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.

ey o d 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 depen(g)500]TJ 80439(-2.1

[†]This version of the paper includes markup in yellow to highlight differences with Wallis (2019), IJCL 24:4.

Some examples of corpus grammars

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'

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

Capturing linguistic interaction in a parsed corpus

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 **ne e feed c oop**

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 **pec f c** 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 coo2(y)500]T39(i)-2.16436(a)3.74(b)-0.295585(t)-26(n8y)161caen to88

[††]Wallis (2019) finds a gradual decline $p(2) < p(1)$.

4. Grammatical interaction between postmodifying clauses

4.1 Embedded 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

4 D e n

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,

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 **e n c co e ence** (possibly revealed by clusters of co-occurring

Capturing linguistic interaction in a parsed corpus

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 **e en** with one based on the retrievability of linguistic **n e c on** . 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 e e ed non c c . p oof of ’

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.

- **A o on** A pattern of linguistic interaction reflects the structure in addition to

position). It should also be possible to measure di

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,

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, $\boldsymbol{p}(\boldsymbol{x})$, is approximately Normal. This holds for large samples where $\boldsymbol{p}$ is not close to 0 or 1. The formula for the Normal ('Wald' or Gaussian) interval approximation is

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

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

However, in our data $\boldsymbol{p}(\boldsymbol{x})$ can indeed be very close to zero. As we have already seen, the sample size, $\boldsymbol{F}(\boldsymbol{x})$, falls rapidly. As $\boldsymbol{p}$ 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

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

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).

A [illegible] nfcnce en cnen p[illegible]

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. $\boldsymbol{p}$ is constant between decision $\boldsymbol{x}$ and $\boldsymbol{x}$+1, $\boldsymbol{p}(\boldsymbol{x}) \approx \boldsymbol{p}(\boldsymbol{x}+1)$.

The simplest approach if we are plotting Wilson intervals is to test if the expected value, $\boldsymbol{p}(\boldsymbol{x})$ is outside the confidence interval for $\boldsymbol{x}$+1. This test, sometimes referred to as 'the $\boldsymbol{z}$ 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 $\boldsymbol{F}(\boldsymbol{x})$. This can be laid out as follows.

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