Natural Language Engineering **14** (4): 471–494. © 2008 Cambridge University Press
doi:10.1017/S1351324908004695 Printed in the United Kingdom

Definitional and human constraints on structural annotation of English[*]

GEOFFREY SAMPSON
Department of Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ, England
e-mail: grs2@sussex.ac.uk

ANNA BABARCZY
Department of Cognitive Science, Budapest University of Technology & Economics, 1111 Budapest, Stoczek utca 2, Hungary
e-mail: babarczy@cogsci.bme.hu

(*Received 20 June 2006; revised 22 March 2007; accepted 15 January 2008; first published online 05 June 2008*)

Abstract

The limits on predictability and refinement of English structural annotation are examined by comparing independent annotations, by experienced analysts using the same detailed published guidelines, of a common sample of written texts. Three conclusions emerge. First, while it is not easy to define watertight boundaries between the categories of a comprehensive structural annotation scheme, limits on inter-annotator agreement are in practice set more by the difficulty of conforming to a well-defined scheme than by the difficulty of making a scheme well defined. Secondly, although usage is often structurally ambiguous, commonly the alternative analyses are logical distinctions without a practical difference – which raises questions about the role of grammar in human linguistic behaviour. Finally, one specific area of annotation is strikingly more problematic than any other area examined, though this area (classifying the functions of clause-constituents) seems a particularly significant one for human language use. These findings should be of interest both to computational linguists and to students of language as an aspect of human cognition.

1 Introduction

In previous work (Babarczy, Carroll, and Sampson 2006) we began to explore the limits to the potential precision of English grammar annotation, through a series of experiments concerned with the special area of word classification. The present paper extends this exploration to parse-tree structure above the level of leaf nodes: we ask how much precision can be included in definitions of the phrase, clause, sentence, and paragraph structure of English as used in real life.

We are not concerned, here, with questions about how well some automatic parser or natural language analysis system performs. The performance of such a system is assessed by comparing its output to a 'gold standard' analysis of the same

[*] A version of this paper was delivered at the second Quantitative Investigations in Theoretical Linguistics conference, Osnabrück, June 2006. We are grateful to conference participants for discussion, and to Sampson's colleague John Carroll for advice.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

language sample(s), but gold standard analyses are not given in Nature. They must be produced by human analysts; this is recognized as difficult work, and there exist many different traditions of grammatical annotation (Bird and Liberman 2001). In consequence, since there is no single universally-agreed 'correct annotation' of any linguistic form, it is hard to get a feeling for how consistent and refined any usable set of annotation conventions can be. If natural language processing research makes heavy use of such conventions, as it does, then it must surely be worthwhile to investigate the ceiling on their reliability.

The fact that alternative traditions of annotation exist may conceal the nature of the problem from some readers. If we find linguists annotating the same examples differently, it is tempting to suppose that this is merely because they have divergent conventions in mind, and that kind of fuzziness or ill-definedness could be eliminated by settling on an agreed 'benchmark treebank' or the like. But that treebank must itself have been produced by analysts subject to the same issues about consistency and well-definedness of annotation practice as affect the work which the benchmark is cited in order to resolve.

It is important to appreciate, also, that there are in principle two questions here. One is how far it is possible to refine an explicit scheme of grammatical annotation, while keeping it well defined and meaningful. The other is how far it is possible for human experts to succeed in conforming their annotation behaviour to the rules of an explicit scheme. In Babarczy *et al.* (2006) we drew an analogy with measuring the size of clouds: one problem is that the inherent fuzziness of clouds limits the degree of precision with which their size can be defined, but another problem might be that humanity lacked the ability in practice to measure clouds even to that degree of precision. For wordtagging, we found that the capacity of human experts to conform to a refined classification scheme lags considerably behind the extent to which such a scheme can be well defined.

Although these questions have often been alluded to in the course of publications chiefly concerned with development of automatic parsers or treebanks, they have not often been addressed directly (for English, or for other languages). One recent relevant paper, concerned with annotation of structure in Chinese, is Xue, Xia, Chiou, and Palmer (2005). Relative to most annotation schemes for English, Xue et al.'s scheme is fairly simple (few node-label distinctions), which probably has to do with the near-perfectly isolating character of the Chinese language – such basic European grammatical contrasts as that between finite and nonfinite clauses have no counterparts in Chinese. Our experience of developing precise guidelines for annotating English has been very different from Xue et al.'s finding that 'very little revision' was needed between a first draft of guidelines and a final version less than two years later. We wish to examine the definitional and human limits to annotation precision with respect to the more intricate structural features which an adequate annotation scheme for English must recognize.

2 Experimental data

The present experiment involved two analysts who are thoroughly familiar with a scheme of annotation which seems to be the most refined and well defined

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

of those currently extant, each applying that scheme independently to a sample of English texts; the discrepancies between their respective analyses were investigated to discover how they were distributed across the various aspects of English grammatical structure, and how far they were caused by limitations of the analytic scheme, as opposed to human error.

The analysts were the present coauthors. The analytic scheme used was the SUSANNE scheme (Sampson 1995). Among various annotation schemes currently used by English corpus linguists, the SUSANNE scheme is distinctive in being explicitly motivated by the aim of maximizing the refinement, well-definedness, and comprehensiveness of the notation (rather than having been developed as an adjunct to the compilation of large treebanks). At the time when the data presented below were generated, the scheme had been refined over a period of about twenty years, with input (often intensive) at different stages from some fifteen or twenty researchers (including one of the world's most distinguished grammarians of English, Geoffrey Leech). Comments by neutral observers suggest that the SUSANNE aim has met with a degree of success: for instance, Dekang Lin (2003: 321) wrote 'Compared with other possible alternatives such as the Penn Treebank … [t]he SUSANNE corpus puts more emphasis on precision and consistency.' Others have made similar remarks. We used the SUSANNE scheme for the present experiment, not because one of the authors happened to be responsible for developing it, but because by common consent it appears to be the best available scheme for our purposes.

The experimental texts were ten diverse files (each of 2000+ words) from the LUCY treebank (www.grsampson.net/RLucy.html), which in turn were based on text extracted from files of the written-English section of the British National Corpus that are categorized in the LUCY treebank as 'polished' (i.e., roughly, edited published) writing, as opposed to informal or ephemeral documents. We shall refer to these files by their LUCY filenames; the following list shows these in the first column, the BNC source files in the second column, and brief indications of text genre in the third column. (Detailed information on the original documents is available in both the BNC and the LUCY files.)

B07	B07	book on librarianship
B09	A8H	city material from *The Guardian* daily newspaper
B18	C9F	cookery book
B22	CKU	art magazine
B27	EFG	women's magazine
B31	FBU	law reports
B39	J7U + J7V	two reviews of books on literary linguistics
B42	HCV	Irish bank staff association newsletter
C03	EE5	memoir of life in the French Foreign Legion[1]
C07	C07	novel set in India

[1] The LUCY finename for this text involves a miscategorization. Filenames beginning with C in the LUCY treebank are for fiction; although BNC file EE5 reads like a novel, the book it came from was published as an authentic memoir.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

The choice of polished writing, rather than informal writing or speech, was a consequence of the fact that we needed to use corpus material which one analyst had already annotated and the other was known not to have looked at (so that there could be no 'contamination' of one analyst's judgements by the other); but it would in any case have been the appropriate choice. Annotating informal writing or speech introduces many issues about how to handle writers' errors, unclarities of speech transcription, and so forth, which are additional to the issues about well-definedness of the notation scheme for English language structure, and obscure the latter. These additional issues are interesting and important in their own right, but for an initial exploration such as that reported here it was advisable to restrict the problem domain by focusing on polished writing.

Thus, although the original SUSANNE annotation scheme of Sampson (1995) has subsequently been extended to handle the writing of unskilled writers (www.grsampson.net/LucyDoc.html) and to handle speech phenomena more fully than the 1995 book permitted (www.grsampson.net/ChrisDoc.html), those extensions to the annotation scheme were not relevant to this experiment. The only additions and revisions to the 1995 scheme treated as operative for this experiment were the relatively small collection listed in §§14–15 of the CHRISTINE treebank documentation file.

Ideally, an experiment of this kind would involve more than two analysts, and a body of analysed material larger than 20,000-odd words. But this is a case where the ideal must bow to what is practical. We did not aim to study the process by which newcomers learn an annotation scheme, but to study how accurately it can be used by individuals who have already mastered it thoroughly; for a scheme of this level of complexity, that means that the analysts had to be people who had years of experience of working with it, and in the circumstances of university research there are never likely to be more than a tiny number of individuals meeting that requirement available at any time. Also, because of the intricate nature of the annotation scheme and the degree of detail in which it was necessary to scrutinize individual annotation discrepancies, even the quantity of material we used represented a serious research challenge; and it proved sufficient to yield a number of rather clear and interesting findings. (Arguably, computational linguistics has suffered from an unbalanced focus on high-level investigations of large-scale data-sets at the expense of close examination of individual cases.)

3 Text complexity

One might expect *a priori* that the ability of independent analysts to produce matching annotations would correlate with the complexity of the texts analysed. So in Table 1 we present measures of the complexity of the ten sample texts. 'Complexity' here is used in the schoolroom sense of the incidence of clause subordination, and is computed in the same way as in Sampson (2001). For each token[2] in a

[2] Computational linguists' choice of the term *token* for the segments of a text associated with leaf nodes of a parsetree seems regrettable, because it invites confusion with C.S. Peirce's useful distinction between *token* and *type*. But the computational-linguistic usage has become so ubiquitous that it is now unavoidable.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

Table 1. *Complexity of the sample texts.*

Text	Mean complexity	Max complexity
B07	1.50	5
B09	1.82	5
B18	1.16	4
B22	1.58	6
B27	1.22	4
B31	2.49	9
B39	1.63	5
B42	1.53	4
C03	1.74	6
C07	1.56	5

language-sample, we count the number of nodes in the lineage of the token (i.e. the path from the corresponding leaf node to the root node of the parsetree for the sample) which are labeled as main or subordinate (finite or nonfinite) clauses.[3]

The second column of Table 1 gives the complexity averaged over the leaf nodes of both analysts' versions of each text. The third column gives the maximum complexity count for any leaf node in the respective texts (the maximum counts were identical for the two analysts' output in each of the ten cases). It was surely predictable that by far the largest average and maximum complexity figures are for the legal text B31; and it seems plausible enough that the lowest average should be for the cookery book, B18.

4 Annotation scheme and tree similarity metric

In order to follow our quantitative findings on inter-annotator agreement, it is necessary to know something both of the general nature of the SUSANNE annotation scheme, and of the metric we used to assess the similarity of pairs of analyses of the same texts. We now consider these two issues in the reverse order.

The metric used to compare analyses is the *leaf-ancestor metric*, which is not specific to the SUSANNE scheme but applicable to any scheme which represents structure in the form of labeled trees. The leaf-ancestor metric was defined in Sampson (2000). In Sampson and Babarczy (2003) we gave experimental evidence to show that it is a successful operationalization of pre-theoretical ideas of parse accuracy, and in particular that it is considerably superior to its best-known alternative, the GEIG or PARSEVAL family of metrics (on which Kübler and Telljohann (2002) comment 'it is well known that these measures do not give an

[3] One complication is introduced in order to help the count to reflect consensus views of grammatical complexity. The SUSANNE treatment of co-ordination is non-standard (a co-ordination *X, Y, and Z* is assigned the structure [*X*, [*Y*], [*and Z*]], that is the second and any subsequent conjunct are treated as subordinate to the first conjunct); most linguists see the elements of a co-ordination as structurally on a level with one another, so in our complexity counts a clause node immediately dominating a subordinate-conjunct clause node is reckoned as adding one rather than two degrees of complexity.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

accurate picture of the quality of the parser's output'; cf. Manning and Schütze 1999: 434–7). We shall not repeat the detailed definition of the leaf-ancestor metric here, but in essence it assesses the parsing of a token by measuring the edit distance between the lineages for that token in a candidate and a gold-standard parsetree, or in the present case in two analysts' parsetrees, neither of which is assumed to be more authoritative than the other. We represent agreement of the respective parses as the complement of edit distance, so that 1 represents identically-analysed tokens and 0 represents no relationship between the respective analyses. Figures for the parsing of successive tokens are averaged to give an overall figure for a text.

Within the SUSANNE scheme, each tagma[4] is classified formally, and constituents of clauses are additionally classified functionally. The system of formtags recognizes 32 main classes of constituent, together with 64 subcategories, various combinations of which can be applied to various main categories. Additionally, certain types of multi-word sequence can be marked as grammatically equivalent to single words of particular classes. The functiontag system recognizes 23 roles for constituents of clauses. The scheme provides for 'ghost' or null tokens which identify the logical place and role of items that in surface grammar occur elsewhere; indices are used to show which ghost element is represented by which surface element.[5]

The purpose of the 499 pages of Sampson (1995) is to provide guidelines which so far as possible eliminate all ambiguities about how this system of notation symbols should be applied, so that for any form of words found in real-life English usage there should ideally be one and just one annotation that conforms to those guidelines. The scheme aims to represent all features and distinctions which are commonly recognized by grammarians, and it seeks to embody uncontroversial, consensus conceptions of language structure; but it makes no claims to be the 'correct' analysis (in terms of speakers' psychological models of their language or in other respects). It is explicitly a classification system imposed on the English language, rather than a scientific theory about the nature of the language; the principle that there should always be a single, predictable notation available for any form of words is given precedence over the aim that the notations of the scheme should mirror theoretical linguists' analyses.

Since most readers will be unfamiliar with the details of SUSANNE notation, in this paper we shall as far as possible discuss our experimental findings using the ordinary grammatical terms which are used in Sampson (1995) to gloss the SUSANNE symbols. But it will be important to grasp that the scheme explicitly eliminates the vagueness or ambiguity which inhere in many of the traditional terms.

[4] The term *tagma* is used for the token-sequence dominated by a parsetree node other than a leaf node. A *constituent* corresponds to any node other than a root node; a *tagma*, to any node other than a leaf node.

[5] In the SUSANNE scheme, indices are three-digit numbers, but there is no significance in the particular number chosen to mark any particular logical-structure/surface-structure linkage. We did not want to count different choices of index number as inter-annotator discrepancies, so before computing tree similarity we replaced each index number with the '^' symbol. In theory this could fail to penalize a case where both analysts identify a logical-structure ghost element but link it to different surface elements. That does not seem to us a likely kind of error, but if it did occur we will not have registered it.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

For instance, the SUSANNE functiontag :i is glossed as 'indirect object'. For some grammarians, an 'indirect object' is a noun phrase which, like a direct object, is marked by no preceding preposition; other grammarians would describe *Harry* as indirect object not only in *I gave Harry a cup of tea* but also in *I gave a cup of tea to Harry*. Again, some grammarians restrict the use of the term to 'second objects' in clauses which also contain a direct object; others call a sole object 'indirect', if the semantic relationship between it and the verb is characteristic of indirect rather than direct objects (as in e.g. *I paid Harry*). §§5.22–28 of Sampson (1995) are designed to remove these and other ambiguities from the use of the SUSANNE :i symbol, and other passages do likewise for each of the other elements of the notation. References beginning with '§' in what follows will refer to the numbered subsections of Sampson (1995).

It will be unavoidable to quote some examples of SUSANNE notation in what follows, so we illustrate its general features by showing the annotation for a simple artificial example chosen to include the main scheme features: a one-sentence paragraph *Mr Jones expected her to admit it*. That would be given the following labeled tree structure:

[O[S[Nns:s *Mr Jones*][Vd *expected*][Nos:O^ *her*][Ti:o[s^ GHOST]
[Vi *to admit*][Ni:o *it*]]].]

This says symbolically that the whole is a paragraph, O, consisting of a main clause, S, and the full stop. The main clause has four immediate constituents (ICs):

- *Mr Jones*, which is a proper noun phrase, Nn, in singular form, s, functioning as surface and logical subject, :s, of the main clause
- a verb group, V (consisting in this case just of one word, *expected*) marked for past tense, d
- *her*, an object-marked singular pronoun, Nos, which is surface but not logical direct object of its clause, :O, and which is linked by an index ^ to its logical counterpart elsewhere in the structure
- an infinitival clause, Ti, functioning as logical object, :o, of the main clause

and the infinitival clause in turn contains:

- a 'ghost' element functioning as logical subject, :s, and co-referential with *her* in the main clause, but realized by no verbal material of its own
- *to admit*, an infinitival verb group, Vi
- a noun phrase headed (and in this case consisting entirely of) the word *it*, Ni, functioning as surface and logical object of the subordinate clause, :o

In this example :s, :O, :o, and the s which labels the ghost node are functiontags; the other label-elements (other than the index, cf. note 5) are formtags.

5 Tokenization discrepancies

The leaf-ancestor approach to parsetree comparison assumes that both analysts' trees are identical with respect to their sequence of leaf nodes. This assumption can

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

be violated in two ways. The text may be tokenized differently; and one analyst may postulate a null 'ghost' where the other analyst has no ghost.

Under the SUSANNE scheme, presence of an inter-word space in writing is a sufficient but not a necessary condition for identifying a division between tokens associated with separate leaf-nodes in the parsetree. Punctuation marks are parsed separately from the adjacent words; many but not all hyphenated words are divided into separate tokens; genitive suffixes are split from the nouns to which they are attached; etc.

In principle, discrepancies between analysts' tokenization decisions could be handled in the same way as other parsetree discrepancies, by working with parsetrees in which the leaf nodes represented individual text characters and the tokens corresponded to the lowest-level bracketings of sequences of characters, labelled with wordtags. Since the character sequences are independent of individual analysts' parsing decisions, tokenization discrepancies would then become simply one kind of discrepancy assessed by the leaf-ancestor technique along with the others. However, we thought it inadvisable to adopt that solution. Discrepancies with respect to tokenization are relatively rare: most SUSANNE tokens coincide with orthographic words, and even when a token boundary does not coincide with a word space it is usually obvious (for instance there will always be a boundary between an alphabetic word and an immediately following comma). So making parsetrees extend down through the tokens to the separate characters of a text would artificially raise the similarity scores between tree-pairs. Furthermore, wordtagging is not the topic of the present investigation (we examined that aspect of annotation in Babarczy *et al.,* 2006). Here, we wanted to focus comparison on structure above the word level, where correct analyses are often not obvious.[6]

Consequently, where tokenization discrepancies did exist we handled them as follows. When one parsetree contains a token corresponding to two or more tokens in the other tree, the former tree is modified to make its leaf nodes match those of the latter, by dividing the single token to correspond to the multiple tokens of the other tree, and attaching the divided tokens directly to whichever node immediately dominated the single token before division. (Since we are not considering in this paper the issue of word classification, no questions about what wordtags would hypothetically be assigned to the divided tokens arise.) In order to quantify this category of inter-annotator discrepancy, we counted the number of places where one analyst but not the other identified a token boundary, as a proportion of all places where either or both analysts identified a boundary. The answer was 7.6 per thousand. However, more than half of all these places occurred in the single text B27, which in other respects also proved to be an outlier (as discussed in Section 6 below). For the nine texts excluding B27, the proportion of unmatched token boundaries to all token boundaries was 3.7 per thousand.

[6] It might seem that we could simply bypass the side-issue of tokenization discrepancies, by pre-processing the experimental texts to impose a particular tokenization and asking the annotators to treat that tokenization as authoritative. However, tokenization decisions sometimes interact with higher-level annotation decisions, so proceeding in this way would have pre-empted some of the annotation issues that we wished to study.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

Table 2. *Inter-annotator agreement figures.*

Text	Mean leaf-ancestor score	Leaf nodes	Leaves with identical lineages
B07	.950	2369	74.8%
B09	.953	2310	74.8%
B18	.944	2332	76.5%
B22	.954	2400	77.2%
B27	.901	2524	75.2%
B31	.934	2374	61.1%
B39	.938	2585	66.8%
B42	.956	2349	77.5%
C03	.970	2439	82.0%
C07	.955	2441	75.9%
mean	.945		74.2%

Unmatched ghost nodes were ignored in comparing successive pairs of leaf lineages. The number of such nodes was 1.6 per thousand as a proportion of all leaf nodes in both sets of analyses, and 10.3 per cent as a proportion of all ghost nodes. Note that ignoring unmatched ghost nodes in comparing lineages does not imply that figures for tree-similarity will fail to penalize these discrepancies. Any ghost node is linked by an index to a 'real' (non-null) node in another clause: when lineages containing that node are compared, some label in one lineage will contain an index symbol having no counterpart in the corresponding lineage.

6 Overall similarity results

Table 2 shows, in the second column, average leaf-ancestor scores for the ten texts. The third column gives the total numbers of leaf nodes (ignoring unmatched ghost nodes), and the fourth column gives the percentage of leaves whose lineages are identical in both analysts' trees. As one would expect, the most complex text has the lowest percentage of identical lineages: the longer the lineages are, the more elements there are in them which may disagree.

Among the leaf-ancestor scores, .901 for text B27 is a clear outlier. The fundamental reason for this seems to have to do with unclarities about how the lively typography of a long feature in a popular women's magazine is rendered into SGML in the BNC, an issue rather separate from the one we aim to investigate. If the leaf-ancestor score is averaged over nine texts, omitting B27, it rises from .945 to .950. In what follows, mean figures for our ten-text sample will routinely be followed in brackets by mean figures for nine texts excluding B27.

In themselves the similarity scores are not very informative, since we have no basis of comparison. It will be more instructive to look at how the discrepancies are divided between different sources of error. There are two questions here: (i) are there particular aspects of English grammar which lead to disproportionate numbers of discrepancies? and (ii) how is responsibility for the discrepancies shared between inadequacies of the annotation scheme, and human error?

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

7 Dividing overall discrepancy between annotation categories

We examined how various aspects of the annotation scheme contributed to the observed level of inter-annotator discrepancy by recomputing the agreement figures while relaxing in various ways the requirements for labels in the respective trees to count as matching, or while discounting failures to match by particular classes of label.

Thus, we can expect many discrepancies to arise from different choices of formtag by the analysts, the formtag system being by far the richest and most complex part of the complete annotation scheme. Let us consider initially just clause and phrase formtags, leaving aside what in the SUSANNE scheme are called 'rootrank' and 'wordrank' formtags (which will be discussed later). If agreement figures are recomputed with pairs of phrase or clause labels counted as matching whenever they agree with respect to any features other than formtags (that is, a pair of such labels are treated as matching if either both labels lack a functiontag or they share the same functiontag, and either both lack an index or both have an index, but one may be formtagged as, say, a plural noun phrase while the other is formtagged as a relative clause), then agreement figures for texts naturally rise. The overall mean agreement figure rises from 0.945 (0.950) to 0.958 (0.963).

For any one text, we compute the proportion of discrepancy attributable to aspect X of the annotation scheme as:

$$\frac{\text{agreement ignoring differences with respect to } X - \text{overall agreement}}{1 - \text{overall agreement}}$$

On this basis the mean proportion of discrepancy attributable to clause and phrase formtagging is 24.4% (24.3%).[7]

It seems likely *a priori* that analysts will be more consistent in deciding main form categories (whether a tagma is a noun phrase, an adverb phrase, a relative clause, a nominal clause, an infinitival clause, etc.) than in deciding subcategories (whether a verb group is perfective, whether a noun phrase is a proper name, etc.; whether a tagma of any main category is coordinated with, or in apposition to, what precedes; etc.). We tested this by recomputing agreement figures ignoring just subcategory information in clause and phrase formtags; in terms of SUSANNE tag structure, that means ignoring any characters other than (in the case of phrasetags) the first character, or (in the case of clausetags) the first character and an immediately-following lower-case character, if any. We find that on average 70.9% (70.7%) of the total clause/phrase formtag discrepancy figure is attributable to discrepancies in subcategories.

[7] These and subsequent figures for proportions of overall discrepancy in our sample attributable to particular aspects of annotation are calculated by applying the above formula to each text separately, then averaging over ten (nine) texts, so that the overall figure does not depend on the minor length-differences among the sample texts. (Consequently the percentages shown here cannot be checked by reference to the figures in the previous paragraph.)

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

Apart from formtags, the other elements that may occur in a tagma label are functiontags, and indices. If we relax the definition of label-matching so that labels which are identical except for having different functiontags match, we find a mean agreement figure of 0.959 (0.965); on average 27.7% (29.9%) of total discrepancy in a text is accountable for in this way.[8]

If indices are ignored, so that otherwise similar labels match despite one having and the other lacking an index, then the mean agreement figure is 0.946 (0.951); the mean proportion of total discrepancy accounted for in this way is 1.7% (1.8%).

For two further aspects of the annotation scheme, it seems likely *a priori* that discrepancies would relate less to different labels for corresponding tree-nodes, than to whether nodes bearing the relevant labels are postulated at all. These are the categories of formtag identified in the SUSANNE scheme (§4.39) as 'rootrank formtags' and 'wordrank formtags.' Rootrank formtags include `O` 'paragraph' and `Oh` 'heading,' one or the other of which labels each root node, together with labels for the categories 'title,' 'quotation,' 'interpolation,' 'tag question,' and 'technical reference.' Wordrank formtags cover 'grammatical idioms' such as *up to date*, treated by the scheme as orthographic word-sequences that function grammatically as single words, together with coordinations between single words.

Where a tagma is regarded as, say, an interpolation, it will be a constituent of some grammatical class and will be formtagged accordingly, but the node bearing that formtag will be dominated by a higher node labeled `I`, for 'interpolation.' Thus a typical inter-annotator discrepancy would be, not an `I` node in one tree corresponding to a node with another label in the other tree, but an `I` node in one tree having no counterpart node in the other tree. Likewise, what one analyst treats as an idiom the other analyst may regard as a sequence of words used in their normal senses, and hence not a tagma at all. Accordingly, the method of relaxing conditions on label-matching is not suitable for investigating discrepancies relating to these aspects of annotation. Instead, we computed agreement figures by relaxing the requirement that labels of the relevant class in one lineage need to match any label in the counterpart lineage. The score for a partial mapping from labels in one lineage into labels in the other lineage is computed as the total number of labels in the two lineages which are *either* mapped onto identical labels *or* belong to the class whose contribution to inter-annotator discrepancy we are examining (or both), divided by the total number of labels in both lineages (and the score for a lineage-pair is then as usual the highest possible score for any partial mapping between the two sets of labels).

Computed in this way, the mean agreement figure ignoring rootrank discrepancies is 0.948 (0.952), and the mean proportion of overall discrepancy attributable to rootrank discrepancies is 3.6% (2.0%). The unusual structural feature of text B27 already discussed relates to rootrank labels, so it is understandable that the mean proportions with and without that text are very different here. For text B27 alone,

[8] In the computation of this paragraph, a label-pair fails to match if one contains some functiontag and the other has none.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

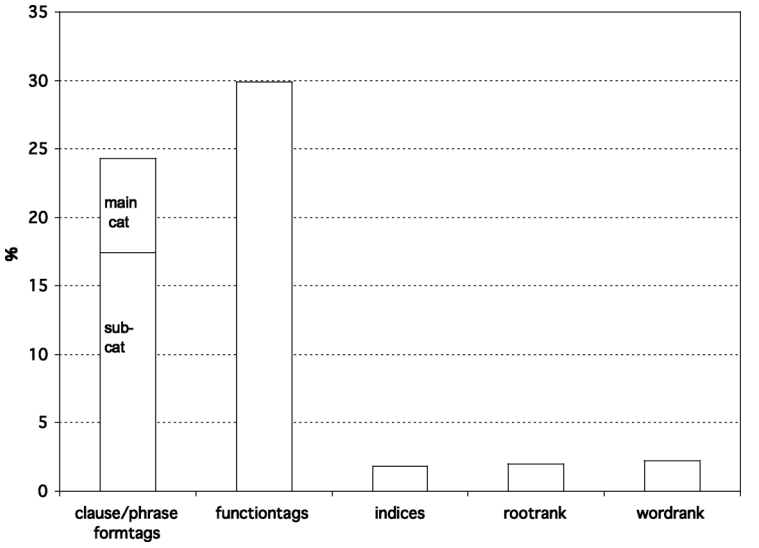


Fig. 1. Contributions to discrepancy, 9-text sample.

17.5% of total discrepancy – almost nine times the average in the other texts – is attributable to rootrank discrepancies.

The mean agreement figure ignoring wordrank discrepancies is 0.947 (0.952); the mean proportion of overall discrepancy attributable to wordrank discrepancies is 2.1% (2.2%).

For the nine-text sample omitting the unrepresentative text B27, the above findings are displayed graphically in Figure 1.

The various categories of analytic discrepancy examined above account between them for about 60 per cent of the total discrepancy figures. A small proportion of the remaining 40 per cent could be due to node-labels simultaneously failing to match in two or more respects: if counterpart labels differ with respect both to formtag and to occurrence of an index, relaxing the matching requirements for either of these annotation elements alone will not convert failure to match into a successful match. But it seems likely that the great majority of the unexplained 40 per cent of total discrepancy relates to the multifarious ways in which analysts' trees may differ not just in labelling but in structure and number of nodes. We have not devised insightful ways of classifying discrepancies of that kind that would allow us to break the 40 per cent figure down into meaningful categories.[9]

[9] To our regret, we also see no easy way to generate a confusion matrix showing which specific tags tend to alternate where annotations are discrepant. The problem is that where a pair of lineages disagree, their lengths are not necessarily the same and, even if they are, differences are not necessarily restricted to a single position. So an automatically-generated confusion matrix would require a definition of correspondence between non-identical nodes across lineages which would be far from trivial to achieve satisfactorily.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

Considering the results for the discrepancy categories we have examined, it is no surprise that the figure for formtagging should be relatively large; almost every node of a parsetree has a formtag (the only exceptions being ghost nodes), and as already said the formtag system provides for far more potential distinctions than any other aspect of the annotation scheme. So there is plenty of room for inter-annotator disagreement in the area of formtags. (On average, 97.8% of nonterminal nodes in a sample text contain a formtag; because of the various constraints on combinations of main category and subcategory symbols, it is hard to say exactly how many distinct valid formtags would be available in principle, but our sample texts contain 461 distinct formtags, which is certainly only a fraction of the total possibilities.) It is more noteworthy that the proportion of total discrepancy attributable to functiontagging is even larger: under the SUSANNE scheme only a minority of nodes are assigned any functiontag, and the distinctions between functiontags are far fewer. (On average, 37.8% of nonterminal nodes in a sample text contain a functiontag, and functiontags are single characters selected from a range of 23 possibilities.)

8 Sampling discrepancies to assign responsibility

The next question to ask is how far inter-annotator discrepancies arise from human error as opposed to vagueness or inadequacy in the explicit annotation scheme.

This question can be addressed only through close attention to specific discrepancies between the two analysts' annotation decisions. To examine every discrepancy in our data in the necessary degree of detail would be a larger task than we could undertake, so we looked at a sample subset. For this purpose, a sample was constructed by considering every twentieth discrepantly-parsed token. (A token is 'discrepantly-parsed' if there is any difference between its lineages in the respective analysts' parse-trees.)

Since much of what follows depends on counting individual discrepancies, we should be explicit about how we have individuated discrepancies in our sample. Not every separate difference between nodes in a pair of lineages is a separate discrepancy, because there are frequently dependencies between the choice of label for one node and the choice of label for its mother or daughter node. For instance, the alternative lineages for *same* in the passage:

> ... it was understandable that the staff were having a tough time of it, pretending that we were all the same ... (C03.00859)

were:

```
D:e ] Fn:o Tg@ Ns:o Fn:s S O
D:e ] Fn:o Tg:h Fn:s S O
```

corresponding to the fact that one of us analyzed the *pretending* present-participle clause (`Tg`) as appositional to the *tough time* singular noun phrase (`Ns`), and hence an immediate constituent of that phrase, while the other treated the *pretending* clause as

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

an IC of the *were having* nominal clause (Fn). A fundamental rule of the SUSANNE annotation scheme is that an IC of a clause receives a functiontag but an IC of a phrase does not; the analyst who treated the *pretending* clause as an IC of the *were having* clause tagged it in accordance with §5.148 as :h, a Manner/Degree adjunct (while in the other lineage it is marked as appositional, @). Thus not only does one lineage contain a label Ns:o having no corresponding label in the other lineage, but the labels for the *pretending* clause contrast as Tg@ versus Tg:h; these are all interrelated consequences of the single choice between treating the *pretending* clause either as appositional to the *tough time* phrase or as a clause IC, so they are jointly counted as a single discrepancy.

On the other hand, a pair of lineages may contain independent discrepancies at different levels in the parse-trees. The lineages for *1976* in the sentence:

> Section 7 . . . refers to "a justice of the peace;" consequently, the procedure under section 7(5) must come within the closing words of section 121(1) of the Act of 1980, namely, that it is a hearing that by virtue of an enactment, namely the Bail Act 1976, may take place before a single justice. (B31.00088)

are:

```
Nns@ ] Ns P:c Fr Ns:e^ Fn@ Np P:h S- S O
[ M@ Nns@ ] Ns P:h Fr Ns:e^ Fn@ Np P:p S- S O
```

Working from left to right, the presence of [M@ at the foot of one lineage and not the other corresponds to the fact that one analyst has treated *1976* as appositional to *the Bail Act* while the other has treated *the Bail Act 1976* as a four-word proper name without internal structure; the contrast between P:c and P:h two levels above the right-boundary symbol means that the *by virtue of* prepositional phrase (P) was seen by one analyst as a Contingency adjunct and by the other as a Manner/Degree adjunct within the *may take place* relative clause; and the contrast between P:h and P:p, five levels higher, shows that the *within* phrase was taken alternatively as a Manner/Degree adjunct or a Place adjunct within the *must come* clause. These represent three independent decisions by the respective analysts, and are counted as three separate discrepancies.

When parse-trees are fairly deep (i.e. lineages are long), the same discrepancy at a high level may show up in successive lines of our sample. For instance, the preceding line in the sample to the one just discussed is for the word *namely* following *the Act of 1980,* earlier in the same sentence, and for this word the sole difference between lineages is the contrast between P:h and P:p for the *within* phrase, already considered. When the same discrepancy between a pair of trees recurs in more than one lineage-pair of our sample, it is counted once only in our statistical analysis. On the other hand, where identical discrepancies occur in the analysis of separate instances of identical wording, each instance is counted separately.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

9 Breakdown of the discrepancy sample

A handful of cases in our sample were set aside for special reasons. In one case, the BNC SGML file represented the typography of the original document in a misleading way, and this seemed to be the main reason for discrepancy between the two analyses; that case consequently can shed little light on the application of the scheme to genuine English text. Also, analysts were expected to correct clear misprints in a text before annotating it, and there were two debatable misprints each of which had been corrected by one analyst and not the other. The ability of analysts to agree on what is or is not a misprint, while an interesting issue, is not the issue we are concerned with here.

Furthermore, there were a number of cases in the sample where parse-tree discrepancies were consequences of differences of opinion between the analysts about what the wordtag for a word should be. An example related to the word *following* in:

> ... only a homemade index of reviews following scanning of newspapers or journals is likely to be effective. (B07.00654)

Out of context, the scheme permits the word *following* to be tagged either as a preposition or as a present participle, and the two analysts have made different choices in this case. If *following* is a preposition, then the sequence it introduces is a prepositional phrase, but if *following* is a present participle then that sequence is a nonfinite clause. The experiment reported in Babarczy *et al.* (2006) has already examined wordtagging performance in depth, so in the research reported here it seemed better to focus on the much larger set of parse-tree discrepancies not explainable in terms of wordtag differences. Similarly, we excluded a couple of cases where the parse-tree discrepancies stemmed from different tokenization decisions by the analysts. Altogether, 16 cases in our sample were eliminated from analysis.

The 294 remaining discrepancies are classified as follows:

A Discrepancy resulting from violation of an explicit feature of the annotation scheme: 173 (58.8%)

- A.i Correction chiefly requires close attention to detail(s) of the scheme: 113 (38.4%)
- A.ii Correction chiefly requires close attention to the meaning of the text: 35 (11.9%)
- A.iii Error represents a typing mistake or careless slip: 25 (8.5%)

B Although the meaning of the text is clear, the scheme does not yield a single, unambiguous annotation decision: 58 (19.7%)

- B.i The scheme is vague about the boundary between alternative annotations: 57 (19.4%)
- B.ii The scheme is contradictory: separate guidelines explicitly require incompatible annotations: 1 (0.3%)

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

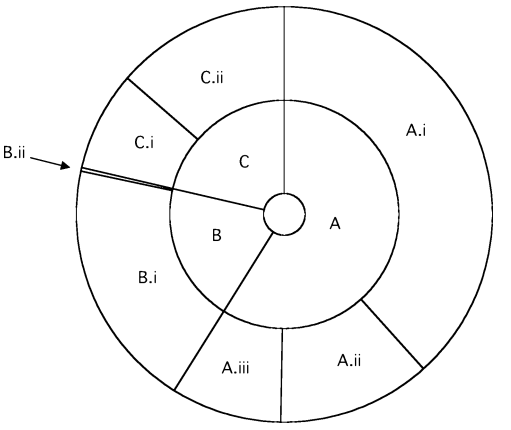


Fig. 2. Proportions of discrepancies of various types.

C Discrepancy corresponds to structural ambiguity in the text;[10] either construal is defensible: 63 (21.4%)

C.i The contrast between alternative interpretations corresponds to a 'real', humanly-significant difference: 23 (7.8%)

C.ii In context the contrast is a logical distinction without a significant difference: 40 (13.6%).

Figure 2 displays these proportions graphically.

Some examples will clarify these categories, and should at the same time give readers a sense of the level of analytic detail at which annotation predictability breaks down:

A.i, discrepancies depending for resolution on details of annotation scheme

> ... put your skincare creams into small, nonbreakable, plastic bottles, ... (B27.00865)

The comma following *small* implies that that adjective is coordinated asyndetically with what follows; but one analyst treated *small, nonbreakable* as a tagma modifying the phrase *plastic bottles*, while the other treated *small, nonbreakable, plastic* as a three-way co-ordination. Arguably, the former analysis better represents the sense, but §4.517 makes the comma after *non-breakable* decisive in favour of the latter analysis.

> Nor is the procedure under section 7(5) "the hearing of complaint." (B31.00096)

[10] Where an example was structurally ambiguous, analysts might happen to choose the same meaning and annotate identically, or they might happen to choose different meanings, yielding a discrepancy; the theoretical possibility that both analysts would notice the ambiguity and provide multiple annotations did not arise in our experiment, because analysts were asked to give just one annotation for any example. Noticing alternative interpretations after one has found an apparently satisfactory way of understanding a passage does not seem to be very natural human behaviour, so if analysts *had* been asked to identify all possible annotations we would not have expected good results.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

The placing of the closing inverted commas after the sentence-final stop is a common publishers' convention, if an illogical one. Our annotation scheme normally treats sentence-final punctuation marks as sisters rather than daughters of the sentence tagmas they bound, but treats paired inverted commas as sisters below the same mother node. In this case, one analyst has placed the sentence boundary before the full stop, the other has placed it after the inverted commas. §4.77 is decisive in favour of the former analysis.

A.ii, discrepancies depending for resolution on subtleties of meaning

> Her dress stuck to the back of her legs with perspiration . . . (C07.00843)

The analysts have functiontagged the *to* phrase respectively as an adjunct of Direction or as a prepositional object. For a car passenger's dress to stick to the back of her legs does not imply movement of the dress toward the legs; the word *to* is determined by the verb *stuck* here rather than carrying an independent meaning of its own, i.e. the phrase is a prepositional object.

> . . . whether the justice had power to adjourn the proceeding before her under section 7(5) of the Bail Act 1976 to the following Monday. (B31.00117)

One analyst treats *before her under section 7(5) of the Bail Act 1976* as a Whiz-deleted relative clause postmodifying *proceeding* – '[which was] before her under . . . '; the other treats the object of *adjourn* as just *the proceeding before her*, and treats the *under* phrase as an adjunct of *adjourn*. In context it appears that the reference to the Bail Act is intended as a potential source of authority for the act of adjourning, it does not describe the proceeding or how it comes to be before the J.P.

A.iii, typing mistakes and careless slips

> It is very rare to have this feeling with another person. (C07.00947)

The phrase *this feeling* should be labelled `Ns:o`, as a formally singular noun phrase which is functioning as direct object of its clause; one analyst has accidentally omitted the colon separating formtag from functiontag, resulting in a notation which is invalid within the scheme.

> But managing director Mr Peter Jarvis admitted . . . (B09.00343)

The subject of *admitted* is the noun phrase *managing director Mr Peter Jarvis*. One analyst has placed the left-hand boundary of the noun phrase so that it includes the word *But*, which in reality is obviously an IC of the sentence.

B.i, discrepancies arising from scheme vagueness

> Bookings can be made at the Society's Offices, 42 Merrion Sq. Dublin 2. tel. 01767053. (B22.00154)

One analyst has included the 'proper name' subcategory symbol (`Nn`) in the label for the noun phrase *the Society* (which in context refers to the Irish Georgian Society), the other has not. The rules for deciding when a noun phrase consisting of words that have uses as common nouns should take the `Nn` subcategory are

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

subtle, and depend heavily on typography, namely whether capitalization would be recommended by a named standard house-style manual (§4.144, and see also §4.153). In the present context it is not clear whether the capital S is required; and the capitalization of *Offices* is non-standard, suggesting that the style of the document tends to overcapitalize so that the capitalization of *Society* should perhaps be discounted.

> Last time the Democratic-controlled Congress sought . . . (B09.00313)

The form *Democratic-controlled* (which by the scheme is divided into three tokens, the hyphen being assigned a leaf node of its own in the parse-tree) is a past-participle clause, within which *Democratic* is functiontagged as Agent of the verb *controlled*. The analysts differ on whether *Democratic* is formtagged as an adjective phrase (`J`) or as a noun phrase with adjective rather than noun head (`Nj`, as in e.g. *the poor*). The agent in a clause will normally be a nominal rather than adjectival element, and *Democratic* here appears to be a reference to the Democratic Party; on the other hand the word is uncontroversially an adjective rather than a noun, and within a hyphenated compound there is inevitably no surrounding material such as a definite article to show that *Democratic* is functioning as a noun phrase. The wording in fact seems slightly unexpected (arguably, *Democrat-controlled* would be the expected usage), and this perhaps excuses the failure of the scheme to resolve the annotation issue.

B.ii, discrepancy arising from scheme contradiction

> An evening had been set aside for a get-together of the Branch's top 50 clients . . . (B42.00204)

One analyst labels the *for* phrase as a Contingency adjunct, the other as a Benefactive adjunct. If considered in isolation, the reference to 'purpose' in §5.158 would justify the Contingency classification, and the *for*-phrase examples in §5.184 would justify the Benefactive classification; but by the rules of the scheme only one of the two can be selected.

C.i, significant structural ambiguities

> Sterling lost ground against a buoyant German mark as dealers took fright at the prospect of a fall in exports leading to a widening of the current account deficit last month. (B09.00259)

This sentence contains two C.i cases: the *as* clause was marked alternatively as a Time or a Contingency adjunct (*as* = 'while' or *as* = 'because'); and the phrase *last month* was either an IC of the *leading* clause, or a postmodifier of *deficit* – if the former, the feared widening would have occurred last month, but if the latter, the fear was that last month's deficit would widen at a later date. Each of these contrasting interpretations seems reasonable.

> I had set my sights on getting a good position in training so that I would be sent . . . (C03.00933)

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

The phrase *in training* was either an IC of the *getting* clause, or a postmodifier of *position*. This makes a real difference: the former structure implies that the aim during training was to get a good position (for the future, after training), the latter would be the appropriate analysis if a 'position in training' is itself something which may be favourable or unfavourable. Perhaps someone more familiar than the present authors with life in the Foreign Legion could reliably resolve this ambiguity, but for us either reading is plausible.

C.ii, nonsignificant structural ambiguities

As pointed out elsewhere (e.g. Sampson 1995: §4.34), structural ambiguities which are real enough logically can often be 'distinctions without a difference' in practice. Consider the following examples:

> At one point the Nawab reached across Olivia to pull down the blind on her window, as if wanting to spare her the sight of all that parched land. (C07.00836)

The *as if* clause was an IC either of the *reached* clause or the *pull* clause. Since reaching across to pull down the blind was a single action with a single purpose, in human terms there is no significance in the question whether the reaching or the pulling seemed intended to spare Olivia.

> The good news is for members in the Republic of Ireland where no premium increases are proposed during 1993. This means that members in the Republic of Ireland who are over twenty five years of age . . . have had no increase in premium for over three years and for all other policyholders there has been no increase for over two years. (B42.00109–10)

The clause *and for all other policyholders there has* … was coordinated alternatively with the *This means* main clause, or with the *have had* subordinate clause. In the former case, only the three-year-plus premium stability for Irish over-25s is claimed to be an entailment of the first sentence, and the two-year lack of increase for other policyholders is an independent assertion; in the second case, both lacks of increase are claimed to follow from the good news in the first sentence. Logically, these interpretations are clearly distinct, but no policyholder will care which is intended; the financial implications are identical.

One might be tempted to feel that the incidence of C.ii-type discrepancies merely demonstrates that aspects of the SUSANNE annotation scheme are too fine-grained, and do not correspond to distinctions that have any reality in the language as used by its speakers. But it would be hard to sustain that argument. What appear to be, structurally, the very same contrasts that in some contexts lead to a C.ii discrepancy, in other contexts make large differences to the sense of examples. The prevalence of C.ii cases, in our view, is telling us something significant about language, not just about a specific annotation scheme. We return to this point in our Conclusions section.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

Table 3. *Inter-annotator discrepancy resolutions compared.*

	Sb	S?	Ss
Bb	2	3	1
B?	7	9	10
Bs	1	9	15

10 Monitoring for bias

The process by which discrepancies were classified as analyst errors or limitations in the annotation scheme is open to criticism, in that it involved one coauthor (GRS) referring to the definition of the scheme in order to resolve differences between his own and the other coauthor's annotation decisions.

As one check on whether the results represented mere bias in favour of one's own views (in which case they would tell us little), we looked at how often resolutions of the 173 type-A cases sided with the respective analysts. The answer was that 109 cases were resolved in agreement with GRS's original decisions, 62 were resolved in agreement with AB's original decisions, and in two cases it appeared on reconsideration that neither analyst's original annotation was correct. There is an imbalance in these figures; but that is to be expected, when one considers that GRS had been the researcher primarily responsible for the annotation scheme for two decades, and the compiler of Sampson (1995), while AB had been working with it for about four years. The fact that in more than a third of these cases GRS concluded that his own original decision was wrong and AB's correct suggests that personal bias was not a major factor in the classification process described in this section.

Nevertheless, as a further check, a random subset of 57 of the discrepancies classified by GRS as type A, B, or C (that is, not discrepancies relating to wordtagging) were independently resolved by AB. The two analysts' resolution decisions are compared in Table 3, in which Bb stands for 'AB regarded her own original annotation as correct and GRS's as incorrect,' B? for 'AB regarded neither analyst's original annotation as clearly more correct than the other,' Bs for 'AB regarded her own original annotation as incorrect and GRS's as correct,' and Sb, S?, and Ss similarly classify GRS's discrepancy-resolutions. If bias by an analyst in favour of his or her own original decision were an important factor, one might expect to find $Bb > Sb$ and $Ss > Bs$ (using these symbols to stand for the totals in the rows and columns they label). In fact, $Bb = 6$ and $Sb = 10$, a large difference in the other direction; and $Ss = 26$ and $Bs = 25$, where the difference is in the predicted direction but proportionally very small. We believe that personal bias can be discounted as a factor affecting our findings.

11 Conclusions

The above findings yield three broad conclusions, all of which are significant for natural language processing, and some of which are also interesting for what they tell

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

us about the nature of human language, independently of information technology applications.

11.1 Human fallibility more significant than definitional limitations

In the first place, it is clear that (for these analysts and this scheme) the limitation on annotation predictability is due far more to human fallibility than to the limited precision of the scheme. Type A discrepancies are almost three times as numerous as those of type B. It is of course logically possible that other individuals, or the same individuals after accumulating even more years of experience, might succeed in reducing their incidence of type-A discrepancies. But, assuming that we shall not encounter people willing to dedicate entire careers to honing their grammatical-annotation skills, we question whether the A versus B differential could be much reduced. As in the case of word classification, so with parsing it seems that (as it were) our ability to make the size of clouds a well-defined property runs well ahead of our ability to measure cloud size in practice.

In the case of word-classification we asked the subsidiary question whether those discrepancies that did arise from annotation-scheme inadequacies might be reduced by revising the scheme; and we reached cautiously optimistic conclusions. We do not feel able to say anything analogous about possible scheme revisions in the case of structural annotation. One can often see how to sharpen the logical boundaries between pairs of wordtags while being confident that the revision will not impinge in unforeseen ways on the rest of the word-classification scheme. In the case of higher-level structure, there are so many linkages between properties of units at different levels that it is hard to feel sure that modifying one definition will not create new uncertainties in other definitions.

11.2 Structural ambiguity often pragmatically nonsignificant

The second general conclusion is that, not only is genuine structural ambiguity quite common in English, but (more surprisingly) there is often no reason to resolve ambiguities because in practice either interpretation amounts to the same thing. Instances of type C.ii are almost twice as numerous as those of type C.i.

This generalization might seem to imply that the annotation scheme is over-refined, postulating artificial structural distinctions that lack real linguistic validity. But the finding cannot be explained away so easily. In these particular contexts, the relevant structural distinction is nonsignificant; but in other contexts, with different vocabulary, the same distinction may make a large difference. Compare the first example under C.ii above (. . . *the Nawab reached across Olivia* . . .) with the following invented examples:

> he crawled out of the tent to mount his horse, as if leaping onto a royal throne – he limped up the aisle to lead the prayers, as if suffering from gout.

The relevant aspects of grammar are the same in all three cases; but in these latter examples, one structural interpretation is clearly correct and the other wrong – *as if*

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

leaping must be an adjunct to the *mount* clause, *as if suffering* must be an adjunct to the *limped* clause. (Or, more precisely, under the alternative structural construals the examples would be saying something quite different and implausible: the crawling would somehow resemble leaping onto a throne, something about the way prayers were led would indicate gout.)

We believe this is representative for the range of C.ii-type discrepancies in our data. The structural differences are ones which a satisfactory annotation scheme must recognize, because they sometimes correspond to important meaning-differences; but in many contexts they do not. Grammar appears to be a tool like, say, a ruler marked off in 32nds of an inch which is commonly used simply to check whether one has picked up a seven-inch or an eight-inch bolt. So far as we are aware, this is not a consideration that has been much noticed by linguists.[11]

The objection that the annotation scheme may be over-refined has sometimes been expressed to us in a different way. Consider the example:

> Scatter the brioche cubes in six small gratin ... dishes. Place four slices of pear on top, then cover each one ... (B18.00918–9)

– where the analysts have functiontagged *on top* respectively as a Place and as a Direction adjunct. In order to place pear slices on top of a dish of brioche cubes, it is evidently necessary that the pear slices are moved toward the dish (Direction); on the other hand, arguably the focus here is on the static end-point of that movement (Place). No example in the relevant scheme sections seems to offer a good precedent for this case, so we classified it as a type-B.i discrepancy. But, the objection runs, if we understand the English word *place* and the phrase *on top* in this example, it is not clear that there remains any further question to answer about the meaning of the clause; the annotation system is forcing us to create spurious facts, rather than merely recording facts that exist independently of the scheme.

At root this seems to be an objection to the general concept of grammatical annotation. Clearly, a speaker of a language will normally understand examples of that language, and this understanding surely involves something beyond knowledge of the meanings of individual words (otherwise the words of a sentence could be freely permuted without affecting meaning). If so, then there must surely be *some* way of modelling symbolically the aspects of meaning that relate to relationships between successive words, rather than to the meanings of the words in isolation. The SUSANNE scheme for English may well not be the ideal approach, but in that case it presumably needs to be replaced by a better scheme – rather than by no scheme at all. (As it happens, language engineers have recently found it worth devoting intensive study to the area of 'semantic role labelling,' from which the above example is taken – see e.g. Gildea and Jurafsky (2002); Xue and Palmer (2004); Màrquez, Surdenau, Comas, and Turmo (2005) – though it seems fair to say that this research has focused on the task of devising algorithms to assign labels automatically, rather than

[11] Though cf. Edward Sapir's analogy (1921: 14) of a dynamo that is capable of powering a lift but used mainly to supply a doorbell.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

on the problem of defining adequate sets of labels and watertight boundaries between them.)

11.3 Functiontagging specially problematic

Thirdly, it seems that assigning functional categories to clause constituents is a specially problematic area of structural classification. This came as no surprise to us who have worked with the SUSANNE scheme over a long period; it has been clear from the outset that developing a satisfactory set of functional categories with watertight boundaries was an unusually troublesome aspect of scheme definition. Nevertheless, it is difficult to understand how the English language works, if some such category-set does not apply to it.

Mathematical logicians distinguish the arguments of a predicate in terms of numerical order, but that cannot be the crucial factor deciding the semantic relationship between an English verb and its arguments: in *he travelled to Wells by car* and *he travelled by car to Wells* the relationship between Wells and travelling, and the relationship between the car and travelling, are the same (Wells is always the destination, the car always the means) rather than interchanged as they would be if numerical order were decisive. But prepositions such as *to* and *by* do not identify argument-relationships unambiguously either; *by* expresses a different relationship in *he stood by the door*, and *to* in *he gave the box to Julie* expresses the same relationship which is alternatively expressed by position in *he gave Julie the box*. Many linguists, from Fillmore (1968) onwards, have believed that the most plausible way to define semantic verb/argument relationships in English and other natural languages is via a limited set of 'cases' or functional categories, onto which prepositions and positional roles can be mapped in a predictable though not straightforwardly one-to-one way. The SUSANNE functiontag set was in fact developed from the most fully worked out version of Fillmorean case theory that we could find when we began work (namely Stockwell, Schachter, and Partee 1973), 'debugged' through the process of applying it to the text samples of the SUSANNE treebank.

If numerical sequence is not crucial, prepositions are frequently ambiguous, and a satisfactory set of case roles cannot be found, there would be a large unanswered question about how English-speakers understand the relationships between verbs and their arguments. This question needs answering for wide-coverage NLP systems that aim to incorporate an inferencing function. But the question is also challenging for pure linguists, because it implies a mystery about a quite central aspect of language structure.

(Scholars associated with Charles Fillmore have recently developed his case theory into a system of 'frame semantics' independent of and more elaborate than the SUSANNE functiontag scheme; see particularly Ruppenhofer, Ellsworth, Petruck, and Johnson (2005). We have not attempted – we are not qualified – to check whether this FrameNet scheme yields closer inter-annotator agreement on unseen language samples than we have obtained for SUSANNE functiontagging, and we are not aware that the FrameNet researchers have studied this issue.)

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press

We believe that the admittedly limited data-set we have described offers findings that should be of interest both to practitioners of computational natural language processing, and to students of linguistics as a human science.

References

Babarczy, Anna, Carroll, J. A. and Sampson, G. R. 2006. Definitional, personal, and mechanical constraints on part of speech annotation performance. *Journal of Natural Language Engineering* **12**: 77–90.

Bird, S. and Liberman, M. 2001. Linguistic annotation. www.ldc.upenn.edu/annotation/

Fillmore, C. J. 1968. The case for case. In E. Bach and R. T. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart & Winston, pp. 0–88.

Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics* **28**: 245–88.

Kübler, Sandra and Telljohann, J. 2002. Towards a dependency-oriented evaluation for partial parsing. In *Proceedings of the Workshop 'Beyond Parseval – Towards Improved Evaluation Measures for Parsing Systems'* LREC 2002, Las Palmas, 2 June 2002, pp. 9–16.

Manning, C.D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Màrquez, Ll., Surdeanu, M., Comas, P. and Turmo, J. 2005. A robust combination strategy for semantic role labeling. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. Vancouver, pp. 644–51.

Ruppenhofer, J., Ellsworth, M., Petruck, Miriam R. L. and Johnson, C. R. 2005. *FrameNet: Theory and Practice*. framenet.icsi.berkeley.edu/book/book.html

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

Sampson, G. R. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics* **5**: 53–68.

Sampson, G. R. 2001. Demographic correlates of complexity in English speech. In G.R. Sampson (ed), *Empirical Linguistics*. London: Continuum, pp. 57–73.

Sampson, G. R. and Babarczy, Anna. 2003. A test of the leaf-ancestor metric for parse accuracy. *Journal of Natural Language Engineering* **9**: 365–80.

Sapir, E. 1921. *Language*. New York: Harcourt, Brace & World.

Stockwell, R. P., Schachter, P. and Partee, B. H. 1973. *The Major Syntactic Structures of English*. New York: Holt, Rinehart & Winston.

Xue, N. and Palmer, Marta. 2004. Calibrating features for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. Barcelona, pp. 88–94.

Xue, N., Xia, Fei, Chiou, Fu-Dong, and Palmer, Marta. 2005. The Penn Chinese TreeBank: phrase structure annotation of a large corpus. *Journal of Natural Language Engineering* **11**: 207–38.

https://doi.org/10.1017/S1351324908004695 Published online by Cambridge University Press