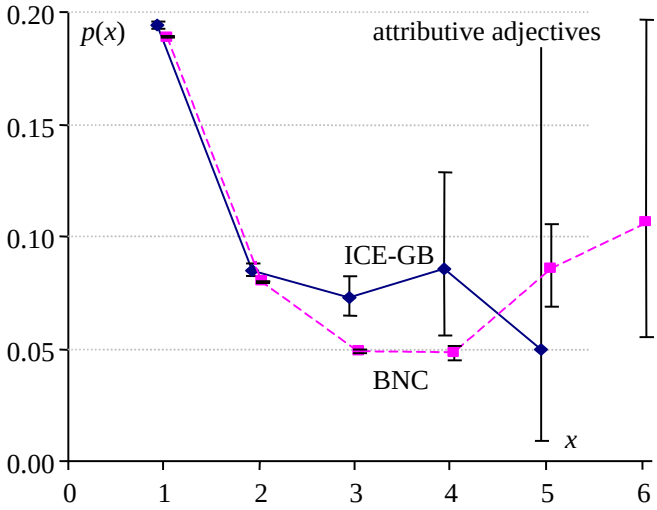


Figure 3. The additive probability for *x* adjectives before a noun in ICE-GB and BNC.

Let us put the parse analysis aside and replace FTF queries with a series of simple sequential word class queries: '<N>' (single noun), '<ADJ> <N>' (adjective preceding a noun), etc. This is, of course, precisely the type of search possible with a tagged corpus. The result (Figure 3, solid line) shows similar evidence of an initial decline, however the probability then appears to fluctuate (the apparent rise is not statistically significant). How do these results compare with those of Experiment 1?

Inspecting the corpus reveals a large quantity of noise with the longer strings. We find 19 cases of a 4-adjective string but only 7 with four attributive AJPs. There are no cases with 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:

(4) *specious ex post facto justification* [S1B-060 #8] (2 AJPs)

(5) *mauvey blue beautiful little thing* [S1B-025 #28] (3 AJPs)

(6) *long long long straight roads* [S2A-016 #29] (4 AJPs)

(7) *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 are 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 fewer than four AJPs! The over-counting 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. Experiment 3 is significant for x=3 only at the 0.05 error level, whereas both observed differences are significant at the 0.01 error level in Experiment 1.

The ability to exploit the parse analysis with adjective phrases improves the reliability of the result, but even without the parse analysis it was still possible to find evidence of the same pattern of declining probability of attributive adjectives. This curve clearly reflects a strong systematic effect involving simple units (adjective phrases), and theoretically less precise methods may still be effective.

As we are simply relying on word class tags, Experiment 3 is reproducible with larger corpora. The tagged *British National Corpus* (BNC, Leech, 1992) contains 100 million words, 10% of which is from spoken sources. Figure **3** shows the results obtained by a simple word class tag sequence search, across all data in the BNC, as a dashed line.[8] Initially the trend seems very similar to that obtained for ICE-GB, although it falls further for x=3. The larger volume of data allows us to penetrate deeper, and reveals a statistically significant *increase* in probability for x=5. However, if we review these cases we find fewer than 60% were correctly classified, and cannot conclude that this increase is genuine.[9]

The overall trend exposed by Experiment 1 is confirmed, but results for longer strings are less reliable due to two factors: an increased risk of classification noise, especially in the case of the unchecked BNC, and the introduction of measurement error by counting adjectives rather than adjective phrases.

3. Experiment 4: Grammatical interaction between preverbal adverb phrases

So far we have identified a general feedback process that appears to act on the addition of attributive adjective phrases prior to a noun head. We also speculated on potential causes. Let us briefly investigate whether the same type of effect can be found in adverb phrases (AVPs) prior to a verb, such as

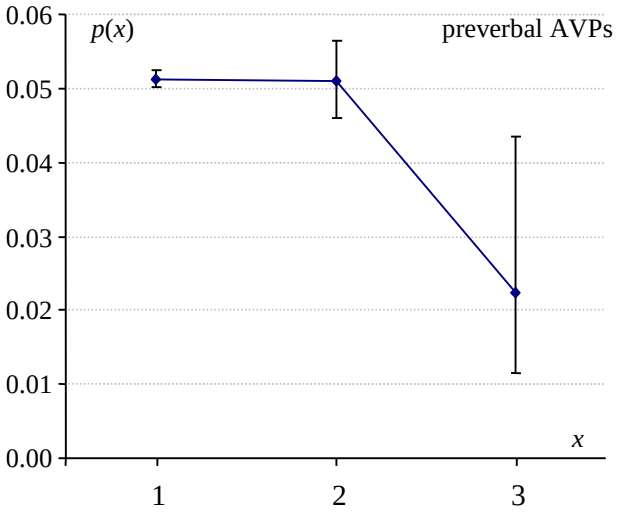


Figure 4. The additive probability of adding a preverbal adverb phrase in ICE-GB.

Examples (8) and (9), where the adverb phrases are underlined.

(8) I think she'd *rather just sing* [S1A-083 # 105]

(9) we can *only sort of work* and see what happens [S1A-002 #109]

We obtain data from ICE-GB using FTFs that capture a single VP; an adverb phrase preceding a VP; two adverb phrases preceding a VP, etc. The probability of adding a second AVP to the first is almost identical (slightly more than 1 in 20) to the probability of adding the first AVP. The probability of adding a third AVP falls significantly at a 95% confidence level. The result is shown in Figure 4.

This result demonstrates an important point. Preverbal adverb phrases do not interact in the same manner as attributive adjective phrases. It follows that a negative feedback loop on repeated addition is not merely some kind of universal law applying to all constructions.

Whereas adjective ordering in English has been studied exhaustively by corpus linguists, adverb ordering (cf. *just rather* vs. *rather just*) has not received the same scrutiny. Our evidence suggests that adverb phrases do not semantically interact to the same degree as adjective phrases, if indeed they interact at all. The same pattern is found in speech and writing, suggesting that the cause is not cognitive. However more research, and more data, is clearly needed.

4. Experiment 5: Grammatical interaction in postmodifying clauses

Noun phrase postmodifying clauses (denoted as 'NPPO, CL' in the TOSCA/ICE grammar) are similar to adjectives in that they add meaning and specificity to the head noun, but constitute entire clauses, such as Example (10).

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

In Example (10), *whose face* is also a noun phrase: it is the subject of the postmodifying clause *whose face has been somewhat rubbed*. In this experiment we investigate the impact of multiple additions of postmodifying clauses containing NPs,[10] in two cases:

i. *sequential* postmodification, where clauses modify one NP head, and
ii. *embedded* postmodification, where clauses modify the immediate prior NP head.

These two types of multiple postmodification are summarised by the pair of two-level FTFs in Figure 6. Since they operate on the same head, we might expect that sequential postmodifying clauses behave similarly to sequential adjective phrases in an attributive position.

In ICE-GB we find cases of postmodifying clauses containing NPs with up to three levels of recursive embedding. An example of two-level embedding is Example (11) below (see Figure 5):

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

There are three methodological issues in extracting data for these types of structure.

i. Matching the same instance multiple times. FTFs count every matching permutation of a query (Nelson *et al.*, 2002: 272), but we need to count unique instances. The lower NP nodes in the FTFs in Figure 6 have no specified function, and could match a subject and object of the same clause, causing the same clause to be counted twice. Consider Example (12), which should count as a single instance.

(12) *the things* [*that the students members of the group say*] [S1A-001 #57]

There is only one NP, *the things*, being modified. But a one-level FTF will generate

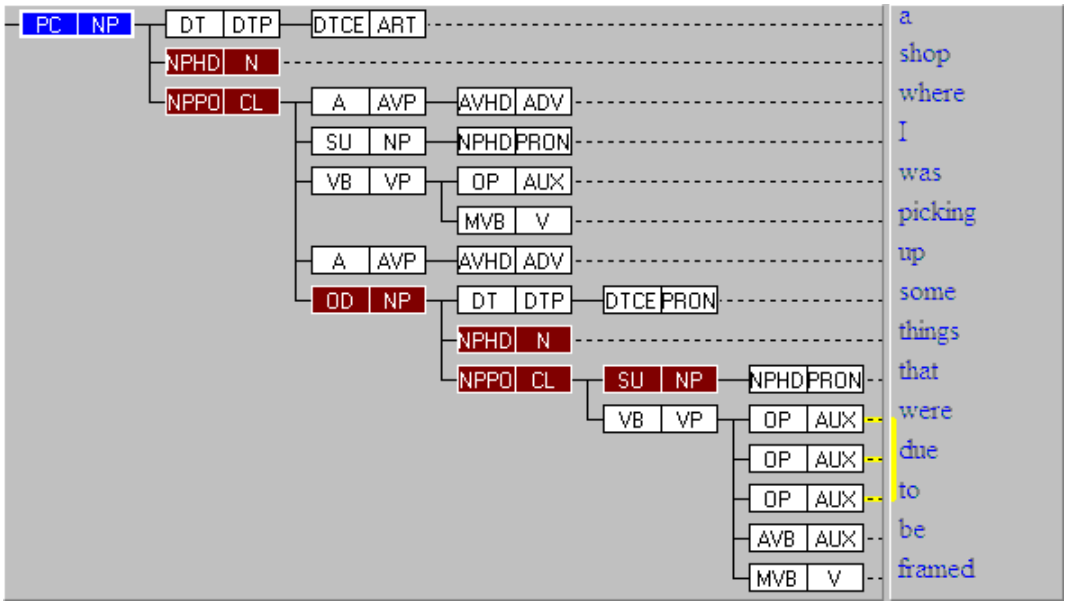


Figure 5. Two levels of embedding in ICE-GB (text unit S1A-064 #132). The shaded nodes are matched by the two-level FTF in Figure 6b.[11]

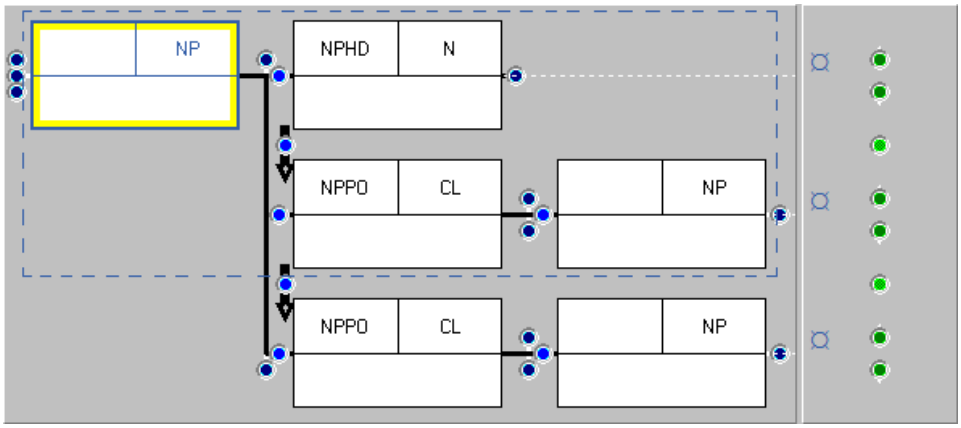


Figure 6a. Sequential FTF (level 2).

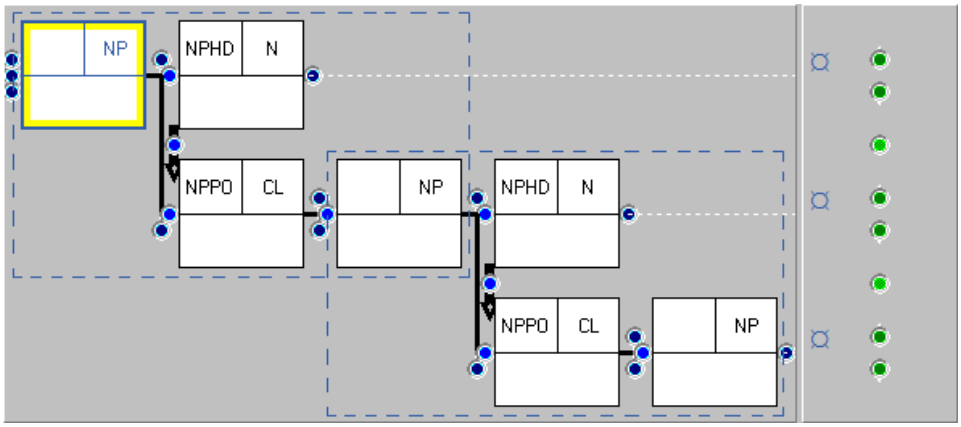


Figure 6b. Embedded FTF (level 2).

Figure 6. Basic Fuzzy Tree Fragments for finding NPs with two postmodifying clauses containing NPs. The dashed lines indicate one-level FTFs.

three matching permutations, one for each NP: direct object *that*, subject *the students*, and parenthetical *members of the group*. We either discount duplicate cases (with the 'next generation' ICECUP IV software this may be automated[12]), or alter the FTF – either to remove the subordinate NP or specify its function (say we are interested only in direct objects, for example). We used ICECUP IV in this study.

ii. Double-counting subordinate FTFs. In Experiments 1-4, there is only one possible head (the 'base' concept), which anchors the FTF. We categorise multiple modifications of that head by the number of modifiers it possesses. A level two FTF (x=2) contains a level one FTF plus an additional step (Figure 6a). However, embedding is subtly different. The two-level FTF (Figure 6b) can be considered as containing *two* one-level FTFs, indicated by the dashed lines in the figures. In two-level examples like Example (11), the case would match a one-level FTF twice:

(11a) *a shop* [*where I was picking up some things*]

(11b) *some things* [*that were due to be framed*]

We need to exclude the second instance (Example (11b)) from our count. The solution turns out to be remarkably simple. We subtract the number of subordinate matching instances, $F(x+1)$, from the total, $F(x)$.[13] The same reasoning applies to two-level matches within three-level cases, etc. This issue is not peculiar to embedding, but may occur in any sequence of additive steps where a one-step sequence is found twice in a two-step sequence.

iii. Matching cases containing co-ordination**.** In the ICE-GB scheme, co-ordination is represented by the introduction of a 'co-ordination node' that sits above, and brackets, conjoin nodes. Either clauses (Example (13)) or noun phrases (Example (14)) may be co-ordinated.

(13) *endorphins* [*which are morphine-like substances* [*which we make ourselves in times of stress or exertion or injury*] ***and*** [*which are our natural home-made opiates*]] [S2A-027 #44]

(14) *an asymmetric building* [*showing frames* ***and*** *walls* [*that are distributed non-uniformly*]] [S2A-025 #18]

Although both (13) and (14) are two-level examples, neither would be found by the FTF in Figure 6b. The co-ordination node in the tree 'blocks' the matching algorithm. To find these cases we create additional FTFs to match these patterns, and categorise cases using ICECUP IV.[14]

Tables 4 and 5 summarise the results, which are plotted in Figure 7a. We find a significant fall in probability in both sequential and embedding cases. The probability of embedding a second clause is significantly higher than that of applying a second (sequential) postmodifying clause to the same head, i.e. the decline is less steep.

However, in the sequential case we discover a subsequent *increase* in probability. This observation requires more discussion.

Table 4. The additive probability for sequential postmodifying clauses within noun phrases reveals a decline and rise.

***x* NPPO sequential clauses**	**0**	**1**	**2**	**3**
'at least *x*' *F*(*x*)	193,135	10,093	166	9
probability *p*(*x*)		0.0523	0.0164	0.0542
significance			s-	s+

Table 5. The same probability applied to embedded postmodifying clauses exposes only a decline.

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

4.1 Sequential postmodfication

We hypothesised that sequential postmodifying clauses operating on the same head would semantically constrain each other in a manner comparable to attributive adjectives. This is not what we found. Logical and semantic exclusion of the kind found between adjectives seems to have limited effect, probably because clauses are an open set. Whereas one adjective, e.g. *blue*, might exclude another, *pink*, or make it very unlikely, adding a clause does not reduce the number of permissible clauses at the next stage. On the other hand, clauses are large units, so we are more likely to see cognitive processing costs reflected in additive decisions. Alternatively, it may be a conscious avoidance of tangents by adapting to the needs of an audience – the result of communicative economy.

What might lie behind the subsequent rise? The increase from *x*=2 to *x*=3 appears anomalous until we consider the relationship between serial postmodification and co-ordination and examine matching clauses.

The grammatical scheme contains a potential ambiguity in the analysis of double postmodification: as two independent postmodifying clauses or as a single co-ordinated pair of postmodifying clauses, as in Example (15).

(15) *…the process* [[*how you turn off*]$_{CJ,CL}$ [*how you turn back on*]$_{CJ,CL}$]$_{NPPO,CL}$ [S1A-050 #109]

We therefore considered the consequences of relaxing the distinction between sequential and co-ordinated postmodifiers. We treat each co-ordinated case as an independent postmodifier of the same head, obtaining the upper line in Figure 7a, marked ‘sequential or conjoined’.

Examining co-ordinated cases indicates that Example (15) is not unusual. The data suggests that, rather than one postmodifying clause limiting the next, the first clause may actually be a template or cue for the construction of the next. In Example (15), *how you turn off* provides an easily-modified template for *how you turn back on*.

Evidence of templating is also found in sequential cases. Compare Example (16), analysed as co-ordinated, and Example (17), which – in the absence of the co-ordinator *and* – is analysed as sequential.

(16) …*his consistent bids* [[*to puncture pomposity*] [*to deflate self-importance*] [*to undermine tyranny*] and [*to expose artifice*]] [S2B-026 #81]

(17) …*one path* [*which was marked… on a …map*] [*which is no longer marked*] [S1B-037 #85]

Perhaps the distinction between co-ordinated postmodification and serial postmodification is not as important as we might otherwise believe. Arguably, the concept of ‘co-ordination’ should be extended to include cases of sequential repetition. In this case, one might reasonably critique the grammatical framework for encoding an unnecessary distinction, or failing to extend the concept of co-ordination to such sequential ‘asyndetic’ cases.

Subdividing the sequential data by speech and writing reveals that this subsequent rise is concentrated in spoken data, whereas the overall rate of successive postmodification in writing falls faster, declining to almost zero (Figure 7b).

This seems to support a communicative thesis. Repetition and emphasis where an audience is present has a communicative function. Moreover, where a process is easier in speech than in writing, cognitive constraints are unlikely to be a credible explanation.

It seems that in spontaneous speech (often, although not exclusively, in the context of an audience), the noun head is serially postmodified to a greater extent than in (mostly non-spontaneous) writing, where an audience is absent. The wide confidence intervals at x=3 and above in Figure 7b prevent us from making a claim that the rate of sequential addition for spoken data exceeds that for written data. At x=2, however, the difference is clear.

4.2 Embedded postmodification

We turn next to embedding. The difference between embedded and sequential patterns is dramatic (Figure 7a). Embedding shows no evidence of a later increase in probability. Observed probabilities are significantly different, as are falls in probability in each case.[15] Sequential results obtain a greater decline at this point (x=2) than embedding. Note that in cases of embedding, we can reasonably assume that decisions are made in word order.

In embedding, each subsequent clause applies to a new head, so neither semantic exclusion or repetition / templating are likely to apply. However, embedding leads to 'garden path' behaviour and increased semantic complexity due to the introduction of new heads, each of which refers to distinct entities. The underlying source of this decline may be cognitive memory / processing limitations, communicative economy, or both.

Whatever the underlying cause, embedding also demonstrates a significant fall in probability, and hence, an interaction between decisions. Adding a postmodifying clause containing a further NP comes with a linguistic cost.

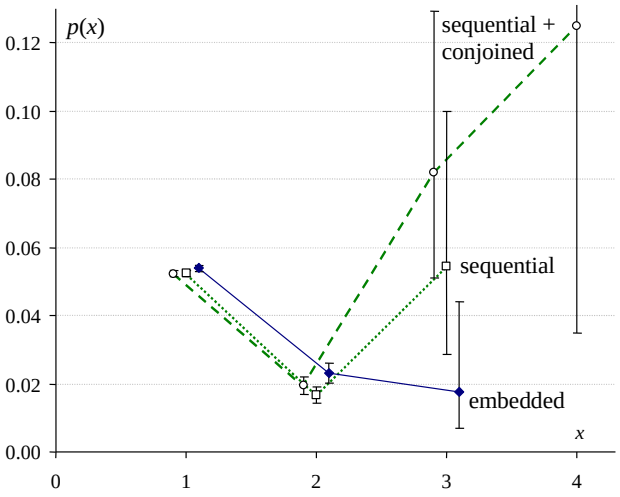


Figure 7a. Embedded, sequential, and sequential or conjoined postmodifying clauses: all ICE-GB data.

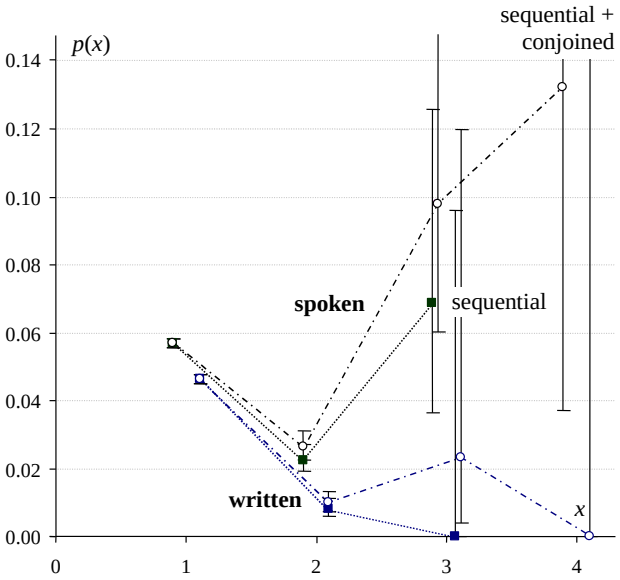


Figure 7b. Sequential, and sequential + conjoined patterns: speech vs. writing.

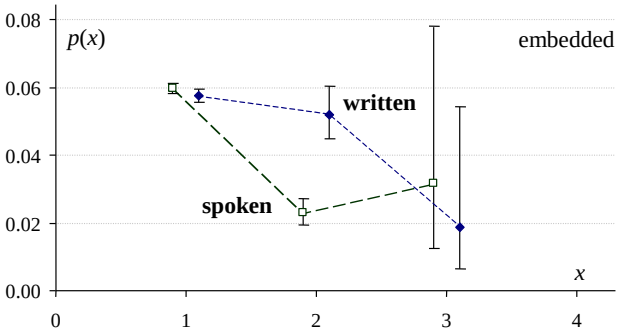


Figure 7c. Embedded patterns: speech vs. writing.

Figure 7: Contrastive additive probability plots for sequential and embedded clauses postmodifying a noun in ICE-GB, with 95% Wilson intervals.

If we subdivide the embedded data into subcorpora of speech and writing (Figure 7c) we find that in writing, the decline in probability due to embedding is delayed compared to speech. This seems to be consistent with the spontaneous nature of much of the spoken data in the corpus compared to the far less spontaneous, and often post-edited, written mode.

4.3 Alternative explanations

In reviewing results such as these we must be careful not to presume our conclusions. Let us briefly consider three possible alternative hypotheses.

AH1. Interaction is a lexical adjacency phenomenon that requires no syntactic explanation. In other words, if grammar did not exist, could we explain our results? Figure 5 illustrates that embedded terms need not be adjacent. Moreover, examining corpus examples finds over 85% of two-level cases contain at least an intermediate VP, NP or clause between the upper and lower NP heads. These are not lexically adjacent. We may reject this hypothesis.

AH2. Observations are artefacts of misclassification. Could embedded clauses be incorrectly attached to trees? Perhaps some embedded cases were misclassified as sequential, explaining the decline? Reviewing the set of two-level sequential cases finds that parsing error cannot explain the results. Of 166 cases of two-level sequential cases we identify a maximum of 9 potentially-embedded ambiguous cases and one incorrectly-attached example. 95% are identifiably correct, or simply cannot be embedded.[16]

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 Example (11) 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, it would not refute the claim that the observed interaction on the embedded axis was grammatical in nature. Conclusion: the fall is genuine, and distinct from serial postmodification.

AH3. Segmentation. During the parsing of the corpus, annotators sometimes broke long spoken ‘sentences’ with segmentation breaks. These were introduced at ‘run-on’ points in long sentences (typically, co-ordinated main clauses). Embedded clauses were not broken up. However, again, our sequential postmodification distribution reveals the opposite trend. If excessive segmentation had an impact on the results, it would be to make lengthy serial postmodification less likely, especially in spoken data. We observe the opposite trend.

4.4 Discussion

Once we have rejected the null hypothesis that each decision to postmodify a noun phrase is independent from previous decisions, we have evidence for the claim that grammatical annotation in the corpus is capturing ‘signature’ artefacts of the language production process alongside communicative constraints.

Since the probability of a speaker choosing to embed a postmodifying clause falls with every subsequent decision, we have statistical evidence of an interaction along this embedding axis.

Similarly, for sequential postmodification we find an initial fall, followed by a rise, but this rise is limited to speech data. Differences between speech and writing could be attributed to the spontaneity of speech or the presence of an interacting audience. A further subdivision into monologue and dialogue may sift this out.

In sequential postmodification we found evidence of ‘templating’, which Tannen (1987) refers to as conscious ‘poetics’ and communicative cohesion principles of lexical repetition. Tannen argues that, far from being redundant, ‘the relative automaticity of repetition facilitates language production in conversation.’ This may explain why we see a subsequent rise in post-modification in speech not seen in writing (see Figure 7b).

Even where lexical items are not repeated, structural self-priming (Pickering and Ferreira, 2008) is evident, as in Example (16). Just like co-ordination, sequential postmodification shows a strong tendency to repeat grammatical structures.

Structural priming is usually discussed in terms of one speaker influencing another, or as an effect spanning several minutes. We find examples of what we might call ‘micro-priming’ within the same utterance, indeed, sometimes within the same noun phrase. This phenomenon might also be seen as an instance of spreading activation, including other kinds of semantic association.

We hypothesised that certain aspects of grammatical analysis could be shown to provide explanatory power in empirical research, and that such a perspective had the potential to be an objective benchmark for comparing frameworks.

In this case, we identified that grammatical analysis sufficient to identify these structures is a requirement to perform this type of experiment. This statement does not mean that every detail of this *particular* grammatical framework is 'correct' – it means that a grammar that represents such structures is required to account for these phenomena. Moreover, we showed that a distinction in the analysis between sequential and co-ordinated postmodification is probably unnecessary. Removing it increases our explanatory power by confirming this templating trend.

5. Conclusions

At first glance, our initial experiments simply provide empirical support for some general linguistic observations, namely that constraints apply between adjectives, phrases and clauses that do not apply between preverbal adverbs. These constraints might 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 compound proper nouns mentioned in Section 2.2.

Our primary aim in these early studies was to exemplify our methodology and (in the case of adverbs) to show that it distinguishes between constructions and, presumably, underlying processes. We were then able to employ the same approach in the evaluation of sequential decisions for embedded and sequential postmodification of noun heads.

In the case of complex constructions such as embedded postnominal clauses, each NP has a different head. Cumulative semantic constraints cannot explain this interaction. We surmise that we have found evidence of communicative and/or psycholinguistic constraints. If we are correct, future research may establish that this method is capable of distinguishing language deficits and fluency. To scale to clinical applications the method would need to be married to a reliable automatic parser for these elements: nonetheless we believe this insight is worthy of further research.

5.1 Implications for corpus linguistics

From the perspective of corpus linguistics, the method proposed is relatively novel. These experiments concern general 'structural' trends between grammatical elements. We have not attempted investigations of why speakers might make particular individual choices.

Examples of this type of research can be found in Nelson *et al.* (2002) and Wallis (forthcoming). Individual choice research often requires a narrow focus to identify a genuine choice (true alternates), and may 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 semantic classes.

Using different means, collocation, colligation, *n*-grams and similar algorithms obtain interaction evidence in a data-driven manner. Identifying lexical patterns does not necessarily demonstrate the existence of semantic constraints, but might allow constraints, *pace* Sinclair (1987), to 'emerge'. However, as we found, a collocation pattern may be due to more than one potential cause. As revealing as they may be, these methods are essentially exploratory in nature, requiring further experimental investigation.

Although both unsupervised collocation algorithms and supervised interaction experiments are very different, each address specific questions concerning particular variables or particular lexical items. For this reason we might term these 'Level I' research.

By contrast, the experiments in this paper are more general in nature, summarising the total pattern of interaction at the level below, rather like Zipf's (1949) famous law of exponential distribution. Zipf does not tell us which items are found where in a sequence, rather that items can be expected to be distributed like this. Likewise, our observed additive probability distributions 'frame' or parameterise Level I research, just as they triggered our review of sentences. Discovering that terms *A* and *B* interact does not tell us why they do so – this is a Level I research question.

'Level II' observations of the type described in this paper collect patterns of interaction expressed over a series of potential linguistic choices. Experiments 1 to 3 with adjective (phrases) provide evidence of a general phenomenon of feedback in language production. Experiment 2 demonstrates that this differs for NPs with common and proper noun heads.

The BNC results found within Experiment 3 appear at first glance to obtain slightly different results, but close analysis causes us to reject this hypothesis. Large tagged corpora, unchecked by human linguists, may contain artefacts of the tagging algorithm that skew the results for longer sequences.

Experiment 4 demonstrates that the observed declining trend of attributive adjective addition is not simply a consequence of the contingent nature of language, but one that arises from *particular* processes. Some speaker choices are less constrained than others.

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

What we might ambitiously call 'Level III' research would concern the validation of grammatical frameworks. Experiments 1 and 3 indicate that the chosen framework – encapsulated as a parse analysis – has a research benefit in more reliably capturing a general trend partially obscured in the lexical sequence. On the other hand, Experiment 5 identifies a result over lexically separate units that can only be explained by the 'deep' grammatical property of embedding. Once we accept that embedding exists, our results allow us to draw a clear distinction between serial postmodification and embedding, while questioning the encoded distinction between serial postmodification and co-ordination.

5.2 Towards the evaluation of grammar

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

Our evaluations throughout this paper have been both analytical – evaluating patterns in abstracted data – and concrete – carefully reviewing individual examples. It would be a mistake to assume that a framework 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.

Experiment 5 demonstrates that interaction may be detected along grammatical axes where assumptions of lexical adjacency and distance would be wholly misleading. We detect distinct patterns of interaction along independent axes: embedded and sequential postmodification. We contrasted speech and writing, which revealed different patterns in each case. We posited that the grammatical model allowed us to detect effects due to both internal psycholinguistic and external communicative constraints.

We are also able to obtain evidence suggesting that a distinction between sequential and co-ordinated postmodification is probably superfluous and misleading.

In conclusion, the evaluation of additive probability along grammatical axes of construction appears to be a useful method for the empirical verification of grammar. We supplement an approach of contrasting frameworks based on the retrievability of linguistic

events with one based on the retrievability of linguistic interactions. From this perspective, a scientifically 'better' grammar is a more predictive theory – one whose explanatory power can be shown to be empirically stronger (reliably explaining more results, or enabling us to capture more phenomena) than another.

Within the scope of this paper, it was not possible to compare two independent grammatical frameworks. We would need to compare closely comparable corpora with different schemes, and ideally, we would wish to compare schemes applied to the same data (a 'multi-parsed corpus', van Zaanen *et al.*, 2004). However, we were able to show the potential for empirical evaluation of step-wise refinement of a scheme.

A multi-parsed corpus is not required to examine variations in the analysis of a particular construction or group of constructions. To evaluate the effect of removing a distinction we may change the definitions of queries (a process termed 'abstraction' or 'operationalisation').[17] If we need to make a new distinction, additional processing or manual annotation may be necessary, but need only be applied to a limited amount of data.

Irrespective of the 'correctness' of one or more frameworks, this demonstration of the explanatory power of a particular framework can be seen as distinctive evidence for the existence of grammatical structure, conceived of as a set of constraints on speaker decisions to construct sentences that obtain a recursive tree structure.

Acknowledgments

Thanks are due to Bas Aarts, Joanne Close, Stefan Th. Gries, Kalvis Jansons, Evelien Keizer, Gerry Nelson, Robert Newcombe, Gabriel Ozón and two anonymous reviewers for comments on versions of this paper and its methodology. Although the method itself is novel, the perspective is based on discussions with linguists and other scientists over several years, and there are too many people to thank for their contributions.

The ability of researchers to generalise from richly-annotated corpora is based on the quality of that annotation. It is to the body of hard-working linguists who construct, annotate and correct our corpora that this paper is dedicated.

Notes

1. For example, suppose the complementation pattern Subject-Verb-Indirect Object (e.g. *he told her*) is found in a corpus. Should it be treated as a legitimate transitive type ('dimonotransitive') to be added to a framework, or should the 'missing' direct object be considered mandatory, so that instances are categorised as incomplete ditransitive complementation patterns? See Wallis (forthcoming) for a discussion.

2. Experiments on corpus data are not 'true experiments' in the sense that conditions cannot be manipulated by researchers. Sheskin (1997: 18) calls these 'natural experiments'. Corpus linguistics 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 method include ecological soundness or 'naturalism'. This does not rule out conclusions from observations, but it is difficult to decide that a specific explanation is the sole cause of an observed distribution. However, we can identify phenomena and explore potential hypotheses for them. Corpus linguistics 'experiments' should be seen as complementary to true laboratory experiments.

3. 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.

4. This confidence interval estimates the likely range of observations obtained by repeated experimental runs with a Binomial statistical model. In other words, were we to repeat our experiment with new sample corpora 100 times, it predicts that in only 5 out of 100 experiments the result would fall outside the error bar.

5. To investigate this hypothesis further we would need to review 200,000 NPs, track entities through a text and eliminate secondary references, a task beyond the scope of this paper.

6. This observation echoes Church (2000), who identified topic words by comparing the recurrence probability of words appearing in a text with their global probability in a corpus. It assumes the increased chance of a second or third instance of the word was primarily due to the topic of the text.

7. Intriguingly, once the decision to add an AJP is made, the proportion appears to *increase* slightly, although the result is not significant. Were this to be substantiated it might represent evidence of a secondary phenomenon, e.g., that the use of serial adjectives is marked (i.e., the speaker is

consciously making a point by using adjectives). As we shall see, we cannot assume that all interactions suppress a trend (i.e., involve negative feedback), or that there is only one underlying process taking place. See also Experiment 5.

8. Results are obtained using the 'Chart' function in Mark Davies' interface to the BNC, available at http://corpus.byu.edu/bnc.

9. The most likely explanation is it is due to the use of adjectives rather than adjective phrases, and the fact that due to the size of the corpus, part of speech tags were not corrected by human linguists.

correct examples	42	(57%)	*small enclosed late seventeenth-century formal garden*
misanalysis	16	(22%)	*red light orange light green light*
repetition	3	(4%)	*white white white white white Christmas*
stuttering	10	(14%)	*c-- c-- c-- b-- b—bill*
self-correction	3	(4%)	*poor performa—poor sorry poor performance*

10. The sequential experiment can be performed by relaxing the constraint that the clause must contain a noun phrase. This obtains very similar results and avoids the first methodological issue. However, cases without NPs cannot be embedded in the same way. In the interests of a contrastive analysis we required that in both cases postmodifying clauses would contain at least one NP.

11. Like FTFs, we have presented this tree left to right, with the sentence down the right hand side. Gloss: PC = prepositional complement, NP = noun phrase, DT = determiner, DTP = determiner phrase, DTCE = central determiner, ART = article; NPHD = noun phrase head, N = noun, NPPO = NP postmodifier, CL = clause, A = adverbial, AVP = adverb phrase, AVHD = adverb head, ADV = adverb, SU = subject, PRON = pronoun, VB = verbal, VP = verb phrase, OP = operator, AUX = auxiliary verb, MVB = main verb, V = verb, OD = direct object, AVB = standalone auxiliary verb.

12. See www.ucl.ac.uk/english-usage/projects/next-gen.

13. The starting point $p(1)$ increases slightly from 0.0523 to 0.0539 as a result. See Tables 4 and 5.

14. ICECUP IV provides a categorisation procedure that works like this. Each outcome type is associated with a number of FTFs. If one of the Type 1 FTFs matches the example, we count this instance as belonging to Type 1. Only if no matches are found to be examine FTFs under Type 2, and so on. This addresses issue (i) and issue (iii). Partly as a result of the research in this paper, the author since adapted the 'standard' ICECUP (3.1.1) to address the third problem by another method: a switch that allows an FTF to match co-ordinated nodes as if they were individual ones. See www.ucl.ac.uk/english-usage/resources/ftfs/ftfs2.htm.

15. Differences in observations $p(x)$ are statistically significant for x=2 and x=3, tested using a Newcombe-Wilson test at an error level of α=0.05 (Wallis 2013). There is also a significant

difference between the two declines, $d = p(3) - p(2)$, when tested for separability of goodness of fit results (Wallis, 2019).

16. A close examination of the 166 cases of double postmodification finds 9 ambiguous cases that might be embedded. See below. 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 ...studying dance][that… found contact work]* [S1A-002 #150]
abstract vs. concrete	16 (10%)	*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%)	*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	7 (4%)	*immensely Christian gentleman [as ever chiselled anybody out of five cents][who taught his Sunday school class in Cleveland]* [S1B-005 #176]
genuinely ambiguous	9 (6%)	*colleague [who's a pensioners' worker][who isn't doing this…]* [S1A-082 #102]

17. In our '3A' model of corpus research (Nelson *et al*., 2002; Wallis, forthcoming), 'abstraction' is a process that maps corpus instances to concepts in the researcher's framework. A concept like a 'complex NP' may not be defined in the annotation scheme, but the researcher may create an FTF, or series of FTFs, to obtain instances of them, and thereby create a dataset that can then be analysed. The researcher is not caught in the 'hermeneutic trap' of depending on the given corpus framework. This observation implies that even without parsing ICE-GB again with another framework (and reviewing and correcting the annotation exhaustively), we may consider, as a thought experiment, whether or not this alternative framework includes the necessary concepts required to extract patterns we find in this paper and elsewhere.

References

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

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.

Beaman, K. (1984). Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.), *Spoken and Written Language: exploring orality and literacy* (pp 45-80). Norwood, N.J.: Ablex.

Böhmová, A., Hajič, J. Hajičová E. & Hladká, B. (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (Ed.). *Treebanks: Building and Using Parsed Corpora* (pp 103-127). Dordrecht: Kluwer.

Carroll, J., Minnen G., & Briscoe, T. (2003). Parser evaluation: using a grammatical relation annotation scheme. In A. Abeillé (Ed.). *Treebanks: Building and Using Parsed Corpora* (pp 299-316). Dordrecht: Kluwer.

Church, K.W. (2000). Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2. In ACL (Eds.), *Proceedings of Coling 2000 Volume 1* (pp 180-186). San Francisco: Morgan Kaufmann.

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

Fang, A. (1996). The Survey Parser, Design and Development. In S. Greenbaum, (Ed.), *Comparing English Worldwide* (pp 142-160). Oxford: Clarendon.

Feist, J. (2011). *Premodifiers in English: Their Structure and Significance*. Cambridge: CUP.

Greenbaum, S., & Ni, Y. (1996). About the ICE Tagset. In S. Greenbaum, (Ed.), *Comparing English Worldwide* (pp 92-109). Oxford: Clarendon.

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.

Leech, G. (1992). 100 million words of English: the British National Corpus. *Language Research 28*(1), 1-13.

Garside, R., Leech, G. & Sampson, G. (Eds). (1987). *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R. & Leech, G. (1991). Running a grammar factory: the production of syntactically analysed corpora or ‘treebanks’. In S. Johansson & A.-B. Stenström

(Eds.), *English Computer Corpora: Selected Papers and Research Guide* (pp 15-32). 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, A., Ferguson, M., Katz J., & Schasberger, B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. *Proceedings of the Human Language Technology Workshop*. San Francisco: Morgan Kaufmann.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*(2), 81–97.

Moreno, A., López, S., Sánchez, F., & Grishman, R. (2003). Developing a Spanish Treebank. In A. Abeillé (Ed.). *Treebanks: Building and Using Parsed Corpora* (pp 149-163). Dordrecht: Kluwer.

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.

Oflazer, K., Say, B., Hakkani-Tür, D. Z., & Tür, G. (2003), Building a Turkish Treebank. In A. Abeillé (Ed.). *Treebanks: Building and Using Parsed Corpora* (pp 261-277). Dordrecht: Kluwer.

Pickering, M. & Ferreira, V. (2008). Structural priming. *Psychological Bulletin 134*, 427–459.

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

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

Sinclair, J. (1987). Grammar in the Dictionary. In J. Sinclair (Ed.), *Looking Up: an account of the COBUILD Project in lexical computing* (pp 104-115). London: Collins.

Tannen, D. (1987). Repetition in conversation: toward a poetics of talk. *Language 63*, 574–605.

Wallis, S.A. & Nelson, G. (1997). Syntactic parsing as a knowledge acquisition problem. In Plaza, E., & Benjamins, R. (Eds.) *Knowledge Acquisition, Modeling and*

Management. Proceedings of the 10th European Knowledge Acquisition Workshop (pp 285-300). Berlin: Springer Verlag.

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 A. Abeillé (Ed.). *Treebanks: Building and Using Parsed Corpora* (pp 61-71). Dordrecht: Kluwer.

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 (pp 738-759). Berlin: Mouton de Gruyter.

Wallis, S.A. (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics 20*(3), 178-208.

Wallis, S.A. (2014). What might a corpus of parsed spoken data tell us about language? In L. Veselovská & M. Janebová (Eds.), *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Olomouc: Palacký University. 641-662.

Wallis, S.A. (2019). Comparing χ^2 tables for separability of distribution and effect. Meta-tests for comparing homogeneity and goodness of fit contingency test outcomes. *Journal of Quantitative Linguistics* (DOI: 10.1080/09296174.2018.1496537).

Wallis, S.A. (forthcoming). Grammar and corpus methodology. In B. Aarts, G. Popova & J. Bowie (Eds.), *The Oxford Handbook of English Grammar*. Oxford: OUP.

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

van Zaanen, M., Roberts, A. & Atwell, E. (2004). A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In Kranias, L., Calzolari, N., Thurmair, G., Wilks, Y., Hovy, E., Magnusdottir, G., Samtiou, A., and Choukri, K. (Eds.) *Proceedings of LREC'04 Workshop on The Amazing Utility of Parallel and Comparable Corpora* (pp 58-61). Lisbon, Portugal: ELRA.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison Wesley.

Investigating the additive probability of repeated language production decisions

Address for correspondence

Name	Sean Wallis
Department	Survey of English Usage
Institution	University College London
Address	Gower Street
City, post code	London WC1E 6BT
Country	UK
Email address	s.wallis@ucl.ac.uk