Are embedding decisions independent?
Evidence from preposition(al) phrases

Sean Wallis
Survey of English Usage
May 2022

Abstract

One of the more difficult challenges in linguistics research concerns detecting how constraints might apply to the process of constructing phrases and clauses in natural language production. In previous work (Wallis 2019) we considered a number of operations modifying noun phrases, including sequential and embedded modification with postmodifying clauses. Notably, we found a pattern of a declining additive probability for each decision to embed postmodifying clauses, albeit a pattern that differed in speech and writing.

In this paper we use the same research paradigm to investigate the embedding of an altogether simpler structure: postmodifying nouns with prepositional phrases. These are approximately twice as frequent and structures exhibit as many as five levels of embedding in ICE-GB (two more than are found for clauses). Finally the embedding model is simplified because only one noun phrase can be found within each prepositional phrase. We discover different initial rates and patterns for common and proper nouns, and certain subsets of pronouns and numerals. Common nouns (80% of nouns in the corpus) do appear to generate a secular decline in the additive probability of embedded prepositional phrases, whereas the equivalent rate for proper nouns rises from a low initial probability, a fact that appears to be strongly affected by the presence of titles.

It may be generally assumed that like clauses, prepositional phrases are essentially independent units. However, we find evidence from a number of sources that indicate that some double-layered constructions may be being added as single units. In addition to titles, these constructions include schematic or idiomatic expressions whose head is an ‘indefinite’ pronoun or numeral.

Keywords: additive probability, interaction evidence, embedding

1. Introduction

In (Wallis 2019), we described a research design which considered the additive probability for repeatedly performing the same construction step. To take a simple example used in the paper, we might evaluate the additive probability of repeatedly adding an attributive adjective phrase to a noun head, according to a canonical scheme that looks like this.

$$\begin{tikzcd}[column sep=small] \mathit{base} \arrow[r] \arrow[dr] & +\ \mathit{term}_1 \arrow[r] \arrow[dr] & +\ \mathit{term}_2 \arrow[r, dotted, no head] & {} \arrow[r] \arrow[dr] & +\ \mathit{term}_n \\ & \varnothing & \varnothing & & \varnothing \end{tikzcd}$$

We examine the probability of adding the x-th term (in this case, an attributive adjective phrase), which we label $p(x)$, to an existing string (a noun head). Thus we obtain results like

the cat
the black cat
the large black cat
etc.

In such a construction, decisions are not necessarily made in the order in which they appear: the speaker conceivably assembled a mental model of the ‘cat’ that they wished to communicate, and then selected adjective phrases before arranging them into a noun phrase sequence. Or they could have simply avoided attributive adjectives altogether and said *the cat that was black and large*.

Nonetheless, we can calculate the probability in our data that a speaker or writer adds an attributive adjective phrase, at point x, before the noun, which we will simply label $p(x)$. So $p(1)$

represents the chance of adding the first adjective phrase, $p(2)$ the chance of adding the second given the first, and so on.

We first obtain a frequency distribution of *at least x* adjective phrases, $F(x)$. We can then simply divide $p(x) = F(x)/F(x-1)$ to obtain the additive probability at each stage.

Each set is a subset of the previous one in the sequence. The set of noun phrases with at least one attributive adjective phrase is a subset of all noun phrases, etc. We test for a significant fall or rise by examining whether one additive probability point $p(x-1)$ is within the Wilson score interval (Wilson 1927) for the next, $p(x)$. This is a 'goodness of fit' test condition.

If the additive probability does not significantly change as repetition x increases, then it means that we have no evidence of an interaction between one decision and the next. With confidence intervals we can also consider the size of effect (maximum and minimum slope) at any point.

In the paper we found a serial and sequential impact with attributive adjective phrases, with a declining probability observed with each repetition. We suggested three possible types of explanation, which are not necessarily exclusive.

1. **Logical-semantic constraints**, such as semantic ordering of adjectives and semantic coherence (so one tends to say *large black cat* rather than *black large cat* or *large small cat*).
2. **Communicative economy**. The communicative environment imposes constraints. For example, a random sample of nouns will include second, third references to previously introduced (and described) concepts. These subsequent references will likely be adjectiveless ([*the*] *cat*), or a pronoun (*he/she/it*).
3. **Cognitive memory/processing constraints**, which were originally primarily conceived of as having a negative impact (hence 'constraint'). However, mental processing may also make certain expressions easier than others to produce.

Note that we are not primarily concerned about the influence of a particular noun on a particular modifier, such as avoiding colour adjectives with abstract nouns (cf. *a black mood*). Rather, our focus is on the hypothesis that the cumulative impact of previous operations causes an additive probability to fall – or, in some cases, rise.

Indeed, in the case of repeated postmodifying clauses following the noun head, there was an initial decline and then a rise. This subsequent increase seemed most likely due to *templating*, a tendency to re-use structures. Consider this example:

(1) …the dream becomes *a text* [*to renarrate*], [*to revise*], [*to listen to*], [*to read*], [*to analyse*]. [W2A-002 #33]

In the case of adjective phrases, the pool of plausible compatible adjectives tended to be used up, suggesting Explanation (1) above. But the same does not appear to apply to clauses following the noun head. Indeed, for communicative purposes, one might imagine someone attempting to convey a particular location or event by repeatedly adding postmodifying clauses if their interlocutor seemed puzzled.

Example (1) is a type of 'asyndetic coordination' (coordination without a coordinator 'and', 'or' or 'but'). In the paper, we took additional steps to count coordinated examples. We would not wish to treat Example (1) differently were the writer to conclude with <u>*and*</u> *to analyse*. Pooling serial postmodification and coordinated cases, it became clearer that the pattern did adopt a 'fall and rise' pattern, suggesting that there was a second phenomenon at work in the case of longer strings.

1.1 Embedding

Wallis (2019) compared sequences comprising serial postmodification of the same head with *embedded postmodification*, i.e. where a postmodifying clause includes a noun head, and *that head* is then postmodified. This appeared to demonstrate a decline – from $p(1)$ to $p(2)$ for spoken data, and from $p(2)$ to $p(3)$ for writing. See Figure 1.

Since each additional term modifies this new head, we may make a default assumption that decisions are made as the construction is assembled in sequence, i.e. in order of increasing depth. The reasoning is that the second addition can only be made after the first has been added.

Unfortunately for our study, postmodifying clauses are not very often embedded, and we rapidly ran out of data, despite a starting point of some 190,000 noun phrases!

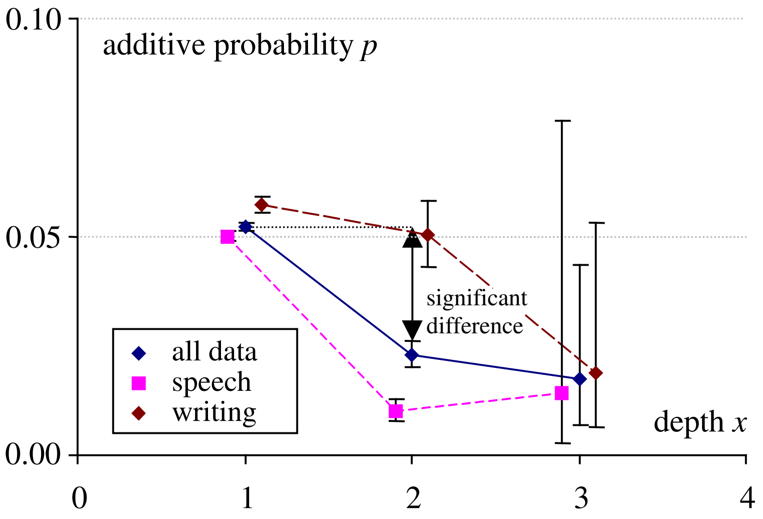


Figure 1. Studying the impact of cumulative cost on embedding, noun phrase postmodifying clauses with head nouns, after Wallis (2019). We have insufficient data to determine whether this trend would continue at further levels of embedding, although we note the difference between speech and writing. A significant difference between additive probabilities in a series is found when $p(x-1)$ is outside the interval for $p(x)$.

The task of extracting and counting embedded structures relies on the combination of a fully parsed corpus, the *British Component of the International Corpus of English* (ICE-GB, Nelson et al. 2002), and an effective search tool, ICECUP. Whereas one can obtain adjective sequences from an unparsed corpus, and possibly even attempt to recover serial postmodification from such a source, retrieving and counting embedded terms requires a parse analysis.

1.2 Why are preposition(al) phrases interesting?

Figure 1 identifies what appears to be a genuine ‘cost’ of embedding, but we cannot determine whether the initial trend is a general one, or is limited to a difference between first and second order embedding. The observed difference in speech and writing might be due to processing cost or differing communicative strategy. It might also be due to topic differences in speech and writing subcorpora, although this appears less likely.

In this paper we suggest that a different structure, *prepositional phrases* (PPs, also called ‘preposition phrases’), may be more fruitful for evaluating the processing costs of embedding.

The first motivation for adopting PPs is that it is possible to find longer chains of embedding in the million-word ICE-GB. Consider Examples (2) and (3), which are viable (and readily interpreted) embedded strings. These are 5-deep structures, i.e. structures with two further levels of embedding than we find for clauses.

(2) So consultation and cooperation with the public <,> as well as speed and *a sympathetic understanding* [*in response* [*to calls* [*for help* [*from the victims* [*of crime*]]]]] and a physical presence on the streets are what the public now seek of the police <,> [S2B-037 #16]

(3) *The introduction* [*of the Independent Police Complaints Authority* <,> [*with its wide powers* [*of intervention and supervision* [*of the investigation* [*of complaints*]]]]] can but reassure the doubting public <,> [S2B-031 #69]

In Examples (2) and (3), each of the noun phrase heads are all nouns, and not pronouns, numerals, nominal adjectives or proforms. In Section 2.4 we examine what happens if we relax the requirement for heads to be nouns,[1] and in Section 3 we allow conjoins in embedded paths.

A second motivation is that, since postmodifying PPs are of higher frequency, we may be able to examine more grammatical subcategories or text categories.

Finally, unlike clauses, which may have a noun phrase acting as a subject, object or complement, only one element, the prepositional complement, can be a noun phrase. The overall model is streamlined and simpler.

2. Experiments excluding embedded conjoins

In this paper we will conduct experiment in two phases. First we will consider a subset of cases where we exclude cases where the embedded sequence contains one or more conjoined element, as in Example (3). These can be identified and counted with ICECUP 3.

We include conjoins in our data in Section 3.

2.1 Obtaining data

Consider the Fuzzy Tree Fragment (FTF) in Figure 2. This is an example of one-level embedding under a node that is not itself a prepositional complement noun phrase (marked '(¬PC, NP)'). As a result, the FTF must be at the start of an embedding sequence. In this first experimental phase, we also exclude *conjoined* complement NPs by creating a second FTF and subtracting these cases.

Note that we can permute the wordclass categories in the first NPHD 'slot' (circled) to obtain a series of different queries. Thus we apply each query across all data in ICE-GB to obtain frequency data in Table 1. We will also distinguish common and proper nouns.

Let us first identify the 'base' element in our scheme. Figure 3(a) depicts an FTF consisting of the two nodes in the first row of Figure 2 ('(¬PC, NP)' plus the designated head), and Figure 3(b) illustrates the parallel structure for subtracting conjoined prepositional phrases ('PC, NP (CJ, NP)'). (To find conjoined cases for Figure 2 we replace the topmost node with these two nodes.)

We can now obtain frequency data for $f = F(1)$ and $n = F(0)$, and compute the additive probability or *modification rate*, $p(1) = f / n$. See Table 2.

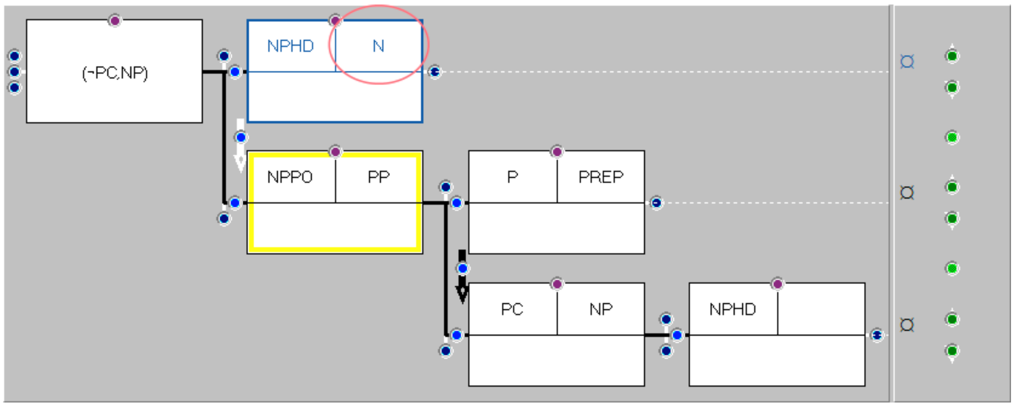
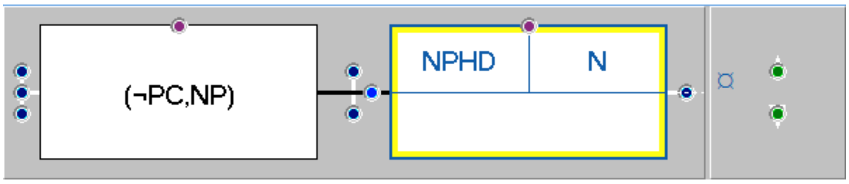


Figure 2. Level 1 Fuzzy Tree Fragment (FTF) for a head noun that is not in a prepositional complement noun phrase, which is followed by a prepositional phrase acting as a noun phrase postmodifier (NPPO, PP), which in turn consists of a preposition (P, PREP) and its complement NP (PC, NP) with a head (NPHD). The FTF is drawn left-to-right rather than top-down for ease of visualisation, and the matching 'sentence' would be read from the top, down on the right.

[1] The ICECUP software matches FTFs against tree structures by assigning a single grammatical node in a corpus tree analysis to a single node in the FTF. But Example (3) illustrates two important exceptions to this rule. It includes a compound proper noun, the *Independent Police Complaints Authority*, which matches a single node. It also includes a co-ordinated pair of noun phrases, *intervention and supervision*, which matches the 'PC,NP' node. This is found by an additional method discussed in Section 3.

(a)



(b)

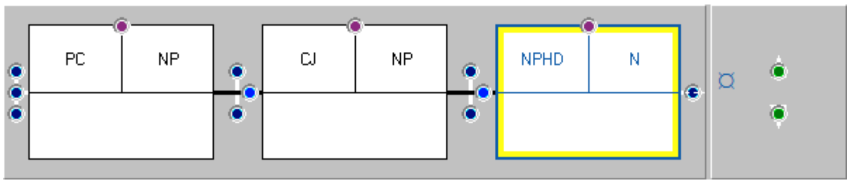


Figure 3. FTFs for obtaining our ‘base’ term, ‘level 0’, restricted by nouns (N). The first query (a) finds all nouns that are noun phrase heads which are not found in prepositional complement NPs. This FTF will be at the start of an embedding sequence, if we also exclude cases matching the second FTF (b), which finds cases of conjoined prepositional complement NPs.

		N(com)	N(prop)	PRON	NUM	NADJ	PROFM
postmodifier frequency	$f = F(1)$	13,441	565	1,489	944	54	1
head frequency	$n = F(0)$	83,079	21,224	89,244	7,132	628	591
additive probability	$p = f / n$	0.1618	0.0266	0.0167	0.1324	0.0860	0.0017

Table 1. Frequency distributions of topmost PP structures across both speech and writing in ICE-GB, subdivided by the wordclass of the initial noun phrase head (cf. Figure 3). Nouns, common and proper, comprise more than 90% of the data. We also calculate the additive probability *p*.

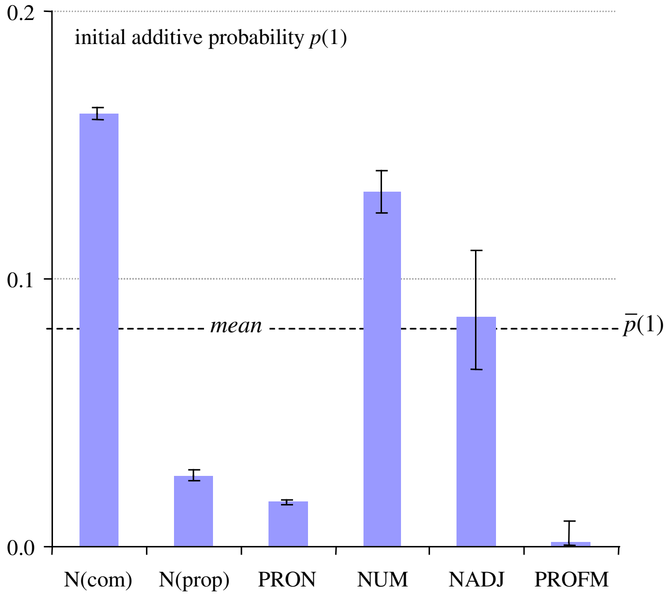


Figure 4. Variable rates of initial postmodification, with 95% Wilson score intervals. Common nouns have the highest rate of PP postmodification, followed by numerals.

We plot this initial rate of postmodification in Figure 4. The mean rate (i.e. the rate for all data unspecified by head) is 0.0817, with a 95% Wilson interval of (0.0805, 0.0829).[2]

2.2 Nouns postmodified by prepositional phrases

First, we create a sequence of FTFs of increasing complexity, which we label level 0, 1, 2, 3 etc. We also create a second FTF with the conjoined prepositional phrase at the top.

[2] For nouns containing nouns, the rate is 0.1385 ∈ (0.1327, 0.1445), which is more than double the equivalent rate for postmodifying clauses (0.0556 ∈ (0.0546, 0.0577)).

The base term, level 0, is obtained by the FTFs in Figure 3, which find all cases of 'NPHD, N' (noun phrase head, noun) not found in a prepositional complement NP or in the conjoined version of the same. By subtracting the second set of query results from the first, we obtain a set of all base items with total frequency *F*(0).

Level 1 is the same as this base term, where the noun is followed by a prepositional phrase acting as a noun phrase postmodifier (NPPO, PP), consisting of a preposition (P, PREP) and its complement NP (PC, NP), which itself has a head (NPHD). See Figure 2. This yields *F*(1) after subtraction of conjoined prepositional complement cases.

For level 2, we use an FTF like Figure 5 for two-layer embedding. This matches examples like (4) below, and (after subtraction of conjoins) obtains *F*(2).

(4) Oh *O* [*with a circumflex* [*over the top*]] [S1A-009 #247]

This 'level 2' FTF consists of two single-layer cases, each indicated by the dashed lines in Figure 5.

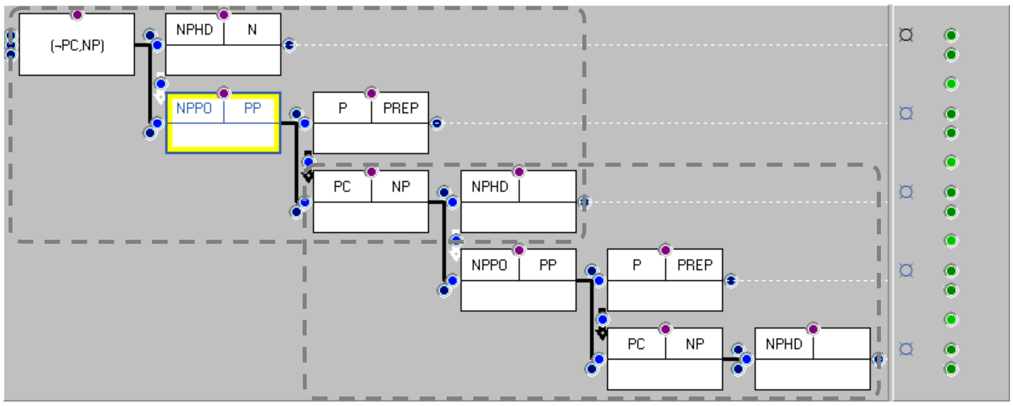


Figure 5. Two-level FTF schema, which shows how further levels are added below.

depth *x*		0	1	2	3	4	5
ICE-GB	*F*(¬PC,NP)	114,938	15,210	2,134	229	18	3
	F(PC,NP (CJ,NP))	10,063	1,154	113	8	0	0
	F	104,875	14,056	2,021	221	18	3
spoken	*F*(¬PC,NP)	58,609	6,559	824	77	5	1
	F(PC,NP (CJ,NP))	4,033	436	45	3	0	0
	F	54,576	6,123	779	74	5	1
written	*F*(¬PC,NP)	56,329	8,651	1,310	152	13	2
	F(PC,NP (CJ,NP))	6,030	718	68	5	0	0
	F	50,299	7,933	1,242	147	13	2

Table 2. ICE-GB raw frequency data matching FTFs of at least depth $x = \{0, 1,\ldots 5\}$ where noun phrase heads are specified as nouns. The final row, *F*, is obtained by subtracting the second query from the first.

We continue to add these units until we have a level 6 FTF, which exhausts the data. Note we have built in some flexibility: the PP is not obliged to immediately follow the noun head (denoted by the white arrow). ICECUP treats compound nouns as matching a single node by default. Subdivided into speech and writing subcorpora, this process obtains the raw frequency data in Table 2. For now, we require that no item within these paths is conjoined.

The additive probability is the chance that having added *x*-1 terms already to the base, the speaker or writer will add another.[3] The distribution expresses the frequency of *at least x* levels of

[3] To analyse this type of data we must consider a number of steps, see Wallis (2019: 506-507). In this phase we restricted data to exclude the prepositional complement, so we have already identified base frequencies and we don't need to subtract embedded matching cases. For the purposes of our study, and to make it simpler for the reader to

embedding. It is the total number of cases in a superset containing all the cases at the level below, *including* the level below that, etc.) We convert these frequencies, $F(x)$, to additive probabilities as noted earlier:

$$p(x) = F(x)/F(x - 1),$$

and compute Wilson score intervals for each observation. This gives us the graph in Figure 5. We see a non-significant difference and then a fall in the written data ($p(3) < p(2)$). The spoken data indicates an initial rise and then a fall.[4]

We can read Figure 6 as evidence for a significant decline from the postmodification rate for level 2 (cf. the FTF in Figure 5) to level 3 in each of the three trend lines, but something different appears to be taking place between level 1 to 2.[5]

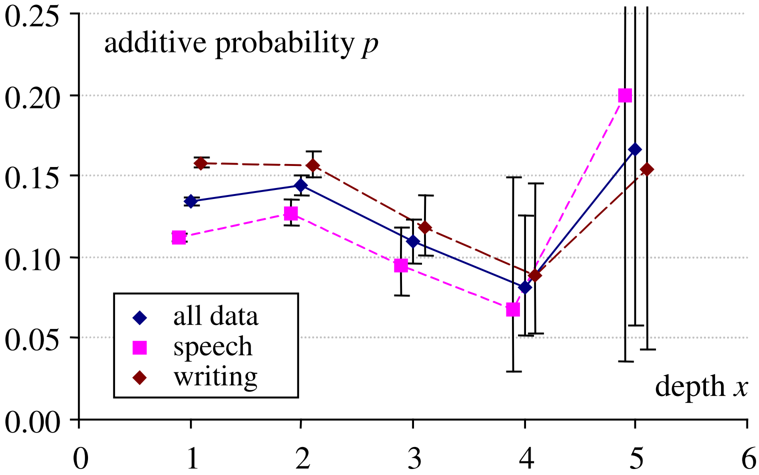


Figure 6. Variation in additive probability p for embedding depth x, nouns postmodified by PPs, with 95% Wilson score intervals.

2.3 Common and proper nouns

Subdividing the data into structures postmodifying common and proper nouns obtains the graph in Figure 7. The general trend for nouns now appears to be due to the sum of an initial rise in additive probability for proper nouns ($p(2) > p(1)$) combined with a secular decline in the rate for common nouns from $p(1)$ to $p(3)$, in writing at least.

The wordclass of the initial postmodified head makes a big difference to the subsequent pattern.

We can read this graph as saying that proper nouns are postmodified overall at a much lower rate (which we might expect), but that the postmodification rate for this subsequent embedded postmodified head converges with the equivalent rate for common nouns.

This should not be surprising, as the principal association of a second order embedding will be with its immediate head, not a previous one. However, there appears to be a residual effect: written data still exhibits a significantly lower rate for proper nouns than common nouns at $x = 2$.

reproduce, we will not include coordinated cases. The method described here is not the most exact, and counts instances of multiple postmodification independently, slightly elevating the postmodification rate. However examination of cases reveal that this has a marginal effect.

[4] We subdivide data into speech and writing for two reasons. Firstly, whereas speech is mainly spontaneous (and even scripted speeches are uttered spontaneously), writing permits editing and review. Superficially at least, we tend to see trees that are shaped differently in speech and writing subcorpora. Secondly, the data sets are independent, and thus each acts as a kind of replication study for the other (Wallis 2021: 201).

[5] For the mean of all data (middle line), the probability of adding the structure once, $p(1) = 0.1340 \in (0.1320, 0.1361)$; and twice, $p(2) = 0.1439 \in (0.1381, 0.1497)$. Since $w^{-}(2) > p(1)$, this is a statistically significant rise.

Note also that proper noun heads in the spoken data have a slightly higher initial probability of being postmodified than in the written.

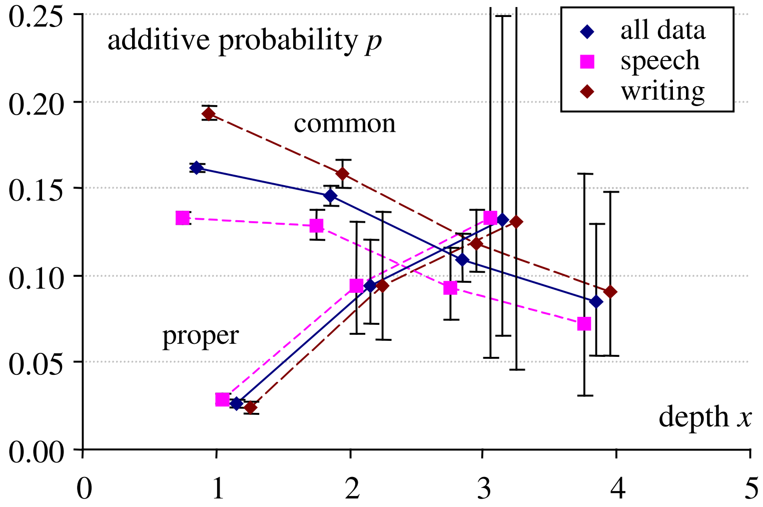


Figure 7. Additive probability trend analysis for embedded PPs following common and proper nouns.

2.4 Heads of any type

As we saw, Wallis (2019) also revealed an initial decline for postmodifying clauses containing nouns, although this phenomenon appeared to differ between speech and writing.

Figure 8 plots the trend for postmodifying NP heads of any type. Again, we see a rise-and-fall effect across the pooled data, which is even more exaggerated than that seen with all nouns. But averages conceal variation, and we should not assume that all subtypes of data behave identically.

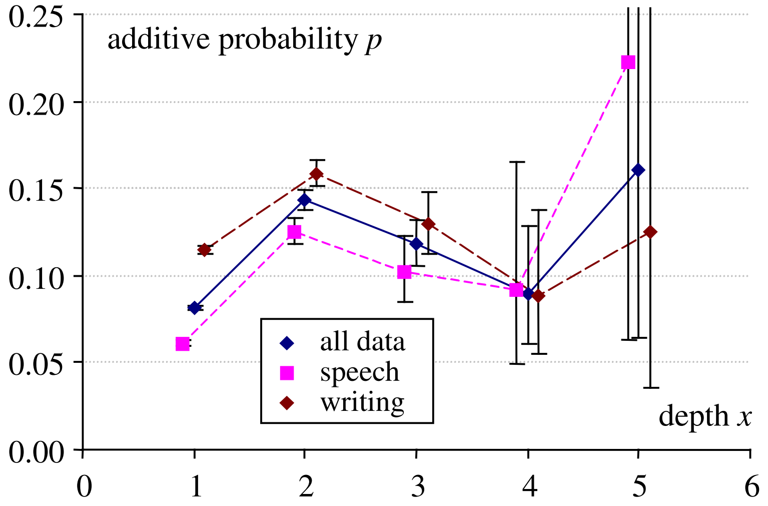


Figure 8. Additive probability analysis of postmodifying PP embedding with initial noun phrase head unspecified.

2.5 Nouns, pronouns and other wordclass types

We have discovered that embedding decisions may interact, but also that level 1 and level 2 decisions (at least) may be influenced by the wordclass (and associated semantics) of the initial node.

What happens if we postmodify pronouns, numerals or nominal adjectives?

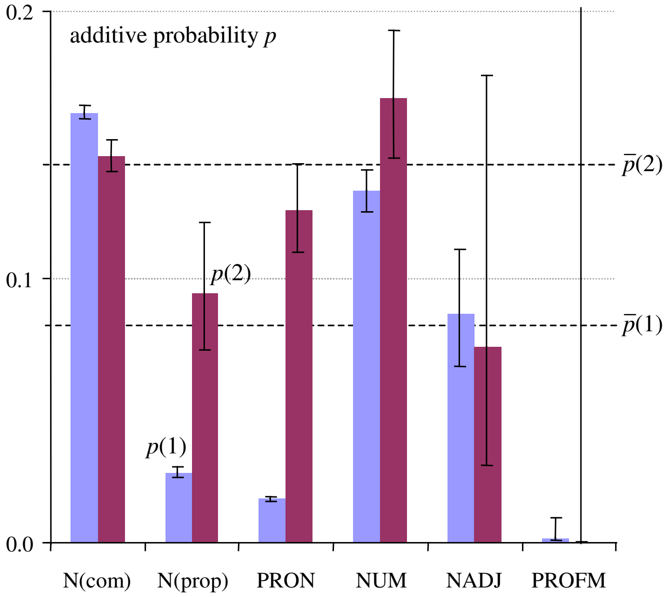


Figure 9. Additive probability rates $p(1)$ and $p(2)$ for different wordclass heads across all ICE-GB data. This indicates a substantial difference between wordclasses in both first and second-level PP embedding. Although rates for $p(2)$ appear to converge, they are far from equal, and several are still significantly different.

Not only do we see differences between postmodification rates for levels 1 and 2, level 2 probabilities are significantly different from each other. Figure 9 reveals that the rate for proper nouns is lower than for common nouns (see also Figure 7).[6] Newcombe-Wilson tests (Newcombe 1998) confirm that the level 2 embedding probability, $p(2)$, also differs for pronouns and common nouns, and pronouns and numerals.

If we continue the same analysis to level 3, a significant fall is found for common nouns and numerals ($p(3) < p(2)$), but the rate does not statistically significantly differ between subtypes.

2.6 Pronoun subtypes

Figure 9 indicates that pronouns have an even more pronounced initial rise than proper nouns. However, PP postmodification varies widely between different classes of pronouns. The main pronoun types that appear to routinely permit postmodification are

- quantifying (41% are postmodified), e.g., *a lot* [*of the problems* [*in singing*]],
- assertive (19%), *some* [*of the implications* [*of this kind of thing*]],
- negative (12%), *nothing less* [*than an arm* [*of the State*]],
- universal (9%), *all* [*of the rocks* [*in the area*]], and
- nonassertive (8%), *anything* [*up to two thousand* [*at a time*]].

These pronouns have one aspect in common: they tend to be unspecific unless postmodified, and may be referred to as ‘indefinite’ pronouns. Similarly, doubly-postmodified numerals are dominated by ‘one’ as in (*one* [*of the strange things* [*about the conference season*]]).

With the exception of quantifying pronouns, where we see a fall in probability from $p(1) = 0.4121$ to $p(2) = 0.1256 \in (0.1043, 0.1506)$, the secondary additive probability, $p(2)$, is not significantly different from the first. So why do we see a rise from $p(1)$ to $p(2)$ for all pronouns?

[6] Where a ‘point’ (end of bar) falls within an interval range they are not significantly different from each other, so e.g., $p(2)$ rates for numerals and common nouns are not significantly different from each other.

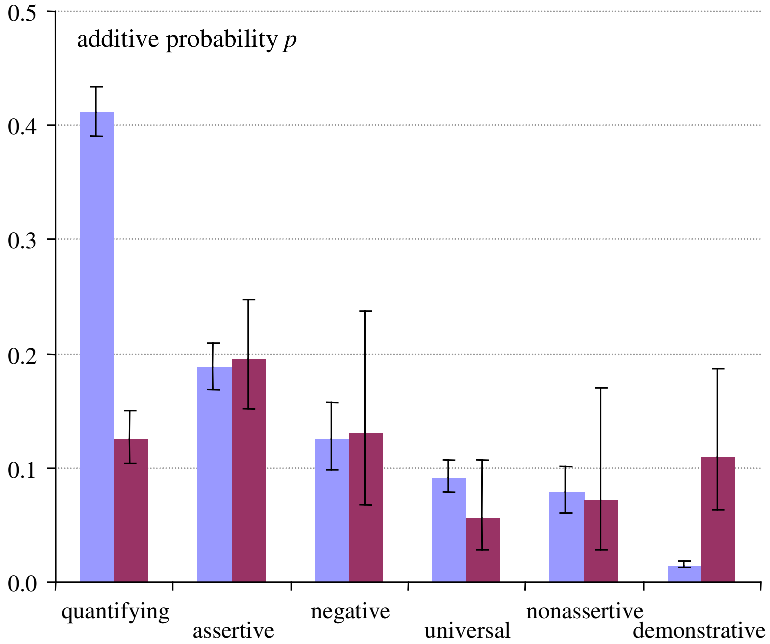


Figure 10. First and second level additive probabilities for pronoun subtypes (all ICE-GB data). With the exception of quantifying pronouns, whose initial rate is high, $p(2)$ is not significantly different from $p(1)$.

The answer turns out to be simple. The most common type of pronoun in the corpus, personal pronouns, is rarely postmodified. Out of 60,465 (67% of pronouns) there are 12 level 1 matches (e.g., *I* [*as the Arts Minister*] or *me* [*on glockenspiel*]) – and *zero* level 2 cases. The initial probability for personal pronouns, $p(1)$, pools all cases of pronouns, so the mean rate is pulled down.

But at the second level, these unmodifiable pronouns are now less than 3% of secondary heads. The rate increases. These rates converge on a mean, as in Figure 7.

In Figure 10 we also include demonstrative pronouns, which have a high rate of double-embedding (e.g., *those* [*of the Court* [*of Appeal*]]) but a low initial rate. Unmodified *that* or *those* is unsurprisingly very common. But once modified, the chance of further postmodification is higher, yielding a similar significant increase that we saw with proper nouns. By inspection of the 11 double-postmodified cases, 9 appear to be independent, with a couple of idiomatic or schematic exceptions (*those* [*with an ear* [*for those things*]], *those* [*under the age* [*of 16*]]).

2.7 Restricting embedded heads

In each case thus far we have not restricted examples according to the type of *embedded* heads. But there is a methodological problem of interpretation if we pool sets of data that behave differently. We saw how some types of initial head, personal pronouns in particular, were almost incapable of postmodification.

Although we might compare subsequent proportions with each other, if the initial category has diverse rates of postmodification (as with all pronouns) it is difficult to fairly compare initial proportions and gradients, where in fact most of the data would be found.

One solution is to restrict each head more strictly. Wallis (2019) insisted that embedded clauses must contain at least one NP with a noun head. Although this reduced the set of cases it ensured that the embedded term was capable of an identical repetition step.

In Figure 11 we repeat the analysis of common and proper noun heads by requiring that every subsequent head is specified as being of the same type as the initial one. Each embedded head in the common noun sequence of FTFs must be a common noun, and each in the proper noun sequence must be a proper noun. Now we can see that although the common noun rate falls slightly for writing between $p(1)$ and $p(2)$, they are not significantly different for spoken data. The previouslyclaimed 'secular decline' (a trend over multiple points) seems more difficult to substantiate when all heads are common nouns.[7]

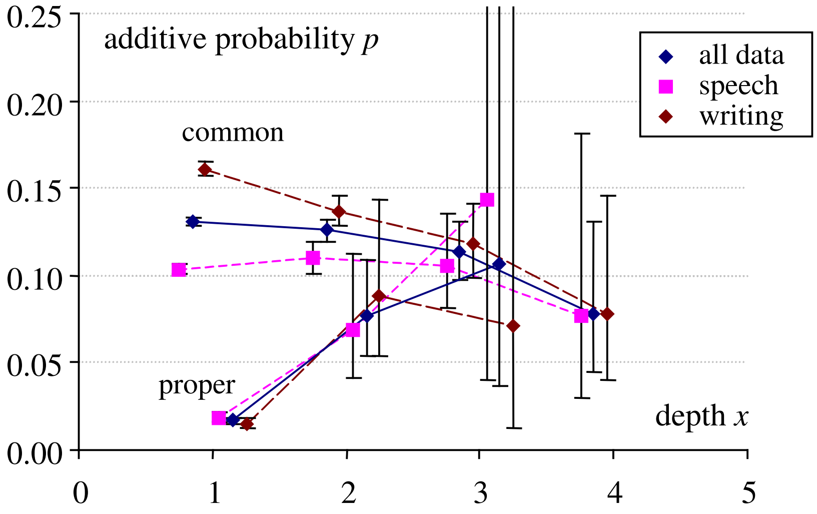


Figure 11. Embedding the same construction type. Common and proper nouns containing common and proper nouns (respectively), and so on. The additive probability for common nouns in writing significantly falls: $p(2) < p(1)$; for proper nouns, it rises: $p(2) > p(1)$, for both categories. Compare with Figure 7, where subsequent heads are not restricted by type.[8]

On the other hand, the rate increases from level 1 to level 2 for proper nouns. Of 27 cases of double-embedding, 18 are full titles (Professor [*of* Politics [*at* Birkbeck]], *the* Royal Society [*for the* Protection [*of* Birds]], etc.) and 11 contain them (e.g., Controller [*of* Radio City [*in* Liverpool]]). Only 2 out of 37 contain arguably independent units (*LA4431* [*about the* *Special Meeting* [*of* *Council*]]). In short, this increase is likely to be due to title compounds.

As a proportion of the data for common and proper nouns *unrestricted* by subsequent embedded type (see Section 2.3), these restricted subsets correspond to around 70-80% of the data for common nouns, and 55-65% of the data for proper nouns. If we generalise to all nouns, the resulting rise-fall pattern is indeed very similar to Figure 6.

If we attempt the same approach for pronouns and numerals we find cases of single-embedding (e.g., some [*of* it], tens [*of* thousands]) but only two examples of double-embedded numerals (two point five [*times* ten [*to* the three / four]]). Like title forms, these appear either idiomatic or schematic.

3. Experiments allowing embedded conjoins

One of the limitations of the experiments we have discussed thus far concerns coordination. Consider Example (5).

(5) So without there being a conspiracy there's been *a concentration* [*of mishandling and ineptitude and lack* [*of regard* [*for the main condition* [*of arts functioning* [*in a great metropolitan city*]]]]] which has led to this [S1B-022 #28]

In this example, the conjoined triple *mishandling and ineptitude and lack* [*of regard…*] would not be identified by the FTFs we have seen thus far. As a result, the method in Section 2 has a tendency to underestimate embedding rates.

[7] This is addressed when conjoins are taken into account. See Section 3.

[8] Where intervals are very wide it means that the true rate in the population is extremely uncertain. Although datapoints on the right may appear to be very dramatic, in fact they express 'location unknown'.

3.1 Obtaining data

The latest ICECUP 3.1.1 software permits users to simply apply a switch to additionally match conjoined nodes. We can use this switch to allow conjoined PPs and NPs to be found in the embedding sequence. See Figure 12.

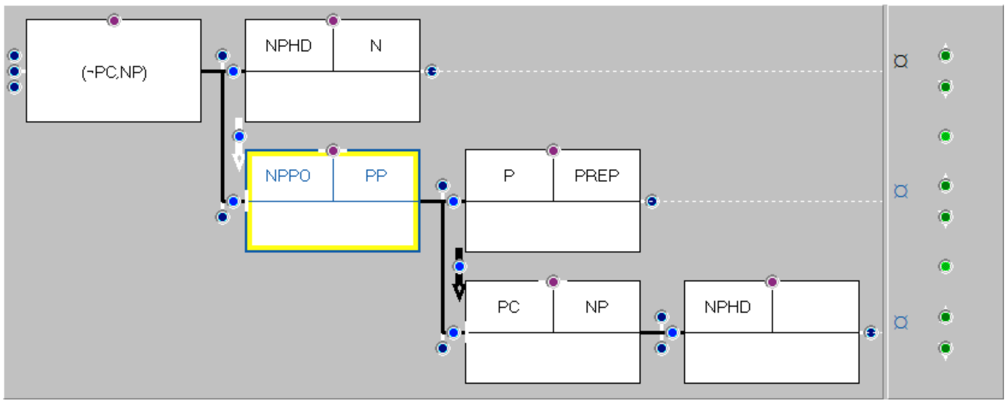


Figure 12. Fuzzy Tree Fragment for single-embedded prepositional phrases (cf. Figure 2), allowing conjoined PPs and NPs to be found in the path (indicated by the broken box).

However, relaxing constraints in ICECUP normally has the effect not just of finding, *but also counting separately*, all distinct matching permutations. Thus for example, Figure 12 finds three matching arrangements that are all found in ostensibly the same case, Example (6).

(6) I certainly believe that *the police system* [*in the United Kingdom*] *and* [*in England and Wales*] should be nationally organised but I also believe that uh many of its services should be locally delivered <,> [S1B-033 #6]

This raises an interesting philosophical point concerning *case overlap*. How should we account for these subsidiary patterns? They are clearly *not* independent from one another, having the same head, *the police system*, but they might differ in other respects. In (Wallis 2022), we identified a large end-weight bias for PP modification of noun phrases. The chance of a final noun phrase conjoin being postmodified (as in Example (5)) was found to be around 6.15 times greater than that for the first conjoin (95% confidence interval: between 4.51 and 8.39 times). For prepositional phrases, this end weight bias was lower, at 4.25 ∈ (2.00, 9.01) times. In both cases we found evidence of 'templating', i.e., where grammatical structure (and cognitive resources) appears to be re-used from conjoin to conjoin. These observations do not really affect our experimental model. Indeed, evidence of templating justifies treating so-called 'independent' conjoin patterns as comprising a single (dependent) case, and counting it once, which we do anyway. The fact that there is an end weight bias is of interest insofar as it tangentially indicates that the introduction of a conjunction within a structure is not a 'stopping condition', preventing further postmodification.

As a result of the additional complexity introduced by coordination, Wallis (2019) used ICECUP IV, which is not available for general use. The standard search facility in ICECUP is *explorative* and exhaustive, that is, it finds every single possible matching arrangement whereby an FTF might match a tree in the corpus. But that behaviour is not what we want in this case.

ICECUP IV includes an alternative search facility designed for systematic data gathering for experiments. The user specifies a distinct single node or group of nodes in a tree, such as the noun phrase head. Every instance matching this node is then classified according to whether or not it is singly-embedded, double-embedded, and so forth. We may have a discrete variable, 'embedding depth' = {0, 1, 2, 3…} with each level specified by a logical combination of FTFs.

Using this method, the noun phrase head *the police system* in (6) counts *once*, and is classed as having a single-level of embedding because none of the noun phrases (*United Kingdom*, *England* or *Wales*) are themselves postmodified.

3.2 Nouns postmodified by prepositional phrases

Let us repeat the analysis performed in Section 2.2, but this time allow cases with conjoins in the embedded path to be included. Data is extracted with ICECUP IV up to and including level 3 embedding. For levels 4 and above we found it was more efficient to use ICECUP 3.1.1 with the relaxation of inheritance rules, and check cases manually.

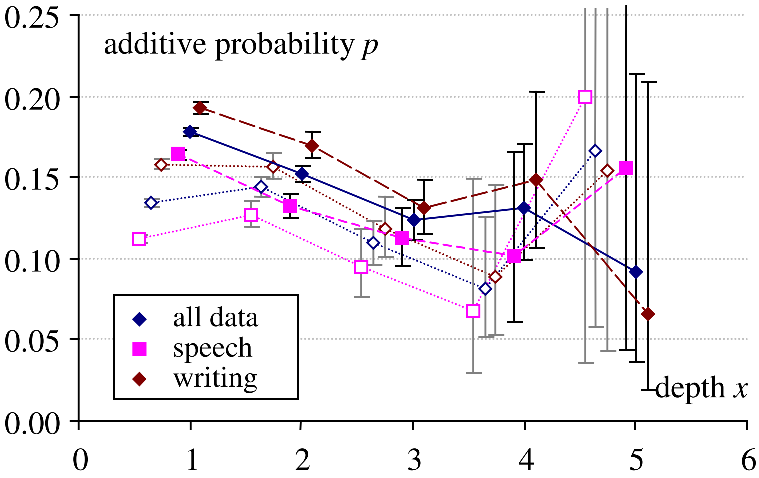


Figure 13. Additive probability for embedding, noun heads, including conjoined cases. For comparative purposes the data from Figure 6 is included (dotted).

Figure 13 reveals that one result of allowing conjoined cases to appear in the path is that the overall rate of addition tends to be higher, as more cases are included (this difference is mostly significant at levels 0 and 1). The downward tendency, which in Figure 6 appeared as a rise and fall, now appears to be a more uniform secular decline from *p*(1) to *p*(3).

As a final experiment, we remove the restriction that the base node not be a prepositional complement, and instead use ICECUP IV and subtract cases to avoid double counting (see Wallis 2019). This increases the dataset by more than 50% again, but obtains a very similar set of results.

3.3 Common and proper nouns

What happens if we subdivide data into those with common and proper nouns head base units, but permit conjoined cases to appear in the path? Reprising Figure 7 (Section 2.3) obtains Figure 14 below. Initial additive probability rates tend to be higher, and common nouns now exhibit a consistent pattern of decline from *p*(1) to *p*(3).

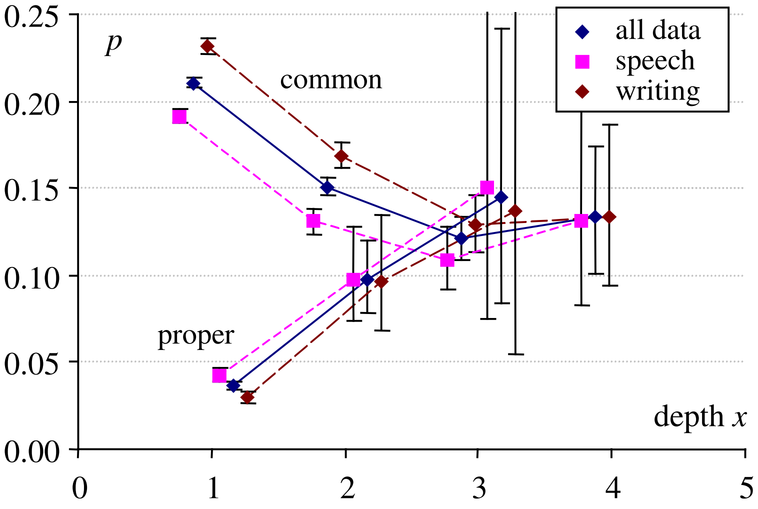


Figure 14. Figure 7 reprised, with conjoins permitted in the path.

In the case of repeating common or proper nouns in the path (see Section 2.7), allowing conjoins in the path does not materially change the outcome. The biggest difference is in the common noun sequences (which are nearly four times more frequent), where we see a sharp decline.

4. Conclusions

In (Wallis 2019), we identified evidence of interaction decisions impacting over two levels of embedding by drawing data from postmodifying clauses. However, data and results were limited. In this paper, we focus instead on postmodifying prepositional phrases. They are both more common and, for the same size of corpus, can be found exhibiting greater embedding depth.

For ease of reproduction and exploratory purposes, we prioritised a simpler system for identifying cases, which focused exclusively on instances without conjoined terms in the path. We employed FTFs using the publicly-available ICECUP 3.1.1 software and ICE-GB Release 2. We also required that the topmost NP node was not a prepositional complement, thereby ensuring that it began the embedded chain of instances being explored.

First, we considered patterns where embedded heads are unspecified. If the first head is also unspecified, then we see a 'rise and fall' effect, $p(1) < p(2) > p(3)$. This was unexpected, and not comparable to the equivalent pattern for postmodifying clauses, which fall in frequency. Whereas grammarians conventionally analyse cases of double-PP embedding as consisting of two structurally independent units, our data suggest that they are not always semantically or cognitively independent.

When we subsequently allowed conjoined noun phrases and prepositional phrases into the embedded path, this had the effect of increasing the number of cases classed as postmodified at level 1, 2 and so on. Initially at least, this research design change increased the identified rate of embedding, causing $p(0)$ and $p(1)$ to increase, and replacing a 'rise and fall' effect with a secular decline.

If we restrict only the first head type by wordclass (common noun, proper noun, pronoun, numeral, nominal adjective and proform), we are able to draw two general conclusions.

Firstly, second order additive probabilities, $p(2)$, tend to *converge* on a mean ($\bar{p}(2) = 0.1428 \in (0.1376, 0.1483)$). This is what we would expect if the decisions were generally independent. Secondly, *significant differences between them* may also be detected. This second order rate represents the chance of adding the same structure in different contexts. Although in many cases decisions are independent, we have evidence of influence across two levels of embedding.

This means that the wordclass (and semantics) of this initial head is having an impact on decisions two-deep. We can envisage two ways that such an influence may operate:

1. **directly** as a single decision 'chunk' (level 0 restricts decisions at both levels 1 and 2, or decisions are made together) or
2. **indirectly** as a consequential impact (level 0 affects decisions at level 1, which in turn influences those at level 2).

Both explanations could be simultaneously true. However, the implication of a grammatical model is that it reflects independent decisions at each level, which favours either no interaction or an indirect one at most. Recall that we said our default assumption was that decisions in an embedded series were made independently in turn. This is our null hypothesis. The alternate 'direct' hypothesis is that decisions are in fact made together.

The set of pronouns exhibits a very diverse rate of initial PP postmodification. The most common pronoun subtype, personal pronouns, are rarely postmodified at all, whereas some subtypes, such as quantifying pronouns, represent a small proportion of cases and are frequently postmodified. If we are to meaningfully refer to 'pronouns' for this purpose, we should recognise this wide variation, and only pool subtypes equally capable of postmodification.

Examining cases of second-level embedding drawn from subtypes of pronouns and proper nouns restricted by type demonstrate a high level of schematic or idiomatic constructions, such as titles. These seem to favour the 'direct' explanation above.

Doubly-embedded pronouns and numerals also show convergence to $\bar{p}(2)$, however, as there are of a lower frequency, these rates are generally not significantly different from each other. Cases restricted to the same subtype are simply too rare to draw conclusions.

Finally, for the most populous categories of nouns, restricting subordinate heads to be of the same type explains approximately 70-80% of the common noun data. This is in line with overall proportions in the corpus (80% of nouns are common, 20% proper). However, with only 20% of nouns overall, the 55-65% share of proper noun data is unexpectedly high, i.e. there is a strong tendency for proper nouns to be postmodified by PPs with proper noun heads. The widespread use of PP postmodification in titles appears to be the main cause, an instance of 'direct' formulaic interaction in double-level embedding.

Wallis (2019) showed that one can evaluate the additive probability over multiple sequential and embedded construction steps. We applied a simplified version of this method to a wide range of conditions on postmodifying preposition(al) phrases, yielding possibly surprising results. We verified these results by relaxing this restriction on coordination using ICECUP IV. These did not obtain substantially distinct results: rather, they strengthened our conclusions.

We also see evidence of a secular decline with common nouns, which make up the vast majority of cases. This is consistent with trends for postmodifying clauses, and indicates a potential mental processing cost or a communicative constraint. However, we also see a very distinctive *increasing* rate for proper nouns and other categories, which on closer inspection appear to be where ostensibly independent decisions are actually routinized into idioms and schema.

References

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

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

Wallis, S.A. 2019. Investigating the additive probability of repeated language production decisions. *International Journal of Corpus Linguistics 24*:4, 490-521.

Wallis, S.A. 2021. *Statistics in Corpus Linguistics Research: A new approach.* New York: Routledge.

Wallis, S.A. 2022. *Directional evidence revisited: End weight bias and templating in conjoined phrase postmodification.* London: Survey of English Usage.

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