An experiment on the upper bound of interjudge agreement: the case of tagging

Atro Voutilainen
Research Unit for Multilingual Language Technology
P.O. Box 4
FIN-00014 University of Helsinki
Finland
Atro.Voutilainen@ling.Helsinki.FI

Abstract

We investigate the controversial issue about the upper bound of interjudge agreement in the use of a low-level grammatical representation. Pessimistic views suggest that several percent of words in running text are undecidable in terms of part-of-speech categories. Our experiments with 55kW data give reason for optimism: linguists with only 30 hours' training apply the EngCG-2 morphological tags with almost 100% interjudge agreement.

1 Orientation

Linguistic analysers are developed for assigning linguistic descriptions to linguistic utterances. Linguistic descriptions are based on a fixed inventory of descriptors plus their usage principles: in short, a *grammatical representation* specified by linguists for the specific kind of analysis – e.g. morphological analysis, tagging, syntax, discourse structure – that the program should perform.

Because automatic linguistic analysis generally is a very difficult problem, various methods for evaluating their success have been used. One such is based on the degree of correctness of the analysis provided, e.g. the percentage of linguistic tokens in the text analysed that receives the appropriate description relative to analyses provided independently of the program by competent linguists ideally not involved in the development of the analyser itself.

Now use of benchmark corpora like this turns out to be problematic because arguments have been made to the effect that linguists themselves make erroneous and inconsistent analyses. Unintentional mistakes due e.g. to slips of attention are obviously unavoidable, but these errors can largely be identified by the double-blind method: first by having two (or more) linguists analyse the same text independently by using the same grammatical representation, and then identifying differences of analysis by automatically comparing the analysed text versions with each other and finally having the linguists discuss the differences and modify the resulting benchmark corpus accordingly. Clerical errors should be easily (i.e. consensually) identified as such, but, perhaps surprisingly, many attested differences do not belong to this category. Opinions may genuinely differ about which of the competing analyses is the correct one, i.e. sometimes the grammatical representation is used inconsistently. In short, linguistic 'truth' seems to be uncertain in many cases. Evaluating – or even developing – linguistic analysers seems to be on uncertain ground if the goal of these analysers cannot be satisfactorily specified.

Arguments concerning the magnitude of this problem have been made especially in relation to *tagging*, the attempt to automatically assign lexically and contextually correct morphological descriptors (tags) to words. A pessimistic view is taken by Church (1992) who argues that even after negotiations of the kind described above, no consensus can be reached about the correct analysis of several percent of all word tokens in the text. A more mixed view on the matter is taken by Marcus et al. (1993) who on the one hand note that in one experiment moderately trained human text annotators made different analyses even after negotiations in over 3% of all words, and on the other hand argue that an expert can do much better.

An optimistic view on the matter has been presented by Eyes and Leech (1993). Empirical evidence for a high agreement rate is reported by Voutilainen and Järvinen (1995). Their results suggest that at least with one grammatical representation, namely the ENGCG tag set (cf. Karlsson et al., eds., 1995), a 100% consistency can be reached after negotiations at the level of parts of speech (or morphology in this case). In short, reasonable evidence has been given for the position that at least some tag sets can be applied consistently, i.e. earlier observations about potentially more problematic tag sets should not be taken as predictions about all tag sets.

1.1 Open questions

Admittedly Voutilainen and Järvinen's experiment provides evidence for the possibility that two highly experienced linguists, one of them a developer of the ENGCG tag set, can apply the tag set consistently, at least when compared with each others' performance. However, the practical significance of their result seems questionable, for two reasons.

Firstly, large-scale corpus annotation by hand is generally a work that is carried out by less experienced linguists, quite typically advanced students hired as project workers. Voutilainen and Järvinen's experiment leaves open the question, how consistently the ENGCG tag set can be applied by a less experienced annotator.

Secondly, consider the question of tagger evaluation. Because tagger developers presumably tend to learn, perhaps partly subconsciously, much about the behaviour, desired or otherwise, of the tagger, it may well be that if the developers also annotate the benchmark corpus used for evaluating the tagger, some of the tagger's misanalyses remain undetected because the tagger developers, due to their subconscious mimicking of their tagger, make the same misanalyses when annotating the benchmark corpus. So 100% tagging consistency in the benchmark corpus alone does not necessarily suffice for getting an objective view of the tagger's performance. Subconscious 'bad' habits of this type need to be factored out. One way to do this is having the benchmark corpus consistently (i.e. with approximately 100% consensus about the correct analysis) analysed by people with no familiarity with the tagger's behaviour in different situations – provided this is possible in the first place.

Another two minor questions left open by Voutilainen and Järvinen concern the (i) typology of the differences and (ii) the reliability of their experiment.

Concerning the typology of the differences: in Voutilainen and Järvinen's experiment the linguists negotiated about an initial difference, almost one per cent of all words in the texts. Though they finally agreed about the correct analysis in almost all these differences, with a slight improvement in the experimental setting a clear categorisation of the initial differences into unintentional mistakes and other, more interesting types, could have been made.

Secondly, the texts used in Voutilainen and Järvinen's experiment comprised only about 6,000 words. This is probably enough to give a general indication of the nature of the analysis task with the ENGCG tag set, but a larger data would increase the reliability of the experiment.

In this paper, we address all these three questions. Two young linguists[1] with no background in ENGCG tagging were hired for making an elaborated version of the Voutilainen and Järvinen experiment with a considerably larger corpus.

The rest of this paper is structured as follows. Next, the ENGCG tag set is described in outline. Then the training of the new linguists is described, as well as the test data and experimental setting. Finally, the results are presented.

2 ENGCG tag set

Descriptions of the morphological tags used by the English Constraint Grammar tagger are available in several publications. Brief descriptions can be found in several recent ACL conference proceedings by Voutilainen and his colleagues (e.g. EACL93, ANLP94, EACL95, ANLP97, ACL-EACL97). An in-depth description is given in Karlsson et al., eds., 1995 (chapters 3-6). Here, only a brief sample is given.

ENGCG tagging is a two-phase process. First, a lexical analyser assigns one or more alternative analyses to each word. The following is a morphological analysis of the sentence *The raids were coordinated under a recently expanded federal program:*

```
"<The>"
        "the" <Def> DET CENTRAL ART SG/PL
"<raids>"
        "raid" <Count> N NOM PL
        "raid" <SVO> V PRES SG3
"<were>"
        "be" <SVC/A> <SVC/N> V PAST
"<coordinated>"
        "coordinate" <SVO> EN
        "coordinate" <SVO> V PAST
"<under>"
        "under" ADV ADVL
        "under" PREP
        "under" <Attr> A ABS
"<a>"
        "a" ABBR NOM SG
        "a" <Indef> DET CENTRAL ART SG
```

[1] Ms. Pirkko Paljakka and Mr. Markku Lappalainen

```
"<recently>"
        "recent" <DER:ly> ADV
"<expanded>"
        "expand" <SVO> <P/on> EN
        "expand" <SVO> <P/on> V PAST
"<federal>"
        "federal" A ABS
"<program>"
        "program" N NOM SG
        "program" <SVO> V PRES -SG3
        "program" <SVO> V INF
        "program" <SVO> V IMP
        "program" <SVO> V SUBJUNCTIVE
"<.>"
```

Each indented line constitutes one morphological analysis. Thus *program* is five-ways ambiguous after ENGCG morphology. The disambiguation part of the ENGCG tagger[2] then removes those alternative analyses that are contextually illegitimate according to the tagger's hand-coded constraint rules (Voutilainen 1995). The remaining analyses constitute the output of the tagger, in this case:

```
"<The>"
        "the" <Def> DET CENTRAL ART SG/PL
"<raids>"
        "raid" <Count> N NOM PL
"<were>"
        "be" <SVC/A> <SVC/N> V PAST
"<coordinated>"
        "coordinate" <SVO> EN
"<under>"
        "under" PREP
"<a>"
        "a" <Indef> DET CENTRAL ART SG
"<recently>"
        "recent" <DER:ly> ADV
"<expanded>"
        "expand" <SVO> <P/on> EN
"<federal>"
        "federal" A ABS
"<program>"
        "program" N NOM SG
"<.>"
```

Overall, this tag set represents about 180 different analyses when certain optional auxiliary tags (e.g. verb subcategorisation tags) are ignored.

3 Preparations for the experiment

3.1 Experimental setting

The experiment was conducted as follows.

1. The text was morphologically analysed using the ENGCG morphological analyser. For the analysis of unrecognised words, we used a rule-based heuristic component that assigns morphological analyses, one or more, to each word not represented in the lexicon of the system. Of the analysed text, two identical versions were made, one for each linguist.

2. Two linguists trained to disambiguate the ENGCG morphological representation (see the subsection on training below) independently marked the correct alternative analyses in the ambiguous input, using mainly structural, but in some structurally unresolvable cases also higher-level, information. The corpora consisted of continuous text rather than isolated sentences; this made the use of textual knowledge possible in the selection of the correct alternative. In the rare cases where two analyses were regarded as equally legitimate, both could be marked. The judges were encouraged to consult the documentation of the grammatical representation. In addition, both linguists were provided with a checking program to be used after the text was analysed. The program identifies words left without an analysis, in which case the linguist was to provide the missing analysis.

3. These analysed versions of the same text were compared to each other using the Unix sdiff program. For each corpus version, words with a different analysis were marked with a "RECONSIDER" symbol. The "RECONSIDER" symbol was also added to a number of other ambiguous words in the corpus. These additional words were marked in order to 'force' each linguist to think independently about the correct analysis, i.e. to prevent the emergence of the situation where one linguist considers the other to be always right (or wrong) and so 'reconsiders' only in terms of the existing analysis. The linguists were told that some of the words marked with the "RECONSIDER" symbol were analysed differently by them.

4. Statistics were generated about the number of differing analyses (number of "RECONSIDER" symbols) in the corpus versions ("diffl" in the following table).

5. The reanalysed versions were automatically compared to each other. To words with a different analysis, a "NEGOTIATE" symbol was added.

[2] A new version of the tagger, known as EngCG-2, can be studied and tested at http://www.conexor.fi.

6. Statistics were generated about the number of differing analyses (number of "NEGOTIATE" symbols) in the corpus versions ("diff2" in the following table).

7. The remaining differences in the analyses were jointly examined by the linguists in order to see whether they were due to (i) inattention on the part of one linguist (as a result of which a correct unique analysis was jointly agreed upon), (ii) joint uncertainty about the correct analysis (both linguists feel unsure about the correct analysis), or (iii) conflicting opinions about the correct analysis (both linguists have a strong but different opinion about the correct analysis).

8. Statistics were generated about the number of conflicting opinions ("diff3" below) and joint uncertainty ("unsure" below).

This routine was successively applied to each text.

3.2 Training

Two people were hired for the experiment. One had recently completed a Master's degree from English Philology. The other was an advanced undergraduate student majoring in English Philology. Neither of them were familiar with the ENGCG tagger.

All available documentation about the linguistic representation used by ENGCG was made available to them. The chief source was chapters 3-6 in Karlsson et al. (eds., 1995). Because the linguistic solutions in ENGCG are largely based on the comprehensive descriptive grammar by Quirk et al. (1985), also that work was made available to them, as well as a number of modern English dictionaries.

The training was based on the disambiguation of ten smallish text extracts. Each of the extracts was first analysed by the ENGCG morphological analyser, and then each trainee was to independently perform Step 3 (see the previous subsection) on it. The disambiguated text was then automatically compared to another version of the same extract that was disambiguated by an expert on ENGCG. The ENGCG expert then discussed the analytic differences with the trainee who had also disambiguated the text and explained why the expert's analysis was correct (almost always by identifying a relevant section in the available ENGCG documentation; in very rare cases where the documentation was underspecific, new documentation was created for future use in the experiments).

After analysis and subsequent consultation with the ENGCG expert, the trainee processed the following sample.

The training lasted about 30 hours. It was concluded by familiarising the linguists with the routine used in the experiment.

3.3 Test corpus

Four texts were used in the experiment, totalling 55724 words and 102527 morphological analyses (an average of 1.84 analyses per word). One was an article about Japanese culture ('Pop'); one concerned patents ('Pat'); one contained excerpts from the law of California; one was a medical text ('Med'). None of them had been used in the development of the ENGCG grammatical representation or other parts of the system. By mid-June 1999, a sample of this data will be available for inspection at http://www.ling.helsinki.fi/ voutilai/eacl99-data.html.

4 Results and discussion

The following table presents the main findings.

Figure 1: .Results from a human annotation task.

	words	*diff1*	*diff2*	*diff3*	*unsure*
Pop	14861	188/1.3%	44/.3%	0	4/.0%
Pat	13183	92/.7%	11/.1%	2/.0%	1/.0%
Law	15495	107/.7%	18/.1%	10/.1%	0
Med	12185	126/1.0%	39/.3%	1/.0%	9/.1%
ALL	55724	513/.9%	112/.2%	13/.0%	14/.0%

It is interesting to note how high the agreement between the linguists is even before the first negotiations (99.80% of all words are analysed identically). Of the remaining differences, most, somewhat disappointingly, turned out to be classified as 'slips of attention'; upon inspection they seemed to contain little linguistic interest. Especially one of the linguists admitted that most of the job seemed too much of a routine to keep one mentally alert enough. The number of genuine conflicts of opinion were much in line with observations by Voutilainen and Järvinen. However, the negotiations were not altogether easy, considering that in all they took almost nine hours. Presumably uncertain analyses and conflicts of opinion were not easily passed by.

The main finding of this experiment is that basically Voutilainen and Järvinen's observations about the high specifiability and consistent usability of the ENGCG morphological tag set seem to be extendable to new users of the tag set. In other words, the reputedly surface-syntactic tag set seems to be learnable as well. Overall, the experiment reported here provides evidence for the optimistic position about the specifiability of at least certain kinds of linguistic representations.

It remains for future research, perhaps as a collaboration between teams working with different tag sets, to find out, what exactly are the properties that make some linguistic representations consistently learnable and usable, and others less so.

Acknowledgments

I am grateful to anonymous EACL99 referees for useful comments.

References

Kenneth W. Church 1992. Current Practice in Part of Speech Tagging and Suggestions for the Future. In Simmons (ed.), *Sbornik praci: In Honor of Henry Kucera*, Michigan Slavic Studies. Michigan. 13-48.

Elizabeth Eyes and Geoffrey Leech 1993. Syntactic Annotation: Linguistic Aspects of Grammatical Tagging and Skeleton Parsing. In Ezra Black, Roger Garside and Geoffrey Leech (eds.) 1993. *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach.* Amsterdam and Atlanta: Rodopi. 36-61.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila (eds.) 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Berlin and New York: Mouton de Gruyter.

Mitchell Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19:2. 313-330.

Randolph Quirk, Sidney Greenbaum, Jan Svartvik and Geoffrey Leech 1985. *A Comprehensive Grammar of the English Language.* Longman.

Atro Voutilainen 1995. Morphological disambiguation. In Karlsson et al., eds.

Atro Voutilainen and Timo Järvinen 1995. Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics.* ACL.