Where Should Annotation Stop?

Geoffrey Sampson
University of Sussex

ABSTRACT

The paper asks how much structural detail it is reasonable to include in a detailed general-purpose grammar annotation scheme. I argue that there is no principled answer to that question; even grammatical distinctions which in general are clear and linguistically central will often be "distinctions without a difference" in particular examples. The discipline which offers the closest intellectual precedent for linguistic treebank-compilation activity, biological systematics, is disanalogous from our work in that respect.*

Detailed v. skeleton analytic schemes

Any scheme for structural annotation of corpora must embody decisions about how much detail to include.

Some groups explicitly aim at "skeleton parsing", marking much less grammatical detail than linguists recognize a language as containing. In many circumstances, this will be a sensible strategy. If one's chief goal is to have as large as possible a quantity of analysed material, from which reliable statistics can be derived, then skeleton parsing is more or less unavoidable. Automatic parsers may be able to deliver skeleton but not detailed analyses, and human analysts can produce skeleton analyses quickly. Furthermore, for some natural language processing applications skeleton analysis may be all that is needed.

But attention also need to be given to detailed structural analysis. All the grammar in a language, surely, serves some function or another for users of the language – it is not just meaningless ornamentation. There are many diverse potential applications for automatic NLP, some of which have scarcely begun to be developed, and it would be rash to assume that this or that aspect of language structure can safely be ignored because it will never be relevant for any practical NLP application. If some minor details of structure might be significant for research in the future, then the sooner we begin devising standardized, explicit ways of registering them in our treebanks (structurally analysed corpora) the better, because the business of evolving usable, consistent schemes of structural classification and annotation is itself a challenging and time-consuming activity.

To draw an analogy from the biological domain, much of the range of very lively research developments currently taking place in genetics and cladistics depends on the fact that biologists have a detailed, internationally-recognized system for identifying living species, the foundations of which were laid down as long ago as the eighteenth century. Linnaeus and his successors could not have guessed at the kinds of research revolving round DNA sequences which are happening in biology nowadays, but modern biology would be hampered if their species-identification scheme were not available.

Since the 1980s, my team has been developing a structural annotation scheme for English (a first draft of which was published as Sampson (1995))

* This research was supported by the Economic and Social Research Council (UK).

which aims at rigorous explicitness and maximum completeness of detail. We have also been compiling and circulating treebanks which apply the scheme to language samples, but the level of detail of the analytic scheme means that the treebanks illustrating it are small compared to some of those nowadays available – we accept this as a necessary cost of our strategy. To quote the documentation file of our SUSANNE Corpus (`ftp://ota.ox.ac.uk/pub/ota/public/susanne/`):

> The SUSANNE scheme attempts to provide a method of representing all aspects of English grammar which are sufficiently definite to be susceptible of formal annotation, with the categories and boundaries between categories specified in sufficient detail that, ideally, two analysts independently annotating the same text and referring to the same scheme must produce the same structural analysis.

Comprehensiveness and rigour of analytic guidelines are ideals which can never be perfectly attained, but there is some evidence that the SUSANNE scheme is recognized as having made a useful advance; for instance, Terence Langendoen (President of the Linguistic Society of America) commented in a review that its "detail … is unrivalled" (Langendoen 1997: 600).

If one's aim is a comprehensive detailed rather than skeleton analytic scheme, then a question which arises and which does not seem to have been much discussed to date is where to stop. How does one decide that one has exhausted the range of grammatical features which are "sufficiently definite to be susceptible of formal annotation"?

The trainability criterion

In practice, one factor that may impose limits on detail is what it is practical to teach annotators to mark reliably. Even if an annotation scheme is limited to standard, traditional grammatical categories, it is hard to overestimate the difficulty of training research assistants to apply it to real-life language samples in a consistent manner. Some annotation projects are explicit about ways in which training considerations shaped their notation scheme. Meteer et al. (1995), defining the dysfluency annotation scheme of the Switchboard Corpus, make remarks such as "annotators were basically unable to distinguish the discourse marker from the conjunctive use of *so*", "*actually* also proved impossible for the annotators to mark consistently and was jettisoned as a discourse marker part of the way through".

But, although what one can and cannot train annotators to do is obviously an important consideration in practice, it is hard to accept it as a principled boundary to detail of annotation. Sometimes, annotators' failure to apply a distinction consistently may be telling us that the distinction is unreal or inherently vague. But there are certainly other cases where the distinction is real enough, and annotators are just not good at learning it (or a principal investigator is not good at teaching it). Usually, leaders of annotation projects are senior and more linguistically experienced than the annotators employed by the projects, so taking trainability as decisive would mean systematically ascribing more intellectual authority to the inexpert than to the expert.

Limits to expert decision-making

In principle, what junior annotators can learn to do is a secondary consideration, which is likely to depend on factors such as time available for training and individual educational background, as much as on the properties of the language itself. More scientifically interesting is the fact that sometimes it seems difficult or impossible to devise guidelines that enable even linguistic experts to classify real-life cases consistently.

If some grammatical distinction is hard for an expert to draw in a majority of cases, then probably we would all agree that that distinction is best left out of our annotation scheme. An example might be the distinction, among cases of the English pronoun *they*, between the original use referring to plural referents, and the newer use, encouraged recently in connexion with the "political correctness" movement, for a singular referent of unknown sex. This probably deserves to be called a grammatical distinction; note for instance that "singular *they*" forms a reflexive as *themself* rather

than *themselves*, as in the following British National Corpus examples:

> *Critics may claim inconsistency, but the person involved may justify themself by claiming total consistency*. FA9.01713

> *… the person who's trying not to drink so much and beats themself up when they slip back and get drunk!* CDK.02462

(These are not isolated oddities; traditionalists may be surprised to learn that 23 of the 4124 BNC texts each contain one or more tokens of the form *themself*, which seems quite a high number considering that singular *they* is unquestionably far less frequent than plural *they*.) But (although I have not checked this) it seems likely that in a high proportion of cases where *they* is in fact being used for "he or she", there will be few or no contextual cues to demonstrate that it is not used with plural reference – sometimes even for the speaker or writer it may be intended as nonspecific with respect to number as well as sex. So I would not want to add a distinction between singular and plural *they* to our annotation scheme, and I imagine few colleagues would advocate this for general-purpose linguistic annotation schemes. (If an annotation scheme is devised for some specialized purpose, there is obviously no saying what distinctions it may need to incorporate.)

More problematic are the many grammatical distinctions which can often be made easily, and which may seem to the linguistic expert (and perhaps to less expert annotators) rather basic to the structure of the language, but which in particular cases may be hard to draw. What proportion of instances of a distinction need to be indeterminate, before we regard the distinction as too artificial to include in our annotation scheme?

Structural ambiguities in spoken language

Much of the recent work of my team has dealt with spoken language (we have been compiling the CHRISTINE spoken British English treebank, `http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html`). Indeterminate structural distinctions are particularly noticeable in speech. Rahman & Sampson (2000) drew attention to a number of cases where distinctions that are fundamental with respect to written English turn out to be blurred in the spoken language. For instance, direct v. indirect quotation is conceptually or logically a very clear distinction, which has considerable human significance (relating for instance to different kinds of accuracy obligations on those who quote). In written English the distinction is made very sharp, not just through wording but through punctuation. Yet in spoken English direct v. indirect quotation is not a yes-or-no distinction at all, but at most a cline. The language has several features which mark material as direct quotation or as reported speech, but it is common for these features to be mixed, so that a quotation is more or less direct but not entirely one or the other. A BNC example discussed in Rahman & Sampson (2000) was:

> *well Billy, Billy says <u>well</u> <u>take</u> that and then <u>he'll</u> come back and then he er gone and pay that* KCJ.01053-5

– where, among the underlined items, the introductory *well*, the imperative *take*, and present-tense (*wi*)*ll* rather than (*woul*)*d* point towards direct quotation, but *he* (rather than *I*, referring to Billy) points towards indirect quotation. In spoken English, this kind of direct/indirect quotation ambiguity is so pervasive that it is tempting to see the distinction as an artificial, unrealistic one (so that, in terms of SUSANNE annotation symbols, no contrast would be maintained between `Fn`, for "nominal clause", and `Q`, for "quotation") – though the distinction is so important logically that we did not take this line in the CHRISTINE Corpus.

However, structural ambiguities in speech are not the most significant cases for present purposes. Applying any annotation scheme to spoken language inevitably leads to numerous unclarities caused by the nature of speech rather than the nature of the scheme. Analysts typically work from recordings with little knowledge of the situation in which a conversation occurred or the shared assumptions of the participants. Often, patches of wording are inaudible in the recording; speakers will mis-speak themselves, producing wording which they would not themselves regard

as good examples of their language; their "body language" will be invisible to the analyst; and if analysts work from transcriptions, even intonation cues to structure are unavailable. In these circumstances there will often be doubt about how to apply even a very limited, skeleton annotation scheme.

Limits to written-language analytic refinement

The real problem relates to unclarities in applying an annotation scheme to published written language, where the wording is as well disciplined as writer and editor can make it, and the only background assumptions shared by writer and reader are those common to members of their society and hence available to annotators too.

Let me illustrate via a range of examples drawn more or less at random from one written BNC text which I happened to be working with (in connexion with our new LUCY project, `http://www.cogs.susx.ac.uk/users/geoffs/RLucy.html`) at the time of writing this paper. (The sample is extracted from a novel about life in the French Foreign Legion. As English prose, I would judge it to be well-written.)

There are in the first place various passages which are genuinely grammatically ambiguous, e.g.:

> *I had set my sights on getting a good position in training so that I would be sent to the* 2ème Régiment Étranger de Parachutistes. EE5.00933

– is the *so that I …* sequence a constituent of the *sent* clause or the *getting* clause (was being sent to the *Deuxième Régiment* the motive for setting sights, or the potential result of getting a good position)?

> *They were kicked senseless and then handed over to the Military Police who locked them up in the roofless regimental prison before they were handed over to the Colonel of the Regiment for interrogation and questioning*. EE5.00912

– is the *before* clause part of the *locked* relative clause, or is it a constituent of the *then handed over* clause near the beginning (is the handover to the Colonel described as following the prison spell or as following the handover to the Military Police)? In both cases, the alternative interpretations would correspond to different annotation structures in the SUSANNE scheme, and surely in any other plausible linguistic annotation scheme.

Where a passage is genuinely ambiguous, we *expect* an expert to be unable to choose between alternative annotations – that is what "ambiguous" means in this context. Consequently, inability to choose in these cases is not a ground for suspecting that the SUSANNE scheme is over-refined. Notice, though, that even though many linguists would agree that the examples are genuinely ambiguous, these are not the kinds of ambiguity which might be resolved by asking the writer "what he really meant" – in the second case, for instance, the handover to the Colonel in fact followed *both* the handover to the Military Police *and* the prison spell, and there is no reason to suppose that the writer intended one interpretation to the exclusion of the other. This is a frequent situation in real-life usage.

In many other cases, the SUSANNE annotation scheme requires the analyst to choose between alternative notations which seem to correspond to no real linguistic difference (and where the choice is not settled by the rather full definitions of category boundaries that form part of the scheme), so that one might easily conclude that the notation is over-refined – except that the same notational contrast seems clearly desirable for other examples. Here are a handful of instances:

Passive v. BE + *predicative past participle*

The SUSANNE scheme (§4.335) distinguishes the passive construction, as in *I doubt … whether the word <u>can be limited</u> to this meaning …*, from cases where *BE* is followed by a past participle used predicatively, as in *… the powers … <u>were</u> far too <u>limited</u>*. What about the following, in a context where earth is being shovelled over a man:

> *When his entire body was covered apart from*

his head, … EE5.00955

I see no distinction at any level between a passive and a *BE* + predicative particle interpretation of *was covered* here; does that mean that it was a mistake to include the distinction even in connexion with "clear cases" such as those previously quoted?

Phrase headship

The SUSANNE system classifies phrases in a way that depends mainly on the category of their head words, which is commonly uncontroversial. In the example:

If we four were representative of our platoon, … EE5.00859

it is clear that *we four* is a phrase, subject of the clause, but I see no particular reason to choose between describing it as a noun phrase headed by *we*, or a numeral phrase headed by *four*.

Co-ordination reduction v. complete tagma

In the example:

He had wound up in Marseilles, sore and desperate, and signed on at Fort St Nicholas. EE5.00855

the first clause contains a pluperfect verb group *had wound*. It is normal for repeated elements optionally to be deleted from conjoined tagmas, so *signed* might be either the past participle of another pluperfect form from which *had* was deleted, or a past tense forming the whole of a simple past construction. This again seems in this context a distinction without a difference. Yet simple past v. pluperfect, and past tense v. past participle, are elementary English grammatical distinctions likely to be recognized by any plausible annotation scheme.

Interrogative v. non-interrogative how

A subordinate clause beginning with an interrogative is commonly either an indirect question (*I know why …*) or a relative clause (*the place where …*). But if the interrogative form is *how*, there is also a usage in which the clause functions like a nominal clause, with *how* more or less equivalent to *that*:

… shouting about the English and how they were always the first to desert … EE5.00902

It was frightening how hunger and lack of sleep could make you behave and think like a real bastard. EE5.00919

The shouting in the first example was presumably not about the manner of English legionnaires' early desertion but about the fact of it. The second example is more debatable; it might be about either the fact of hunger and no sleep affecting one's psychology, or about the insidious manner in which this occurs. This is an instance where the SUSANNE scheme avoids recognizing a distinction which is arguably real; the scheme does not allow *how* to be other than an interrogative or relative adverb, and therefore treats the *how* clauses as antecedentless relative clauses with *how* functioning as a Manner adjunct, even in the former example. But I could not give a principled reason for failing to recognize a distinction here, when other distinctions that are equally subject to vagueness are required by the annotation scheme.

Multi-word prenominal modifiers

Where a sequence of modifying words precedes a noun head, if the SUSANNE scheme shows no structural grouping then each word is taken as modifying the following word or word-sequence (Sampson 1995: §4.9). But a noun can be premodified by a multi-word tagma, in which case the modifier will be marked as a unit: cf. *He graduated with* [Np [Ns *first class*] *honours* [P *in oil technology*]] … GX6.00022 – *first class* is a noun phrase, the word *first* is obviously not intended as modifying a phrase *class honours*. However, consider the examples:

the nearby US Naval base at Subic Bay EE5.00852
… handed over to US Immigration officials …

The words *US Naval* could be seen as the adjectival form of *US Navy*, which is a standard proper name; and *US Immigration* is perhaps also current as a way of referring to the respective branch of the American public service. Yet at the same time, the base at Subic Bay is a naval base, and among naval bases it is a US one; and similarly for US immigration officials. If there are no grounds for choosing whether or not to group premodifying words in these cases, does that make it over-refined to recognize such a distinction in cases like *first class honours*?

(In fact the SUSANNE annotation scheme contains an overriding principle that only as much structure should be marked as is necessary to reflect the sense of a passage, and this principle could be invoked to decide against treating *US Naval*, *US Immigration* as units in the examples above. But ideally one would hope that an annotation scheme should give positive reasons for assigning a particular structure and no other to any particular example, rather than leaving the decision to be made in these negative terms.)

It would be easy to give many more examples of structural distinctions which are clear in some cases but seem empty in other cases. Perhaps the examples above are enough to illustrate the point.

I have no definite solution to the problem posed by cases like these. I do not believe that any neat, principled answer is available to the question of how refined a useful general-purpose structural annotation scheme should be; it seems to me that the devising of such schemes will always be something of a "black art", drawing on common-sense rules of thumb and instinct rather than on logical principles.

But, if that is true, it is as well that those of us involved with corpus annotation should be aware that it is so. People who work with computers tend often to be people who expect a logical answer to be available for every problem, if one can find it. For treebank researchers to put effort into trying to establish the "right" set of analytic categories for a language would lead to a lot of frustration and wasted resources, if questions like that have no right answer. The main purpose of the present paper is to urge any who doubt it that, unfortunately, there are no right answers in this area.

Annotation practice and linguistic theory

One group of academics might suggest that there are right answers: namely, theoretical linguists. For theoretical linguists it seems axiomatic that what they are doing in working out the grammatical structure of a language is not devising a useful, workable set of descriptive categories, but *discovering* structure which exists in the language whether linguists are aware of it or not. What makes the structure "correct" is either correspondence to hypothetical psychological mechanisms, or (for linguistic Platonists such as J.J. Katz, e.g. Katz 1981) the fact that languages are seen as mathematical objects with an existence independent of their users.

For some of the problem cases discussed above, it is plausible that linguistic theorizing might yield answers to classification questions which I described as unanswerable. It would not surprise me if some linguistic theory of headship gave a principled reason for choosing one of the words of the phrase *we four* as head. (It would also not be surprising if another linguist's theory gave the opposite answer.) For some other cases it is less easy to envisage how linguistic theory might resolve the issue.

But linguistic annotation ought not to be made dependent on linguistic theorizing, even in areas of structure where theoretical linguists have answers. That would put the cart before the horse. The task of linguistic annotation is to collect and register data which will form the raw materials for theoretical linguistics, as well as for applied natural language processing. If linguistic theory is to be answerable to objective evidence, we cannot wait for the theories to be finalized before deciding what categories to use in our data banks.

The most we can reasonably ask of an annotation scheme is that it should provide a set of categories and guidelines for applying them which annotators

can use consistently, so that similar instances are always registered in similar ways; and that the categories should not be blatantly at odds with the theoretical consensus, where there is a consensus. We cannot require that they should be the "correct" categories. To return to the biological analogy: studies of DNA sequences at the end of the 20th century are giving us new information about the theoretically correct shapes of the "family trees" of animal and plant kingdoms. It would have been unfortunate for the development of biology if Linnaeus and his colleagues had waited for this information to become available before compiling their taxonomic system.

A disanalogy with biology

I have alluded to the analogy with biological systematics; questions about how many and what grammatical categories treebankers should recognize have many parallels with questions about how many and what taxa should be recognized by biologists. Since our treebanking enterprise is rather a new thing, it is good to be aware of old-established parallels which may help to show us our way forward.

But although the classification problem is similar in the two disciplines, there is one large difference. We are worse placed than the biologists. For them, the lowest-level and most important classification unit, the species, is a natural class. The superstructure of higher-level taxa in Linnaeus's system was not natural; it was a matter of common-sense and convenience to decide how many higher-order levels (such as genus, phylum, and order) to recognize, and Linnaeus did not pretend that the hierarchy of higher-order groupings corresponded to any reality in Nature – he explicitly stated the contrary (cf. Stafleu 1971: 28, 115ff.). But for most biological purposes, the important thing was to be able to assign individual specimens unambiguously to particular species; the higher-order taxonomy was a practical convenience making this easier to achieve. And species are real things: a species is a group of individuals which interbreed with one another and are reproductively isolated from other individuals. There are complications (see e.g. Ayala 1995: 872-3, who notes that in some circumstances the objective criteria break down and biologists have to make species distinctions by "commonsense"); but to a close approximation the question whether individuals belong to the same or different species is one with a clear, objective answer.

In grammar, we have no level of classification which is as objective as that. So far as I can see, whether one takes gross distinctions such as clause v. phrase, or fine distinctions, say infinitival indirect question v. infinitival relative clause, we always have to depend on unsystematic common sense and *Sprachgefühl* to decide which categories to recognize and where to plot the boundaries between them.

It feels unsatisfying not to have a firmer foundation for our annotation activity. Yet anything which enables us to impose some kind of order and classification on our bodies of raw language data is surely far better than nothing.

References

Ayala, F.J. (1995) "Evolution, the theory of". *Encyclopaedia Britannica*, 15th ed., vol. 18, pp. 855-83.

Katz, J.J. (1981) *Language and Other Abstract Objects*. Rowman & Littlefield (Totowa, New Jersey).

Langendoen, D.T. (1997) Review of Sampson (1995). *Language* 73.600-3.

Meteer, Marie, et al. (1995) *Dysfluency Annotation Stylebook for the Switchboard Corpus*. `http://www.ldc.upenn.edu/myl/DFL-book.pdf`

Rahman, Anna & G.R. Sampson (2000) "Extending grammar annotation standards to spontaneous speech". In J.M. Kirk, ed., *Corpora Galore: Analyses and Techniques in Describing English*, pp. 295-311. Rodopi (Amsterdam).

Sampson, G.R. (1995) *English for the Computer: the SUSANNE Corpus and Analytic Scheme*. Clarendon Press (Oxford).

Stafleu, F.A. (1971) *Linnaeus and the Linnaeans*. A. Oosthoek's Uitgeversmaatschappij (Utrecht).