Grammatical mark-up: Some more demarcation disputes[1]

David Denison

The University of Manchester

Abstract

A selective tour of annotation in historical corpora begins with extra-linguistic mark-up: how far can it alert the corpus user to usage which is atypical of the variety being sampled? Several syntactic fossils are discussed, and a playful use of foreign and pseudo-foreign words. Are they a kind of code-switching? In the former case the answer No is given, in the latter a partial Yes.

As for grammatical mark-up, with few exceptions a given scheme must privilege one particular analysis for each word, sentence or other unit of analysis. Special tags are available in the CLAWS and Penn Treebank tagsets for cases which remain ambiguous but which are in principle decidable. Grammatical mark-up remains essentially a matter of synchronic analysis, and the guiding principle is to be as specific as possible; tagsets routinely deploy a much finer set of distinctions than traditional word classes. Historical corpora like the Penn family aim also for consistency of analysis. I argue that both principles can be problematic.

Consider first the push towards a unique POS tag for every word. I propose that certain kinds of word are vague as to their word class not because of a failure of analysis but because they are genuinely underdetermined. Vagueness is not ambiguity, so ambiguity tags would be inappropriate – at least with their currently intended values. Secondly, the desideratum of consistency does not allow for patterns which arguably have dual analyses synchronically, nor for items which are in transition or which have changed over the time-span of a historical corpus. Among the data discussed are the POS-tagging and parsing of adjectives derived from passive participles (*interested, amused*), multi-word prepositions (*on behalf of*), phrasal and prepositional verbs (*run over*), proper-to-common-noun conversions and noun-to-adjective transitions (*BandAid*), countable-to-mass conversions (*He looked at me across a vast expanse of table*) and the converse (*two coffees*). A brief conclusion argues that while some of the problems considered are statistically unimportant, others demand greater flexibility of mark-up.

[1] I am grateful to Hans Martin Lehmann, Gerold Schneider and Nick Smith for help in the preparation of the oral version of this paper, and to Marianne Hundt for commenting on a written draft as well. The usual disclaimers apply.

1 Introduction

In a previous paper (Denison 2007) I offered five detailed case studies of morphosyntactic tagging in English corpora. The focus there was on areas of English lexis and grammar which posed problems for tagging because items fell near category boundaries. Here I will briefly take up similar issues with different data and corpora and extend the discussion to metadata or extra-linguistic annotation. I look at kinds of variation which are not generally well served by corpus mark-up and ask how – or whether – the annotation could be made more helpful. Most of the problems identified are consequences of language change, but even corpora specifically designed for diachronic research are not immune.

2 One variety at a time?

Texts in corpora are generally labelled by date, genre, and so on, and information may be given on dialect, speaker, etc. Nevertheless, this hides much variation within a text. For example, speakers may tell jokes in an accent other than their own, novelists may attempt to recreate a period, sometimes earlier (or later!) than the present, and so on. The extra-linguistic mark-up cannot follow all such twists and turns. Most will admittedly be of minor importance for studies looking for statistical effects across a large corpus, but non-native analysts may be misled in the discussion of individual examples.

2.1 Date

Language changes over time, but not homogeneously. Corpus texts, just like everyday speech, can be littered with novel usages which go beyond the norms of their time, and equally they may harbour usages which are – strictly speaking – no longer current. Two related concerns, then, are the effect on our understanding of linguistic history, and how far linguistic mark-up can or even should reflect such chronological layering. Consider this simple sentence from the British National Corpus (BNC):

(1) How goes it, Bruce? (AB9 7)

This apparent example of V2 syntax appears in a text dated '1985-93' in BNC. It is clearly a fossil – a self-conscious archaism or perhaps foreignism, now established as a kind of salutation. The usage may well be supported by another *How V ...* inversion type:

(2) How come you're homeless anyway? (A0F 1551)

Nevertheless the V2 pattern for interrogatives is one which has generally disappeared for most lexical verbs since the seventeenth century, and both (1) and (2) are idioms which are too idiosyncratic to tell us much about the productive syntax of Present-day English (PDE).

Fossil syntax is surprisingly common. An apparent example of the so-called sentence brace is seen in:

(3) A chemical does not a product make (PV 564)

Example (3) is a creative variation on a fossil, a familiar proverb, (4), in turn a fairly common variant of the more normal (5) (ignoring spelling variations), which was translated from Aristotle into English by the 16th century; see here Speake (2003):

(4) One swallow does not a summer make.

(5) One swallow doth not make/does not make/maketh not (a) summer.

Starting, then, from some form of the proverb like (5), the variant (4) is probably a misquoted poetic archaism[2] of long standing, and example (3) is what is now styled a "snowclone" (Pullum 2004) – that is, an adaptation of a voguish phrase (whether archaic or not) by the substitution of different lexical items in a fixed template. It is far from obvious how to mark snowclones linguistically in a corpus, as it is the template that is in effect a prefab rather than any one idiomatic string.

The point here is that (3) is somewhat inconvenient. The sentence brace was current in prose until the early Middle English period, still fairly common in later Middle English but in steep decline in prose by the 16th century (van der Wurff & Foster 1997). Corpus users surely expect to find a clear marking of date for the examples in a corpus, but the existence of such diachronic layers within a synchronic grammar adds an undesirable complication which is not easily conveyed in metadata.

2.2 Code-switching

Switching from the base language into a foreign language is routinely marked in many corpora, for example the Helsinki Corpus of English Texts:

> Words and phrases in languages other than English was annotated by surrounding it with the code (\...\) in the original version. In the TEI XML Edition, this code is replaced by the foreign element. (Marttila 2011: §3.2.4)

This is obviously helpful. If at some point the language stops being English, users need to know – whether in order to discount the foreign word(s) or to study the process of code-switching. However, it is not always straightforward to add such annotation. This example comes from a small corpus I directed:

(6) I think if I can work that incident up a little it will form a very fitting dénoument to my unhappy "Mme de V." wh: <foreign>(en passant)</foreign> I may mention is likely to be fair copied about the A.D. 1900. This must stand, <foreign>mon cher</foreign>, for the Sunday edition & entreats an answer. (1890 Ernest Dowson, from Corpus of lModE Prose [1994] mark-up altered to XML type)

Dowson playfully Frenchifies his English, and as corpus compilers we had to decide which of his lexical choices, and indeed which of his sometimes fanciful spellings, to code as "foreign". How much mark-up is appropriate?

Arguably some fossils and the kinds of creative usage to be discussed in section 5 below could be marked as code-switching too. Could switching out of 1980s English into what is apparently a different English be seen as the same in principle as switching into a foreign language? Probably not: unlike normal code-switching, comprehensibility for the wider speech community is maintained, not just for the immediate interlocutor. Anyway, given that language is **always** a mixture of rule-governed productivity, pre-fabs and creative extensions of rules, it is a reasonable abstraction to say that overall a corpus text "**is**" (an example of) the language of a certain date, genre, dialect – that is, that it can be taken to represent the range of possibilities of what is essentially one variety. (We

[2] EEBO records 'Yet the old prouerbe long agoe thus spake, |One swallow yet did neuer summer make' from William Painter, *Chaucer newly painted* (1623), while LION has 'One swallow (they say) no Sommer doth make. | Some swallow (I say) till great heat they take' from John Davies, *The scourge of folly* (1611).

should note too that the advent of World Englishes makes it even more impractical to treat creativity as code-switching.)

3 Underspecification

3.1 Vagueness vs. ambiguity

Grammatical tagging aims to assign word classes precisely; in fact tagsets routinely label forms even **more** specifically than the usual parts of speech; CLAWS C5 has 57 basic tags, for example, and C8 rather more. Ambiguity is the situation when the hearer/reader cannot be sure which of two or more readings was intended by the speaker/writer but does know that it must have been one or other, and the distinction affects the interpretation of the sentence. Now taggers are like reader/hearers in that they too have to figure out the correct interpretation and analysis of a sentence, and sometimes they cannot be sure. Some tagsets allow for this eventuality. BNC has 30 **ambiguity tags** (28 listed), including AJ0-NN1 and NN1-AJ0 (adjective or noun), AV0-AJ0 and AJ0-AV0 (adjective or adverb), but these are intended as stopgaps, for use 'when the probabilities assigned by the CLAWS automatic tagger to its first and second choice tags were considered too low for reliable disambiguation' (Leech & Smith 2000). The detailed discussion of **disambiguation** suggests that in principle, manual post-editing could replace an ambiguity tag with the correct single tag. Apparently similar in concept are the **multiple tags** in the Penn Treebank tagset (Marcus, Santorini & Marcinkiewicz 1993: 316).

In Denison (in prep.) I am proposing that some words and longer grammatical strings do not have a unique word class, not because of a failure of analysis but because they are genuinely underdetermined: they are syntactically **vague**. Examples include certain occurrences of

(7) *diverse, various, certain, several* (A ~ D)
(8) *(look) sad, (look) sadly, ...* (Adv ~ A)
(9) *near, worth, like, ...* (A ~ P)
(10) *fun, key, draft, genius, ...* (N ~ A)

In the appropriate contexts the word class of the above items is underdetermined between the two classes indicated in the brackets, so the analysis of the containing sentence is also vague. Whereas the producer of an ambiguous sentence must have intended one or other of the possible readings, a vague sentence is syntactically underdetermined for both producer and recipient. Vagueness and ambiguity are quite distinct.

It looks at first as if the Penn Treebank does recognise vagueness:

> We do not distinguish between verbal and adjectival uses of present and past participles, tagging both uses as VAG and VAN, respectively. (Santorini 2010)

But the fuller quotation implies that this is more likely to be avoidance of ambiguity resolution than a claim that two analyses are indistinguishable in principle:

> We have tried to plan our system so that at each stage of the annotation, information is added in a monotonic way. In particular, we want any future revisions of the bracketed structures always to add information, never to change it. This goal requires us to avoid subjective judgments since they are extremely error-prone. So, for example, we do not

distinguish adjectival from verbal passive participles, nor do we attempt to implement the argument-adjunct distinction.

Here are two analyses from the Penn Parsed Corpus of Modern British English (PPCMBE) with, respectively, a verbal and an adjectival use of *pleasing*, both marked with the POS tag "VAG":

(11) and devoted herself to **pleasing** and entertaining him (YONGE-1865,180.535)
[$_{PP}$ [$_{P}$ to] [$_{IP-PPL}$ [$_{VAG}$ [$_{VAG}$ pleasing] [$_{CONJ}$ and] [$_{VAG}$ entertaining]] [$_{NP-OB2}$ [$_{PRO}$ him]]]]
(12) with the most **pleasing** astonishment (GIBBON-1776,1,357.31)
[$_{PP}$ [$_{P}$ with] [$_{NP}$ [$_{D}$ the] [$_{ADJP}$ [$_{QS}$ most] [$_{VAG}$ pleasing]] [$_{N}$ astonishment]]]

The distinction is made in parsing at the phrasal level – IP-PPL vs. ADJP – rather than by tagging at the word level.

3.2 Vagueness of word class

I now turn to an example of word class vagueness. In BNC, *dinosaur(s)* is always tagged as a common noun, either NN1 (sg) or NN2 (pl) (except for the post-punk band *Dinosaur Jr* , which is correctly marked as NP0, a proper noun, when it appears!). The nominal tag NN1 seems perfectly reasonable even for an example like (13):

(13) Are they secretly debunking today's short-sighted rave fashions by reviving the **dinosaur** antics of Tangerine Dream and Focus? (BNC CK5 1043)

The syntactic slot occupied by *dinosaur* in (13), premodifier of a noun, is one which can be filled by a noun.

What then would the CLAWS tagger have made of the following example, had it occurred in the BNC?

(14) Richard represents views that myself and those who work in the business of football find **totally dinosaur**. (2011 Karren Brady, *London Evening Standard*)

Here we see a recent, perhaps nonce development of clear Adjective syntax for *dinosaur*. I argue that N > A changes of this kind come about through stepwise changes, not abrupt, involving "bridge examples" which are systematically vague in category and cannot be definitively assigned either to N or to A (Denison 2001, 2008, in prep.). The word *dinosaur* in the incipient new sense 'embarrassingly outdated' is a suitable candidate. Example (14) is not a bridge example: the N > A trajectory has reached a clear endpoint. If the wholly adjectival use of (14) spreads to more speakers, they would no longer have clear grounds for deciding whether *dinosaur* as premodifier in the BNC example, (13), was Adjective or Noun. When using attributive *dinosaur* 'embarrassingly outdated', such speaker/writers and their hearer/readers would not need to decide between the N and A classifications, as nothing at all hinges on the distinction. In short, the existing pattern (13) would become morphosyntactically **vague**, at least for speakers who have both N and A entries for *dinosaur* in their lexicon.

There are two important points being made here. One is that corpus mark-up does not recognise word class vagueness even in principle – and maybe it should. The other is that there may be unique analyses, previously uncontroversial, which ought to be revisited and retrospectively reclassified as vague when a new possibility enters the grammar.

4 Alternative analyses

Corpora with grammatical mark-up do not generally offer alternative analyses of the same sentence within a given annotation scheme.[3] The aim in principle is to find "the" correct analysis. Unique analyses may not always capture the whole truth about the syntax of a sentence, however. I discuss two such patterns and only briefly raise the question of whether alternative structural analyses can involve vagueness rather than ambiguity.

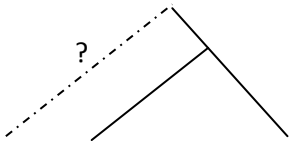
4.1 Prepositional verbs

Here once more are two examples from PPCMBE:

(15) I **sent for** one of these doctors (Reade 1863)
(16) But no Doctor was **sent for** till Monday (Nightingale 189x)

This time the tagging is not controversial, but the parsing is open to question. Most syntactic tests suggest that P is a constituent of PP. The PPCMBE stays with the PP analysis:

(17)



[NP-SBJ I] [VBD sent] [PP for [NP one of these doctors]]

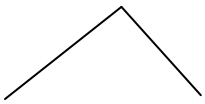
(18)

[NP-SBJ-1 no Doctor] [BED was] [VAN sent] [PP [P for] [NP *-1]]

The tags and parses shown are those of PPCMBE, with partial trees added to draw attention to the constituency of the preposition *for*. The 2nd edition of the International Corpus of English, Great Britain (ICE-GB2)analyses prepositional verbs in a similar way.

That is not the only possible analysis. The lexical unity of the V + P pair and the existence of a passive lead some scholars to suggest reanalysis (for example Mitchell 1958; Vestergaard 1977; Denison 1985; Quirk, Greenbaum, Leech & Svartvik 1985):

(19)



[NP-SBJ I] [PREP VB sent for] [NP one of these doctors]]

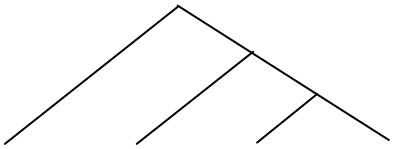
As far as I know, such an analysis (or reanalysis) of prepositional verbs has not been used in corpus parsing schemes.

4.2 Complex prepositions

In PPCMBE the phrase *on behalf of his country* (1888 Trollope) is parsed as follows:

[3] There is also a quite different (and irrelevant) situation, namely where a whole corpus has been processed more than once by different tagging programs. The ANC is supplied with three different stand-off tag schemes, while members of the English Department in Zurich can view certain corpora with a choice of tagsets and parses.

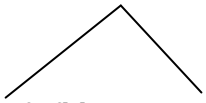
(20)



[$_{PP}$ [$_{P}$ on] [$_{NP}$ [$_{N}$ behalf] [$_{PP}$ [$_{P}$ of] [$_{NP}$ his country]]]]

That is, the preposition *on* is head of a PP, with an NP headed by *behalf* as complement. In contrast ICE-GB2 treats *on behalf of* as a complex preposition with the three words *on, behalf* and *of* "ditto-tagged" (because they function grammatically as a single unit):

(21)



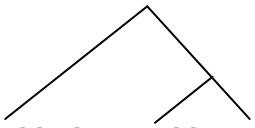
[$_{PP}$ [$_{PREP}$ on behalf of] [$_{NP}$ the people of Britain]]

This is a familiar dilemma, discussed by many scholars (especially Hoffmann 2005). Quirk et al. claim that eight out of nine indicators support a complex preposition analysis (1985: 670-3). Hardliners, on the other hand, find no **syntactic** grounds for recognising such strings as complex prepositions (e.g. Huddleston & Pullum 2002; Aarts 2007). Perhaps this is another context where a certain principle can usefully be appealed to (Denison 2010: 122):

(22) WYSIWYTCH
What You See Is What Your Theory Can Handle

Now the *on behalf* construction has a variant with a possessive before *behalf*. Even ICE-GB2 treats *behalf* as a noun in that case, with no ditto-tagging:

(23) engaged **on our behalf** in military action (S2B-030 099)



[$_{PP}$ [$_{P}$ on] [$_{NP}$ [$_{PRON}$ our] [$_{N}$ behalf]]]

Here there is no choice of analysis and no complex preposition.

Returning to PPCMBE, we find that its creators insist on consistency as a guiding principle:

> Although our treatment of fused forms generally reflects their phrasal origin, certain such items must be treated as unitary because of their syntactic distribution. For instance, UNDERHAND must be treated as an adjective because it can appear as a prenominal modifier. [...] Once an item is treated as unitary in one context, it is treated that way consistently. (Santorini 2010)

In their corpus *behalf* is always treated as a noun. Table 1 lists the relevant occurrences of *behalf*:

Pattern	N
in behalf of X	6
in the behalf of X	2
in X's behalf	6
in that behalf	16
on behalf of X	11
on the behalf of X	1
on X's (own) behalf	6
Total	48

Table 1: *behalf* in PPCMBE

Now as it happens, the 11 occurrences of the string *on behalf of* constitute less than a quarter of the 48 occurrences in the corpus. Whatever the motivation, Table 1 suggests that it may have been a good decision not to give *on behalf of* a multi-word analysis in this corpus but always to analyse *behalf* as a separate lexical item: not only is there a choice between *of-X* and *X's*, there is no single fixed form for the *of* pattern.

In BNC, there is even a rare plural (*behalfs* ×2, *behalves* ×3) as against 4014 singular *behalf*. However, the string *on behalf of* occurs 2708 times in BNC and vastly outnumbers *on the behalf of, in behalf of,* etc. The pattern *on X's behalf* (including *on my/our/his behalf*) occurs over 1100 times. Does this too argue against the complex preposition analysis? – we could simply be observing the usual choice between poss-s and poss-of constructions (as in *the book's cover* vs. *the cover of the book*), which would be the null hypothesis here. However, as I have argued elsewhere (Denison 2010: 118-22), the variation between poss-s and poss-of in the case of the *on behalf* string is not free variation, because common nouns prefer *of X*, while the examples with *X's* nearly all involve possessive determiners and proper nouns. The incipient complementary distribution is confirmed in the spoken part of BNC and in the Diachronic Corpus of Parsed Spoken English (DCPSE). The two alternative patterns (*on behalf of X* and *on X's behalf*) are increasingly dissociated from each other, and there is indeed increasing lexicalisation of the fixed string *on behalf of*.

What kind of mark-up should be used? The BNC is in my opinion particularly good here. Every occurrence of the string *on behalf of* is tagged in two different ways at different levels of XML mark-up: [4]

(24) a. [$_{PRP}$ on] [$_{NN1}$ behalf] [$_{PRF}$ of]
b. [$_{PRP}$ on behalf of]

That is, in (24)a we find three words tagged individually as PRP (preposition) + NN1 (singular common noun) + PRF (preposition *of*), whereas in (24)b the whole string is treated as a "multiword" (Leech & Smith 2000) and tagged as a preposition.

The Corpus of Historical American English (COHA) runs from 1800 or so to the present. It uses the same CLAWS tagger as the BNC but without the same post-processing, and the tagging of *on behalf of* that is displayed online is the multiword type of (24)b. The PPCMBE does not cover much of the twentieth century, stopping at 1914. As we have seen, it effectively tags *on behalf of* analogously to (24)a. In my view, a diachronic corpus covering lModE to the present day or the near future should not be required to apply the same tagging/parse to *on behalf of* throughout the

[4] I am grateful to Sebastian Hoffmann for clarification of this point (p.c. 1 May 2012).

period, *contra* Santorini's principle of consistency quoted above (2010), since the evidence in favour of a multiword analysis has been increasing over time.

Some equivocal syntactic patterns – typically the locus of change – merely involve under-determination of word class (and therefore also of phrasal projection). In other cases, however, I argue for dual analyses (cf. dual inheritance in a Construction Grammar framework). This cannot easily be accommodated in mark-up. The two synchronic situations correspond to diachronic changes that do not and do involve structural change, respectively.

5 Language change

One crude dichotomy in diachrony is between abrupt and gradual change. On the whole, grammatical mark-up copes better with abrupt change.

5.1 Abrupt change: count nouns and mass nouns

Here we have a different problem: a kind of rapid linguistic change involving an important morpho-syntactic distinction which is rarely traceable via linguistic mark-up. A count noun can be singular or plural and when singular cannot normally form a grammatical NP without a determiner. A mass (non-count) noun has no plural and can form an NP without an overt determiner. The syntax and semantics are significantly different. However, as is well known, there is productive conversion of certain mass nouns to count:

(25) Bring me two **coffees**. (BNC A73 2535)

The converse is also found. Here are some BNC examples of count nouns with mass noun syntax, following the hints in Matthews (1979: 29-31):

(26) It was real **mood-swing**. (C86 479)
(27) who did not give the impression of a mind of exceptional ability – there was not enough **knife** in the mind (A68 1139)
(28) He knew his son was all **mouth and trousers** (FBG 265)
(29) 'It's slit up each side,' she said showing an expanse of **thigh**. (ACK 604)

Given the possibility of nouns switching allegiance between count and mass subcategories, and given that many NP contexts do not serve to distinguish them at all, it is not surprising that corpus annotation schemes generally do not attempt to mark countability on nouns. Here is what is said about the BNC tagset:

> We make no special distinction between common nouns that can be mass (or 'non-count') nouns (eg *water, cheese*), and other common nouns. All are tagged NN1 when singular and NN2 when plural. (Leech & Smith 2000: §2)

(Nouns are marked NN0 if they are morphologically invariant for number, as with *sheep*.) Other tagsets are similar. The CLAWS tagger used for the BNC subcategorises nouns by verb concord, or the potential for it, which is easier to operationalise than a distinction based on NP syntax.

The change proper noun > common noun is an abrupt change, a kind of conversion. The BNC tags *Xerox (Corporation)* as NP0 = proper noun, *xerox* as NN1 = singular common noun. With *Band-aid*, *band-aid* 'wound dressing' it generally uses NN1, even though it is a proprietary name dating back to 1924. This is the familiar process whereby certain brand-names get turned into common

nouns. For *Band-Aid* or *BandAid* referring to charity fundraising concerts, it sometimes uses NP0, which is a curious chronological reversal.

5.2 Gradual change

Gradual change is often represented as

(30) X
X and Y
Y
⇓ time

The two co-existent states *X* and *Y* are generally thought of as different forms, but they may equally be underlying analyses, identical in surface form. Mark-up is often less sensitive to gradual change of this type (cf. complex prepositions, N > A, etc.), until the earlier pattern *X* has almost disappeared.

We have already looked at the development of a common noun usage for *band-aid*. A further development gives the word an adjective use. Examples (31) are internet data from WebCorp dated 2005-9:

(31) a. Keeping the heater core for "cooling" is a very **bandaid** approach to [...]
b. it's a very **bandaid** solution to a big problem.
c. OMG..that is so **bandaid**!

Unlike the nonce example of adjectival *dinosaur* in (14), *bandaid* is more firmly established in full adjectival use, as illustrated in (31)a,b at least. Mostly it is a noun, but in some examples it can only be an adjective. As suggested earlier, the simultaneous existence of two different word class analyses for such a word has consequences for the "bridge examples" where the two word classes are neutralised. They become systematically vague in category and should not be definitively assigned either to N or A.

How do corpora of present-day English and reference works deal with such matters? *OED* recognises *Band-Aid* as an adjective (though in fact all its examples are premodifiers that are vague between an N and an A analysis). BNC calls it a noun:

(32) the sort of **band-aid** solution (HHX 3069)
[$_{NN1}$ band-aid] [$_{NN1}$ solution]

So does COCA and the other BYU corpora, which use a crude form of the tagging applied to the BNC – though *Bandaid* is only tagged as an Adjective in one instance where in fact it has been converted to a **verbal** participle!

(33) All we're doing is **Band-Aiding** ourselves (1986 *Time Magazine*).

We encounter similar problems with phrasal verbs. Most tagsets have a special tag for the particle of a phrasal verb, e.g. "RP" in PPCMBE, "AVP" in BNC:

(34)and the Gib was **run up** (PPCMBE holmes-trial-1749) [*gib* = *jib* (sail)]
[$_{NP\text{-}SBJ}$ [$_{D}$ the] [$_{N}$ Gib]] [$_{BED}$ was] [$_{VAN}$ run] [$_{RP}$ up]

As Santorini explains, the VP is flat: 'The trees in the corpora are simply underspecified' (2010). Now tagging always distinguishes phrasal verbs from prepositional verbs, and parsing at least

distinguishes absence/presence of PP. However, diachronically they are not always distinct. Compare the treatment in PPCMBE of passive *run through* and *run over*:

(35)and all things being **run through** which I think necessary to be premised (PPCEME boethpr-e3-p2)
[$_{NP-SBJ-2}$ [$_{Q}$ all] things] [$_{BAG}$ being][$_{VAN}$ run] [$_{PP}$ [$_{P}$ through] [$_{NP}$ *-2]]
(36)he'd been nearly **run over** by a hackney coach (PPCMBE dickens-1837)
[$_{NP-SBJ}$ [$_{PRO}$ he]] [$_{HVD}$ 'd] [$_{BEN}$ been][$_{ADVP}$ [$_{ADV}$ nearly]] [$_{VAN}$ run] [$_{RP}$ over]

This is perhaps modern intuition: ***rún*** *through* (prepositional) vs. *run* ***ó****ver* (phrasal). ARCHER (tagged by Nick Smith with CLAWS and the Template Tagger) and DCPSE treat passive *run over* in essentially the same way, although their tagsets are different:

(37)Mr. Kenyon Parker, Q.C. [...] was **run over** by a Hansom cab yesterday afternoon in Chancery-lane, and seriously injured. (ARCHER 1866pal2.n6b)
[$_{VABDZ}$ was] [$_{VVN}$ run] [$_{RP}$ over] by a Hansom cab
(38)each of them looks as if they've been **run over** by a steam roller (DCPSE DI-B78 0048)
[$_{PRON}$ they] [$_{VP}$ [$_{AUX}$ 've [$_{AUX}$ been] [$_{V}$ run]] [$_{AVP}$ [$_{ADV}$ over]] [$_{PP}$ by a steam roller]]

In writing, both *run through* and *run over* are ambiguous syntactically. Historically, *run over* started off as a prepositional verb, as in the following example:

(39)I wish you had been poked into cells, and black holes, and **run over** by rats and spiders and beetles. (1865 Dickens, *Our Mutual Friend*, II.ii.268)

It was reanalysed as a phrasal verb. The syntactic reanalysis corresponds to a semantic change. Earlier, as a preposition, *over* referred to the trajectory of a vehicle or horse passing over a victim; later, as a particle, *over* came to be resultative, referring to the position of the victim.[5] Once again, therefore, it is not obvious that consistent tagging and parsing of the *run over* combination is desirable right across a diachronic corpus.

Here is another case, the participle. Past participles like *interested, amused, concerned* used to be verbal, as shown by the typical co-occurrence with intensifier *much*. Examples (40)-(42)from PPCMBE illustrate this:

(40)Once I sat between him and Miss Ellen Tree after dinner, and was much **amused** at their conversation and his stories (FAYRER-1900)
[$_{BED}$ was] [$_{NP-MSR}$ [$_{Q}$ much]] [$_{VAN}$ amused][$_{PP}$ at their conversation and his stories][6]
(41)He will be very much **interested** to hear of you. (YONGE-1865)
[$_{ADJP}$ [$_{QP}$ [$_{ADV}$ very] [$_{Q}$ much]] [$_{VAN}$ interested] [$_{IP-INF-SPE}$ to hear of you]]
(42)Woke early , much **vexed** at having to go away again. (BENSON-190X)
[$_{IP-PPL}$ [$_{NP-MSR}$ [$_{Q}$ much]] [$_{VAN}$ vexed] [$_{PP}$ [$_{P}$ at] …]][7]

More recently they have come to be adjectival, modified by *very*.

[5] Note that with example (38) there is a mismatch between the older semantics and the PDE syntax, since the point of the comment, about figures in certain artists' paintings, is not that they look prone and injured but that they look **flattened**, as if a steamroller has passed over them!
[6] In (40) NP-MSR = measure noun phrase, VAN = passive participle (verbal or adjectival).
[7] In (42) and (44) IP-PPL = participial clause, but ?not complement of V.

As for *Ving*, it can be a clear adjective – and be so tagged in corpora. Consider example (43), from ARCHER, which some users have tagged with several different programs. The first two taggings mark it with the code for adjective, but the third does not.

(43)It pays, though it may seem **boring**. (1961evan.j8b)
it [$_{VM}$ may] [$_{VVI}$ seem] [$_{JJ}$ boring] – CLAWS (Nick Smith)
it [$_{MD}$ may] [$_{VB}$ seem] [$_{JJ}$ boring] – ZH TREETAG (also *willing, unwilling, uninteresting, surprising*)
it [$_{Vmod}$ may] [$_{inf}$ seem] [$_{ING}$ boring] – ZH ENGCG2 (also *surprising, willing*, whereas *unwilling, uninteresting* are tagged as adjectives)

If we bring historical knowledge to the question we find that certain *Ving* forms were once more verbal than they are now, occurring where now only adjectives can (allegedly) appear:

(44)we began to Clamber up those Hills , which **seem hanging** over the Road of Gombroon (FRYER-E3-H,II)
which [$_{VBP}$ seem] [$_{IP\_PPL}$ [$_{VAG}$ hanging] [$_{PP}$ over the Road of Gombroon]]
(45)The long crisis in Laos **appeared nearing** a showdown today. (Brown A21)
The long crisis in Laos [$_{VBD}$ appeared] [$_{VBG}$ nearing] a showdown today. (TREETAG annotation)
The long crisis in Laos [$_{Vpast}$ appeared] [$_{ING}$ nearing] a showdown today. (ENGCG2 annotation)
(46)Large and agonizing drops **seemed forcing** their way to his [eyes] (1799lee-.f4a)
(47)the shrill shrieks of owls, the loud cries of the wolf, and mournful screams of panthers, which were redoubled by distant echoes as the terrible sounds **seemed dying** away (1797blee.f4a)
(48)I have tried to remember its teachings, but of late they **seemed slipping** from my mind. (1876roe-.f6a)

What does all this tell us? Participles – both present and past – show many changes over the last 300-400 years, both in word class and distribution. Attempts to be consistent in tagging mask such changes, and uncorrected tagging can produce bizarre results.

6 Does it matter?

Two answers can be given:

Arguably, No. Some of the problems discussed are fairly peripheral. Mark-up is an aid, not an end in itself, and mark-up that is "good enough" – allowing the user to find patterns most of the time with adequate precision and recall – is a reasonable aim.

Arguably, Yes. What's convenient for the POS tagger is not necessarily convenient for the user. I take the position that it **does** matter. The God's Truth fallacy, whereby a corpus 'may easily create the erroneous impression that it gives an accurate reflection of the entire reality of the language it is intended to represent' (Rissanen 1989: 17), applies to grammatical mark-up too: misclassified examples will mislead students. Experienced researchers can find misclassified examples if they already have suspicions, but if not, relevant examples may be missed.

For a word of vague (that is, underdetermined) class, I would prefer tagsets to include tags that explicitly signal indeterminacy between two categories; they could be something like an ambiguity tag in form. In other cases I wish tagging could make distinctions that are deliberately avoided in corpora with which I am familiar. Stand-off tagging allows different mark-up schemes for the same material, as with Zurich Corpus Navigator 2.0 (Hans Martin Lehmann) or American National Corpus 2, but these are essentially different tagsets and taggers and not simultaneously available. Software

which offers “layers” of user mark-up (cf. Julia Richling’s and Anke Lüdeling’s papers at the New Methods conference) might allow alternative mark-up to be exploited more easily. The way that BNC can offer alternative taggings of multiword lexical items is pleasing (section 4.2 above), but it is not clear how that would translate to parsing, and in any case it would break down when faced with multiply overlapping prefabs like *those sort of, those sort, what sort, some sort of, sort of thing, that sort of thing*, etc. (Denison 2007: §2.4).

The balance between too much and too little in corpus annotation is always a delicate one. My brief survey of metalinguistic and grammatical mark-up suggests to me that it is the latter where it would be particularly worth aiming for something more – and indeed something different.

References

Corpora and databases

ARCHER = A Representative Corpus of Historical English Registers
BNC = British National Corpus
COHA = Corpus of Historical American English
DCPSE = Diachronic Corpus of Present-Day Spoken English
EEBO = Early English Books Online (26 Mar. 2012)
ICE-GB2 = International Corpus of English, Great Britain
LION = Literature Online (ProQuest) <http:lion.chadwyck.co.uk> (26 Mar. 2012)
lModE Prose = A Corpus of late Modern English Prose
OED = Oxford English Dictionary
PPCEME = Penn Parsed Corpus of Early Modern English
PPCMBE = Penn Parsed Corpus of Modern British English

Secondary references

Aarts, Bas. 2007. *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford: Oxford University Press.

Denison, David. 1985. Why Old English had no prepositional passive. *English Studies* 66: 189-204.

Denison, David. 2001. Gradience and linguistic change. *Historical linguistics 1999: Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999*, Laurel J. Brinton (ed.), 119-44. Amsterdam and Philadelphia PA: John Benjamins. [Current Issues in Linguistic Theory 215].

Denison, David. 2007. Playing tag with category boundaries. *VARIENG e-Series* 1, *Annotating variation and change (Proceedings of ICAME 27 Annotation Workshop)* ed. Anneli Meurman-Solin & Arja Nurmi. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG). http://www.helsinki.fi/varieng/journal/volumes/01/denison/

Denison, David. 2008. Category change in English with and without structural change. Paper presented at NRG4, University of Leuven.

Denison, David. 2010. Category change in English with and without structural change. *Gradience, gradualness and grammaticalization*, Elizabeth Closs Traugott & Graeme Trousdale (eds.), 105-28. Amsterdam and Philadelphia: John Benjamins. [Typological Studies in Language 90].

Denison, David. in prep. *English word classes: Categories and their limits*. Cambridge University Press. [Cambridge Studies in Linguistics].

Hoffmann, Sebastian. 2005. *Grammaticalization and English complex prepositions: A corpus-based study*. London and New York: Routledge. [Routledge Advances in Corpus Linguistics 7].

Huddleston, Rodney & Geoffrey K. Pullum *et al*. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Leech, Geoffrey & Nicholas Smith. 2000. The British National Corpus (Version 2) with Improved Word-class Tagging (BNC2 POS-tagging Manual). http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19: 313-30.

Marttila, Ville. 2011. *Manual to the Helsinki Corpus TEI XML Edition*.

Matthews, P. H. 1979. *Generative grammar and linguistic competence*. London: George Allen & Unwin.

Mitchell, T. F. 1958. Syntagmatic relations in linguistic analysis. *Transactions of the Philological Society*: 103-6.

Pullum, Geoffrey K. 2004. Snowclones: Lexicographical dating to the second. http://itre.cis.upenn.edu/~myl/languagelog/archives/000350.html

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16-9.

Santorini, Beatrice. 2010. Annotation manual for the Penn Historical Corpora and the PCEEC. http://www.ling.upenn.edu/hist-corpora/annotation/index.html

Speake, Jennifer (ed.) 2003. *The Oxford Dictionary of Proverbs*. Oxford University Press.

Vestergaard, T. 1977. *Prepositional phrases and prepositional verbs: A study in grammatical function*. The Hague: Mouton. [Janua Linguarum series minor 161].

van der Wurff, Wim & Tony Foster. 1997. Object-verb order in 16th century English: A study of its frequency and status. *Language history and linguistic modelling: A Festschrift for Jacek Fisiak on his 60th birthday*, vol. 1, Raymond Hickey & Stanislaw Puppel (eds.), 439-53. Berlin and New York: Mouton de Gruyter. [Trends in Linguistics. Studies and Monographs 101].