Natural Language Engineering **12** (1): 77–90. © 2005 Cambridge University Press
doi:10.1017/S1351324905003803 First published online 6 December 2005 Printed in the United Kingdom

Definitional, personal, and mechanical constraints on part of speech annotation performance[*]

ANNA BABARCZY,[1] JOHN CARROLL
and GEOFFREY SAMPSON
Department of Informatics, University of Sussex,
Falmer, Brighton BN1 9QH, UK

(*Received 31 December 2003: revised 4 October 2004*)

Abstract

For one aspect of grammatical annotation, part-of-speech tagging, we investigate experimentally whether the ceiling on accuracy stems from limits to the precision of tag definition or limits to analysts' ability to apply precise definitions, and we examine how analysts' performance is affected by alternative types of semi-automatic support. We find that, even for analysts very well-versed in a part-of-speech tagging scheme, human ability to conform to the scheme is a more serious constraint than precision of scheme definition. We also find that although semi-automatic techniques can greatly increase speed relative to manual tagging, they have little effect on accuracy, either positively (by suggesting valid candidate tags) or negatively (by lending an appearance of authority to incorrect tag assignments). On the other hand, it emerges that there are large differences between individual analysts with respect to usability of particular types of semi-automatic support.

1 Introduction

A number of authors (Voutilainen 1999; Brants 2000) have explored the ceiling on consistency of human grammatical annotation of natural-language samples. It is not always appreciated that this issue covers two rather separate sub-issues:

(i) how refined can a well-defined scheme of annotation be?
(ii) how accurately can human analysts learn to apply a well-defined but highly-refined scheme?

The first issue relates to the inherent nature of a language, or of whichever aspect of its structure an annotation scheme represents. The second relates to the human ability to be explicit about the properties of a language. To give an analogy: if we

[*] A version of this paper was presented orally at the workshop "Empirical methods in the new millennium: Linguistically Interpreted Corpora" (LINC-01), at the 34th Meeting of the Societas Linguistica Europaea, Leuven, Belgium, 28 Aug–1 Sep 2001. The research was supported by the Economic and Social Research Council (UK) under award no. R00023 8146.

[1] Babarczy now works at the Budapest University of Technology and Economics.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

aimed to measure the size (volume) of individual clouds in the sky, one limitation we would face is that the fuzziness of a cloud makes its size ill-defined beyond some threshold of precision; another limit is that our technology may not enable us to surpass some other, perhaps far lower threshold of measurement precision.

The analogy is not perfect. Clouds exist independently of human beings, whereas the properties of a language sample are aspects of the behaviour of people, including linguistic analysts. Nevertheless, the two issues are logically distinct, though the distinction between them has not always been drawn in past discussions.[2] The main aim of the series of experiments reported here was to begin to explore the quantitative and qualitative differences between these two limits on annotation consistency. The domain of these experiments was English wordclass tagging.

A further aim was to investigate how far annotation accuracy is affected by particular types of semi-automatic support for human annotation. (Note that we are not, in this paper, concerned in any way with the separate question of what levels of accuracy are achievable by automatic annotation systems – an issue which has frequently been examined by others, e.g. Leech, Garside and Bryant 1994; Oliva 2001; Tateisi and Tsujii 2004.)

Our experiments involved independent application to the same language samples of a highly-refined wordtagging scheme (the SUSANNE scheme of Sampson (1995), supplemented with the further refinements of Sampson (2000, sections 13–14)) by two analysts who are particularly well-versed in the details of that scheme (i.e. the first and third co-authors).

The SUSANNE scheme is one of many extant schemes of linguistic annotation, including wordclass annotation, but it is atypical with respect to the priority it gives to precise definition of maximally-refined analytic distinctions; various commentators have made remarks that tend to confirm this (e.g. "the detail . . . is unrivalled" (Langendoen 1997: 600), "Compared with other possible alternatives such as the Penn Treebank . . . [t]he SUSANNE corpus puts more emphasis on precision and consistency" (Lin 2003: 321). The tagset comprises 333 tags, excluding tags for punctuation (the latter are normally uniquely determined given the punctuation mark to be tagged), and excluding the additional wordtags defined in Sampson (2000), which were ignored in this experiment because they mainly relate to speech phenomena such as swearing, while our experiments used written material.

Since we wish to study the ceiling on human annotation performance (we are not interested, here, in how easy or difficult beginners find it to acquire skill in applying an annotation scheme), it is necessary to use analysts who know the scheme about as well as anyone has ever known it – which severely limits the candidates available.

[2] The distinction we are drawing is not the same, for instance, as Dickinson and Meurers' (2003) distinction between "ambiguity" and "error": by "ambiguity" Dickinson and Meurers are referring to cases where a linguistic form (their example is English *can*), taken out of context, is compatible with alternative annotations but the correct choice is determined once the context is given. We are interested in cases where full information about linguistic context and annotation scheme may not uniquely determine the annotation of a given form. On the other hand, Blaheta's (2002) distinction between "Type A" and "Type B" errors, on one hand, and "Type C" errors, on the other, does seem to match our distinction.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

Our only source of data comes from disagreements between analysts on the tagging of particular words, so we need at least two analysts. Ideally we might learn much more from the output of a larger team, but the practicalities of academic research make it unlikely that many more than two suitable analysts will ever be available at the same time.

The language samples used in the experiments consist of nine extracts each of 2000+ words (i.e. 2000 words plus a few more so that extracts begin and end at natural breaks such as paragraph boundaries), taken from randomly-chosen locations within files drawn from the written-English section of the British National Corpus (Leech 1992).[3] The files were all categorized as "informal" rather than "polished" – that is, they were either unpublished documents, or if published they appeared not to have been subjected to the copy-editing processes commonly applied in the production of books and wide-circulation magazines. Within the informal written subset of the British National Corpus, the choice of particular files was random. (They comprised three business letters, a reader's letter in a health club magazine, an informal essay on after-school activities, two private letters, and two formal reports on matters of education.)

It was desirable for our experiments to use written rather than spoken language, so that the facts which interested us would not be complicated by irrelevant problems relating to dysfluencies and transcription ambiguities. The choice of informal rather than polished writing was a consequence of practical exigencies: we had extracts from both categories of written document to hand, but we had already worked on the polished extracts – the experiments reported here required language samples that neither analyst had examined previously.

Each of the nine extracts was divided into six equal-length (ca 333-word) sections; successive experiments used material containing section(s) from each extract, as detailed below, in order to make the experimental results as directly comparable as possible.

2 Software support techniques

Various semi-automated approaches have been used by different research groups to ease the burden on the annotator and help him or her to work more efficiently. Using a semi-automated system may have the advantage not only of reducing the tedium of annotation and speeding up the process, but also of allowing the annotator to concentrate on the difficult cases while letting the program deal with the easy ones, thus increasing accuracy. On the other hand, a semi-automatic system may create the risk of predisposing the human annotator to accept whichever tag is proposed by the system, even when it is incorrect.

Our experiments examine the trade-off empirically. We contrast the effectiveness (in terms of both accuracy and annotator time) of

(i) wholly *manual* annotation

[3] For the British National Corpus, see www.natcorp.ox.ac.uk

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

(ii) semi-automated *dictionary look-up* annotation: manual selection of tags from a menu of candidate tags for each word generated by an automatic system that incorporates a dictionary
(iii) *postediting*: manual post-editing of an automatic tagging of a sample, using a program that guesses a single tag for each word in context.

Our "dictionary look-up" system incorporates an electronic dictionary (derived ultimately from Hornby (1974), with additions) listing candidate SUSANNE tags for each word in the dictionary. For instance, if the word to be tagged is *can*, the candidate tags are VV0t VMo NN1c, representing transitive verb, modal verb, or singular count noun respectively. When a text is tagged using this tool, successive taggable words of the text are presented on screen in a window of text, together with the candidate tags (if any) listed for the relevant word in the dictionary. The annotator can select a candidate by pressing sufficient keys to identify one of the candidates uniquely; for *can*, for instance, pressing the "n" key selects NN1c, pressing "v" and "m" selects VMo. Alternatively, if the word is not in the dictionary, or the appropriate tag in context is not one of those listed in the dictionary for that word (perhaps because it is a proper name which happens to coincide in spelling with a dictionary word), the annotator can choose to enter a tag manually. If the context shown is insufficient, the annotator has the option of scrolling the window up or down the text before choosing a tag (and a tag once input can be changed at any time if a mistake is noticed).

"Manual" annotation involves using the dictionary look-up tool but with the electronic dictionary disabled, so that each word is treated as if it were a word missing from the dictionary, whose tag must be written in by the annotator. Although the dictionary is disabled, the user interface with its facilities for efficient insertion of wordtags in text files is identical to the dictionary look-up interface – this is relevant when annotation speeds are compared, in section 6 below.

In the "post-editing" technique, annotators use that same interface to correct manually the output of an automatic tagger. Unfortunately no existing automatic tagger is designed to tag in perfect conformity with the SUSANNE scheme, but we used one which comes close to that ideal (and the differences are taken into account, below, in assessing the experimental results). The automatic tagger used assigns tags drawn from a "coarse" tagset produced by eliminating final lower-case letters from those SUSANNE tags that contain them. Some of these represent categories such as countable v. uncountable v. both, among singular nouns, or transitive v. intransitive v. both, among verbs; others relate to "encyclopaedic knowledge", for instance classification of proper names as country names, province names, town names, etc. At the coarse level, 160 tags (plus punctuation tags) are distinguished. The automatic tagger also embodies small systematic differences from the SUSANNE tagging norms, allowing alternative tags for certain word-forms for which the SUSANNE scheme recognizes only one valid tag: for instance, some *-ing* words, such as *disturbing*, can be only present participles in the SUSANNE scheme, while the automatic tagger permits any present participle form to be coded as an adjective if it functions adjectivally.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

The automatic tagging program, developed originally by Elworthy (1994), is based on a first-order Hidden Markov Model. It uses a dictionary containing 51,500 word-forms with candidate tags; it selects among candidate tags for a word in context by reference to trained bigram part-of-speech statistics derived from the SUSANNE Corpus and trained unigram (i.e. word/tag frequency) statistics derived from the same source, supplemented by further open-class unigram statistics derived from the LOB Corpus. Words not in the dictionary are assigned a tag by reference to a statistical model derived from the character-sequence make-up of the dictionary entries.

In all cases, annotators could manually consult the definition of the tagging scheme on paper in relevant sections of Sampson (1995, 2000), and view the electronic dictionary file used by the dictionary look-up system, in order to decide on correct tags.

3 Experiment 1: The nature of inter-annotator discrepancies

Our first experiment investigated the extent of agreement between the two experienced annotators tagging the same material, and the reasons for disagreement where disagreements occurred – in particular, whether these represented failings by the annotators or limitations of the annotation scheme.

The material annotated independently by the two annotators in Experiment 1 comprised the first two out of six chunks in each of the nine BNC extracts: 7155 lines, containing about 6000 words. (The number of lines in our "one-word-per-line" files is always greater than the count of "words" in the everyday sense, because some lines correspond to punctuation marks and to orthographic indicators, for instance paragraph-boundary markers.)

For Experiment 1, both annotators used the dictionary look-up tool. Table 1 displays the level of inter-annotator agreement in Experiment 1 at three levels of precision. At the "fine" level of comparison, annotators were counted as agreeing on the tagging of a word only if they assigned fully-identical SUSANNE tags: 97.4% agreement was achieved at this level. At the "coarse" level, as discussed above, the lower-case letters occurring at the end of some SUSANNE tags were ignored; inter-annotator agreement reached 98.0%. Finally, at the "major part-of-speech" level of comparison just 18 tag categories were distinguished; at this level inter-annotator agreement reached 98.5%.[4]

These figures are broadly comparable to those obtained by other investigators. Church (1992: 3) said that various researchers at that time were reporting figures in the range 95%–99%, and he regarded the former figure as more realistic than the latter. Voutilainen (1999) reported an initial 99.08% agreement[5] between annotators

[4] Where a discrepancy related to "tokenization", that is one annotator treated a sequence as a single taggable unit but the other split it into two or more units, we count that as one discrepancy. If, instead, a discrepancy were counted for each of the divided units, the "fine" agreement figure would fall to 97.0%; we have not done the comparable calculation for the other comparison levels.

[5] Misprinted as "99.80%" in Voutilainen's paper, but the correct percentage can be calculated from his Figure 1.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

Table 1. *Inter-annotator disagreements by level of comparison*

Level of comparison	Number of discrepancies	% agreement
Fine	183	97.4
Coarse	141	98.0
Major part of speech	106	98.5

Table 2. *Inter-annotator disagreements by cause*

		cases	%
1	caused by experimental situation	21	11.5
2	annotator misjudgements/slips	53	29.0
3	participle discrepancies	29	15.8
4	routine scheme extensions needed	6	3.3
5	insufficient information available to annotator	59	32.2
6	scheme inconsistent/vague	15	8.2

on a task that seems to have compared most closely with our "coarse" comparison level (Voutilainen's tagset had about 180 members). Brants (2000) obtained an initial 98.57% agreement in a German-language experiment using a 54-member tagset (intermediate in size between our "coarse" and "major part of speech" levels). We do not know why investigators have obtained different figures, but this could easily be a consequence of the particular distinctions made by the respective tagsets (see, for instance, the discussion of proper names in section 5 following Table 2), the quality and level of detail of the respective annotation guidelines, the particular properties of the experimental texts, or any combination of these. For present purposes we are less interested in our raw figures on agreement levels than in analysis of the nature of inter-annotator disagreements.[6]

We investigated the nature of disagreements by examining the 183 discrepancies at the "fine" level of comparison. A broad categorization is as follows:

1 Discrepancies caused by experimental situation

Most cases (fifteen) under this heading related to a user-interface feature of the dictionary look-up software tool, whereby working at speed it is easy to skip a word and leave it untagged without realizing one has done so. This is of no theoretical interest and could be cured, for instance, by adding a warning feature drawing attention to skipped words. Most other cases in this category (four) related to misprints in the texts being tagged. The annotators had agreed in advance that

[6] Because of the large size of tagsets, wordtagging is a task where it would be very difficult to move beyond simple percentages as a measure of inter-annotator agreement to use the more sophisticated kappa measure discussed e.g. by Carletta (1996). (Reliable estimates of individuals' frequency of usage would be hard to form for low-frequency tags.) But because our focus is on the nature, rather than the overall number, of disagreements, this is of little importance.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

where misprints occurred, they would tag the word that should have appeared rather than attempting to tag the character-string (whether word or non-word) actually found, but they did not always correct the text in the same way. While relevant to the issue of inter-annotator agreement, this seems a rather separate matter from the topic of the current investigation.

2 Annotator misjudgements/slips

This large category relates to cases where the materials available to the annotators implied a particular tag choice, but one of the annotators made the wrong choice. Some cases were straightforward careless slips, for instance choice of the tag TO (infinitival marker) rather than IIt (prepositional *to*) in the phrase *to Norman*. Others were subtler. In the phrase … *is worth four or five years of experience in the field*, one annotator tagged *worth* JA (attributive adjective) rather than II (preposition) — *worth* is not a core example of a preposition and might not be classified as such by traditional linguists, but the SUSANNE scheme (Sampson 1995: 121) defines it as such in this usage. In the year-span *1992–93*, tokenized[7] as *1992–93*, both digit-sequences were tagged MCy (numerical year name) by one annotator, but an easy-to-overlook (and, arguably, not very logical) clause in the scheme (Sampson 1995: 111) specifies that *+93* should instead be MCn (general numeral).

Whether straightforward or subtle, in all these cases the scheme does seem to prescribe an explicit and unambiguous tagging decision.

One subdivision under this heading (eight cases) relates to instances where a word-sequence which the scheme treats as a "grammatical idiom" equivalent to a single word was instead tagged word-by-word, or alternatively where some word sequence that can function as an idiom was tagged as such in a context where the individual words were being used in their normal sense (e.g. *as usual* is often an adverb idiom, tagged RR21 RR22, but in *much as usual* the same words should be tagged IIa JJ).

3 Participle discrepancies

A special problem related to tagging of present and past participles in contexts where they function as adjectives or (in the present-participle case) as nouns, rather than as part of verb groups. The SUSANNE scheme makes choice of wordtags for open-class words depend on some agreed dictionary. Our experiments necessarily used the electronic dictionary we had available; this does allow adjective and/or noun tags as alternatives to verb-participle tags for many participle forms, but it contains no definitions or examples, and hence gives no guidance about which uses of these forms are "nouny"/"adjectivy" enough to justify those tags. The annotators frequently disagreed on whether words like *training*, *swimming*, *schooling*, *crossing* should be tagged as nouns or participles where they functioned as nouns but retained a close semantic connexion to the verb sense. This problem would probably

[7] The plus sign is used to indicate that no whitespace intervenes between a token and the preceding token.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

be eliminated by using a published dictionary whose definitions and examples would indicate which uses were being classified as non-verbal.

4 Routine scheme extensions needed

In a few cases the dictionary-dependence of the SUSANNE scheme, combined with use of a small and somewhat dated dictionary, required annotators to choose a tag which violated the spirit of the scheme, and discrepancies arose because one annotator conformed to the requirement while the other chose the "right" tagging. Most of these cases related to grammatical idioms not listed in either Sampson (1995) or our electronic dictionary: for instance, *in the least* or *better off* have as much claim to idiom status as items listed in the scheme as idioms (and the scheme explicitly prescribes, p. 101, that the idiom list will be extended as new precedents are encountered). For purposes of this experiment we agreed not to add new idioms to the list, but an annotator sometimes found the temptation irresistible.

Likewise, one annotator tagged *navy* in *navy linen trousers* as an adjective, denoting a shade of blue, although our dictionary happens to assign it only the tag NNJ1c, appropriate for the sense parallel to *army*.

Most if not all these discrepancies could be eliminated by using a fuller and more up to date dictionary.

5 Insufficient information available to annotator

This largest category of discrepancies covers cases where the definitions of the scheme appear to lead to a unique tagging, but applying the definitions to the word in question required knowledge not easily available to the annotator. The largest subcategory (25 cases) were proper names. The SUSANNE tagging scheme makes many distinctions among proper names, for instance personal surnames, organization names, town names, etc. have separate tags. However, names are often extended from people to places or organizations, from places to organizations located in those places, and so forth, and in context it can often be meaningless to ask whether a name refers to an organization, to the site of the organization, or to the person after whom place or organization was named. Consequently the rules of the scheme (Sampson 1995: 88) achieve determinism by prescribing, in essence, that any occurrence of a proper name is tagged by reference to its original bearer. For little-known proper names, the first bearer can be hard to ascertain. (Since proper names are frequent, this is likely to be one reason for the slightly lower inter-annotator agreement figures in our experiment relative to other recent investigations using different tagging schemes.) Thus, the experimental material contained references to the *Medau Society*, an organization promoting a form of physical exercise; it emerged that the correct tag for *Medau* is NP1s (surname), rather than NP1j (organization) as chosen by one annotator, because the exercise system was named after the man who pioneered it.

The second-largest subcategory under this heading (eleven cases) was abbreviations, where the SUSANNE scheme makes the tagging depend on the word(s) abbreviated, but an annotator sometimes did not recognize the abbreviation. For instance *HNC* was tagged by one annotator NP1j as an organization name, but in fact stands for *Higher National Certificate* (an educational qualification).

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

The third subcategory (eight cases) related to subclassification of verbs for transitivity or nouns for countability where the verb or noun in question was not included in our electronic dictionary. Under the SUSANNE scheme, if (say) a verb is classified as transitive that means not that the token being tagged functions transitively but that all uses of that verb in the language are transitive. This is hard even for an experienced linguist to judge.

6 Scheme inconsistent or vague

There remain fifteen cases where the fault for the discrepancy lay with the formulation of the annotation scheme: the two annotators' different decisions were both defensible in its terms.

Ten of the fifteen cases related to acronyms. The definition of the tag NP1z, "code name", on pp. 113–14 of Sampson (1995) implies that an acronym functioning as a countable noun takes this tag; however p. 133ff., especially p. 134, imply that such a form should be tagged NN1c, singular countable common noun, or another suitable non-proper tag. Thus *A.G.M.* (*Annual General Meeting*) and *AGR* (*advanced gas-cooled reactor*) were tagged NP1z v. NN1c by the two annotators, and *NVQs* (*National Vocational Qualifications*) was tagged NP2z v. NN2 (plural code name v. plural common noun).

Another problem related to foreign but transparent words in a proper name. An organization called *Associacion Casa Alianza* had its component words tagged by one annotator FW (foreign word) and by the other as NN1c (singular countable common noun). The discussion of the FW tag in Sampson (1995) does not seem to resolve this difference.[8]

Two further problems under this heading related to a plus sign used to stand for *plus* as a linguistic conjunction (rather than part of a mathematical formula), and the interpretation of *sorry* in *sorry I've taken* . . . either as a discourse item like *hello* or *please*, or as head of an adjective phrase ([*I'm*] *sorry* [*that*] *I've taken* . . .).

With this last category of discrepancies, we have reached an area where an annotation scheme which aims to be maximally comprehensive and precise proves not to be totally so. But this category is a small proportion of all discrepancies. Even if we set aside "discrepancies caused by the experimental situation" as irrelevant, the cases in category 6 are fewer than ten per cent of the remainder, or 2.5 per thousand words of text.

True, we have repeatedly appealed to the fact that use of a fuller dictionary would resolve many of the discrepancies in categories 2 to 5, yet we know that no dictionary is perfectly complete. Doubtless even using the most suitable dictionary extant, some of those individual discrepancies would remain unresolved, and ought therefore to be added to category 6. But it is hard to believe that these would amount to more than a small fraction of the 90% of discrepancies currently covered by categories 2 to 5. Indeed, the fact that two-thirds of the cases in category 6 relate to the same specific inconsistency in the SUSANNE scheme suggests that it might be easy to

[8] We ignore here the subsidiary point that *Associacion* appears to be a mistaken form, since the Spanish for "association" is *asociación* with one S.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

Table 3. *Inter-annotator disagreements using different annotation methods (manual v. semi-automatic)*

Level of comparison	Number of discrepancies		% agreement	
	Expt (a)	Expt (b)	Expt (a)	Expt (b)
Fine	159	108	95.5	97.0
Coarse	79	75	97.8	97.9

introduce a new rule eliminating that particular inconsistency and thereby rendering determinate a large proportion of tagging decisions which the scheme as it stands fails to resolve. We are sure that the most comprehensive annotation scheme that could be devised would leave some residue of undecided cases, but it might be that that residue could be even less than 2.5 per thousand words.

Thus Experiment 1 shows that, with a tagging scheme recognized as being as highly refined, or more so, than alternative schemes, inter-annotator discrepancies stemming from imperfections or ambiguities in the scheme are few, relative to discrepancies arising from failure of an annotator to apply the scheme properly to an individual case. In terms of the cloud-measuring analogy of section 1, it is as if the precision with which cloud-size can be defined considerably outruns our ability to measure cloud sizes in practice.

4 Experiment 2: Manual v. semi-automatic annotation

Experiment 2 compared the accuracy of dictionary look-up with manual tagging, with the two annotators using different techniques on the same texts.

The test material in Experiment 2 consisted of two texts each of ca 3000 words (3562 and 3637 lines), made up (a) of the third and (b) of the fourth 333-word chunk of each of the nine BNC extracts. In the first part of the experiment, Annotator 1 used the dictionary look-up system to tag text (a) while Annotator 2 independently tagged the same text manually. In the second part of the experiment the roles were reversed in tagging text (b).

As can be seen from the results in Table 3, inter-annotator agreement may slightly decrease when annotators use different techniques. In contrast with the 97.4% fine-level agreeement rate of Experiment 1, in Experiment 2a we find only 95.5% fine-level agreement – though in Experiment 2b the corresponding figure is 97.0%. At the coarse level, though, the results of Experiments 2a and 2b are both comparable to the results of Experiment 1.

This suggests that the lower fine-level agreement rate in Experiment 2a was caused by Annotator 2 occasionally failing to verify the fine-grained function or usage properties of words (such as transitivity or countability class), while these were automatically supplied for Annotator 1 by the dictionary look-up system. It is often tempting for an experienced linguist to suppose that he knows whether (for instance) a given verb functions in English only transitively, only intransitively,

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

or both ways, so that it is unnecessary to go to the effort of checking against the dictionary which the scheme treats as definitive; but, as we saw above, in practice this confidence can be misplaced.

Thus it seems that automatic support can marginally improve the consistency of tagging performance, but with respect to fine-grained decisions only. The chief reason for using semi-automatic rather than manual techniques, unsurprisingly, is speed (see section 6 below) rather than accuracy.

5 Experiment 3: Alternative support tools

Experiment 3 investigated the extent to which tagging performance is affected by the nature of the software support available. We compared performance using the dictionary look-up system with performance using the postediting technique. Since postediting involves deciding to accept or change a single tag proposed by the automatic tagger for each word-token, whereas the dictionary look-up system shows the range of alternatives, an obvious hypothesis is that a human annotator might be unduly predisposed to accept the automatic tagger's decision, so that overall accuracy would be worse with that system than with dictionary look-up.

Even the dictionary look-up technique involves an element of possible bias in favour of agreeing with the machine. It can happen that the correct tag for a particular token is not among the candidates offered by the dictionary look-up tool: in which case the annotator might well be predisposed to accept one of the displayed candidates rather than choosing the correct, "write-in" tag. But that situation arises only rarely, and where it does arise it is usually because the token is a proper name which coincides with a common word. For instance, *Bush* as the American president's surname is listed in the dictionary only as a countable common noun. In these cases it is easy for an annotator to anticipate the limitations of the system. With the automatic tagger, on the other hand, the candidate tag proposed automatically quite often needs to be changed by the annotator, and such cases do not regularly fall into patterns that can be recognized as easily as the *Bush* case.

For Experiment 3, two texts each of ca 3000 words (3662 and 3607 lines) were constructed by taking respectively the fifth (text a) and sixth (text b) 333-word chunk of each of the nine BNC extracts. In the first part of the experiment, text (a) was tagged by Annotator 1 using dictionary look-up and was postedited by Annotator 2 following automatic tagging. In the second part of the experiment Annotator 1's and 2's roles were reversed in tagging text (b). Table 4 shows inter-annotator agreement levels for Experiment 3.

Since the automatic tagger did not supply fine-grained classifications (e.g. transitivity or countability classes) and the postediting annotator did not attempt to modify this aspect of its output, discrepancies at the "fine" level were inevitably very numerous; these figures are not included in Table 4. The first row of the table shows discrepancies at the "coarse" level of comparison. The figures of 96.3% and 95.4% are below the corresponding figure of 98.0% in Table 1. This might suggest that the automatic tagger was indeed creating an undue bias in favour of accepting its output.

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

Table 4. *Inter-annotator disagreements using different annotation methods (alternative software tools)*

	Number of discrepancies		% agreement	
	Expt (a)	Expt (b)	Expt (a)	Expt (b)
raw discrepancies	135	165	96.3	95.4
systematic differences discounted	69	97	98.1	97.3

Table 5. *Inter-annotator speed differences by annotation method*

	Mean tagging time (minutes/3000 words)		
	manual	dict. look-up	post-editing
Annotator 1	315	199	90
Annotator 2	325	100	109

However, examination of specific disagreements suggested that this was not a generalized bias, but related specifically to the systematic differences mentioned above between the electronic dictionaries incorporated in the dictionary look-up and automatic tagging software systems, such as the fact that the automatic-tagger dictionary allows words like *disturbing* to be coded as adjectives while the SUSANNE scheme requires them to be coded as present participles even when functioning adjectivally. This systematic difference is arguably more like a clash between two slightly different schemes of tagging than like the unpredictable errors in automatic tagger output that arise when the automatic-tagger dictionary contains the same candidate tags for a wordform as the dictionary look-up dictionary but the tagger picks the wrong one.

The second row of Table 4 shows the agreement levels achieved, if discrepancies relating to these systematic differences are not counted as errors. Now the figures of 98.1% and 97.3% are about the same as the 98.0% in Table 1.

It seems therefore that human annotators are well able to resist bias in favour of accepting an automatic tagger's erroneous output, at least provided this relates to mistaken choices among the same candidate tags as defined by the scheme which the annotator is seeking to apply.

6 Speed differences

A further question about alternative tools to support tagging is how far they increase annotator efficiency. Since we have seen that accuracy does not seem to be heavily affected by choice of system, it will be sufficient to compare speeds with the different techniques. Table 5 gives tagging times per 3000 words of text for our two annotators using the three techniques. (The dictionary look-up technique was used in all three experiments, so the figures in that column are averages.)

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

It is clear from Table 5 in the first place that both automatic support tools give large advantages in speed over purely manual tagging. This is unsurprising: a tagging scheme as refined as the SUSANNE scheme contains far too much detail for any human to hold in his or her head, so manual tagging requires a great deal of time-consuming consultation of paper documents.

More remarkable are the differences between annotators in the two rightmost columns. For one annotator, it emerges, the two support tools allow similar speeds, both much faster than manual tagging, and dictionary look-up is slightly faster than post-editing. For the other annotator, post-editing is far faster than dictionary look-up, which for that annotator is only a halfway house in efficiency terms between post-editing automatic tagger output and pure manual tagging.

(A possible explanation for the difference is that dictionary look-up, which requires a distinctive keyboard input for almost every word, involves a greater challenge than postediting to annotators' manual dexterity, which varies from person to person. But speculation about explanations is less interesting than the simple existence of such large personal differences with respect to the "habitability" of alternative annotation support tools.)

Before doing these experiments, we might have predicted that the important differences between different tagging support tools would be the different levels of accuracy promoted by tools of different types. So far as our investigation has gone, we have found only minor differences in that respect. But it turns out that, even as between annotators having similar levels of experience with the specific annotation task and conventions, there are large individual differences in the extent to which a particular type of support tool suits an individual's work style.

We are not aware of any earlier research directed at this issue, which seems potentially quite important for the activity of linguistic resource compilation.

7 Conclusion

The foregoing has explored a number of issues relating to ceilings on annotation performance, looking at the wordtagging domain. In future work we intend to extend the investigation to the domain of higher-level phrase and clause annotation.

References

Blaheta, D. (2002) Handling noisy training and testing data. *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing*, Philadelphia, pp. 111–16; faculty.knox.edu/dblaheta/papers/dpb-emnlp02.pdf

Brants, T. (2000) Inter-annotator agreement for a German newspaper corpus. *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC-2000*, Athens; www.coli.uni-sb.de/~thorsten/publications/Brants-LREC00.pdf

Carletta, J. (1996) Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* **22**: 249–254. Reprinted in G. R. Sampson and D. McCarthy, editors, *Corpus Linguistics: Readings in a widening discipline*, Continuum, 2004, pp. 335–339.

Church, K. W. (1992) Current practice in part of speech tagging and suggestions for the future. In: C. F. Simmons, editor, *Sborník Práci: in honor of Henry Kučera*, Michigan Slavonic Studies, pp. 13–48; www.research.att.com/~kwc/kucera.text.ps

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press

Dickinson, M. and Meurers, W. D. (2003) Detecting errors in part-of-speech annotation. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, Budapest, pp. 107–14; www.ling.ohio-state.edu/~dm/papers/dickinson-meurers-03.pdf

Elworthy, D. (1994) Does Baum-Welch re-estimation help taggers? *Proceedings of the 4th Association for Computational Linguistics Conference on Applied Natural Language Processing*, Trento, pp. 53–58.

Hornby, A. S. (editor) (1974) *Oxford Advanced Learner's Dictionary of Current English* (3rd ed). Oxford University Press.

Langendoen, D. T. (1997) Review of Sampson (1995). *Language* **73**: 600–603.

Leech, G. N. (1992) 100 million words of English: the British National Corpus. *Language Research* **28**: 1–13.

Leech, G. N., Garside, R. G. and Bryant, M. (1994) CLAWS4: the tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, Kyoto, pp. 622–628.

Lin, D. (2003) Dependency-based evaluation of Minipar. In: A. Abeillé (editor), *Treebanks: Building and Using Parsed Corpora*, Kluwer, pp. 317–329.

Oliva, K. (2001) The possibility of automatic detection/correction of errors in tagged corpora: a pilot study on a German corpus. In: V. Matoušek, P. Mautner, R. Moucek and K. Taušer (editors), *Proceedings 4th International Conference on Text, Speech and Dialogue (TSD 2001: Lecture Notes in Computer Science*, vol. 2166, Železná Ruda. Springer, pp. 39–46.

Sampson, G. R. (1995) *English for the Computer: The SUSANNE Corpus and analytic scheme*. Clarendon Press.

Sampson, G. R. (2000) CHRISTINE Corpus, Stage I: Documentation. www.grsampson.net/ChrisDoc.html

Tateisi, Y. and Tsujii, J.-I. (2004) Part-of-speech annotation of biology research abstracts. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, pp. 1267–1270; www-tsujii.is.s.u-tokyo.ac.jp/%7Eyucca/papers/LREC2004-528.pdf

Voutilainen, A. (1999) An experiment on the upper bound of interjudge agreement: the case of tagging. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, pp. 204–208; acl.ldc.upenn.edu/E/E99/E99-1027.pdf

https://doi.org/10.1017/S1351324905003803 Published online by Cambridge University Press