COMPUTER-AIDED GRAMMATICAL TAGGING OF SPOKEN ENGLISH

Jan Svartvik, University of Lund

Department of English
Helgonabacken 14
S-223 62 Lund, Sweden

Abstract

The paper presents an outline of a system for grammatical tagging of the London-Lund Corpus of spoken English consisting of some 450 000 words. The material, all of which will be available on magnetic computer tape, and part of which is now available in both machine-readable and printed form, has been transcribed orthographically with prosodic marking for tone units, nuclei, stresses, pauses, etc (see Samples 1 and 2). Whereas there is now considerable agreement on the usefulness of a tagged corpus, there is as yet no consensus on the best type of tagging, let alone the procedure involved. The analysis proposed here is of course specifically aimed at tagging spoken English, but should be largely applicable also to written English.

The syntactic tagging will initially be based on surface properties, since we are interested in gaining information that is directly available through the signals that hearers use for decoding a message, i e their perceptual strategies. In this respect, the plan is no innovation. One computer discourse model which is intended "to tackle problems that a speaker evidently tackles" has recently been reported by Davey (1978.4). His model, however, is designed to produce, not understand. Another and more important difference between the SSE system and the Davey model and most other computer discourse models is that the latter have been devised to handle restricted and artificial universes of discourse, such as describing games or moving blocks. However, the work of Winograd (1972), for example, is directly relevant to our task, since it deals with wider aspects of language and makes impressive use of Halliday's systemic grammar for producing parsing algorithms.

One of our aims is to make the tagging procedure as automatic as possible. Specifically, we would like to see how far it is possible to carry out syntactic analysis based on graphic words and prosody (provided by the material) and word class tags (provided by a general-purpose dictionary). Given that no fully automatic system for grammatical tagging exists, we propose to implement an interactive, semi-manual mode of analysis.

The paper will present word class tagging of types from the Longman Dictionary of Contemporary English, disambiguation of tokens and phrase tagging by means of a set of parsing algorithms. The basic unit of analysis will be the tone unit. In a previous study of Survey material of spoken English, it was found that the overall average length of a tone unit was 5.3 words and that "there was considerable correlation between the length of tone units and their grammatical contents" with a "high degree of co-extensiveness between tone units and grammatical units of group, phrase, and clause structure" (Quirk et al 1964).

The search for grammatical phrases will be from right to left within the tone unit. Since this search sequence is definitely unorthodox, some explanation may be called for. By and large, English phrase structure typically has the head to the right, as in

Verb phrases:	will be DOING
Noun phrases:	the nice little DOG
Adjective phrases:	stunningly BEAUTIFUL

Assuming that a good number of the tone units consist of, at least, grammatical phrases, the nucleus will occur within the phrase and, more often than not, within the head of the phrase. Thus, it is likely that it will be linguistically rewarding as well as computationally economical to search from right to left. It seems that a left-to-right search method also runs into difficulties with solving left-recursion structures and predicting numerous alternatives.

The phrase recognition rules are to be applied in the following order:

(VPH)	Verb phrases
(APH)	Adverb phrases
(JPH)	Adjective phrases
(NPH)	Noun phrases
(PPH)	Prepositional phrases

The typical features of this system are: taking tone units as the basis of grammatical analysis, choosing a general-purpose dictionary for word class tagging, making extensive use of phrase structure rules which are applied in a certain order and cyclically, and partly adopting an interactive mode of analysis.

Sample 1. Computer version of Text S.1.1: TUs 71-102.

```
 81: 1 1 5  710 1 2 A  11  [ə:m] - - De^larey`s the Ca:n\adian . st/udent      /
 82: 1 1 5  710 1 1 A  11  (re^m/ember#)#                                      /
 83: 1 1 5  720 1 1 A  11  ^last y/ear#                                        /
 84: 1 1 5  730 1 1 B  11  ^[mh/m]#                                            /
 85: 1 1 5  740 1 1 A  11  [ə:] he ^should have had his . dissertation \/in#  /
 86: 1 1 5  750 1 1 A  11  ((at the)) be^ginning of M\/ay# .                   /
 87: 1 1 5  760 1 1 A  11  ((but)) the ^damn thing ((hasn`t)) c/ome# -        /
 88: 1 1 5  770 1 1 A  11  [ə:] I ^did get a !p\ostcard fr/om him# - -        /
 89: 1 1 5  780 1 1 A  11  ^saying that [ə:m] the !thing is now :r/eady# .    /
 90: 1 1 5  790 1 2 A  11  and that he will ^send it by the :end . of         /
 91: 1 1 5  790 1 1 A  11  :J\une# .                                          /
 92: 1 1 5  800 1 1 A  11  ^that`s what he !s\ays# .                          /
 93: 1 1 6  810 1 2 A  11  ^now . !A he may not . send it . quite as soon as ./
 94: 1 1 6  810 1 1 A  11  :th/at#                                            /
 95: 1 1 6  820 1 1 A  11  and ^=ə#                                           /
 96: 1 1 6  830 1 1 A  11  it ^may take a hell of a long time to !d\one# .    /
 97: 1 1 6  840 1 1 A  11  ^it he !puts it into the :diplomatic b\ag#         /
 98: 1 1 6  850 1 1 A  11  ^as [ə:m] - !wh\at`s his _name# .                  /
 99: 1 1 6  860 1 1 A  11  Mickey ^O\/ohr _did# .                             /
100: 1 1 6  870 1 1 A  11  ^then ((it`s)) not so b\ad# -                      /
101: 1 1 6  880 1 1 A  11  ^but [ə:] !how are y\ou going to be pl/aced#      /
102: 1 1 6  890 1 1 A  11  ^for *((!h\aving#))*                               /
103: 1 1 6  900 1 2 B  11  *[ə:] -* I ^wouldn`t want it before the :end of    /
104: 1 1 6  900 1 1 B  11  June :\anyhow R/eynard#                            /
105: 1 1 6  910 1 1 B  11  be^cause I`m !going to Madr\id# .                  /
106: 1 1 6  920 1 1 B  11  on the ^t\/enth#                                   /
107: 1 1 6  930 1 1 B  11  and ^coming back on the twenty-n/inth# -          /
108: 1 1 6  940 2 1 B  21  *[ə:]* . I +shall+ "^not                           /
109: 1 1 6  950 1 1 B  11  *^I s/ee#*                                         /
110: 1 1 6  960 1 1 A  11  +^y/es#+                                           /
111: 1 1 6  940 1 1 B  11  b=e#                                               /
112: 1 1 6  970 1 1 B  11  a^way f/om home :th\/en#                           /
113: 1 1 6  980 1 1 B  11  un^t\/il#                                          /
114: 1 1 6  990 1 1 B  11  at ^\any rate#                                     /
115: 1 1 7 1000 1 1 B  12  the ^\end of *-* a^bout the end of \August# - -   /
116: 1 1 7 1010 1 1 A  11  *^[=m]#*                                           /
117: 1 1 7 1020 1 1 A  20  [ə:]                                               /
```

Sample 2. Printed version of Text S.1.1: TUs 71-136.

A [71] [ə:m] - - De‖laney's the CA▵NÀDIAN · STÚDENT {RE‖MÉMBER■} ■ [72] ‖last YÉAR■

B [73] ‖[mhḿ]■

A [74] [ə:] he ‖should have had▁his · dissertation ŶN■ [75] «at the» be‖ginning of MÁŶ■ · [76] «but» the ‖damn thing «hasn't» CÓME■ - [77] [ə:] I ‖did get a ΔPÒSTCARD FRÓM him■ - - [78] ‖saying that [ə:m] the Δthing is now ▵RÉADY■ · [79] and that he will ‖send it by the ▵end · of ▵JÙNE■ · [80] ‖that's what he ΔSÀYS■ · [81] ‖now · ΔA he may not · send it · quite as soon as · ▵THÁT■ [82] and ‖ū■ [83] it ‖may take a hell of a long time to ΔCÒME■ · [84] ‖if he Δputs it into the ▵diplomatic BAG■ [85] ‖as [ə:m] - ΔWHÀT'S his ▻name■ · [86] Mickey ‖CÓHN ▻did■ · [87] ‖then «it's» not so BÀD■ - [88] ‖but [ə:] Δhow are YÒU going to be PLÁCED■ [89] ‖for ☆«ΔHÀVING■»☆

B [90] ☆[ə:] -☆ I ‖wouldn't want it before the ▵end of June ▵ÀNYHOW RÉYNARD■ [91] be‖cause I'm Δgoing to MADRÌD■ · [92] on the ‖TÉNTH■ [93] and ‖coming back on the TWENTY-NÍNTH■ - [94] ☆[ə:]☆ · I +shall+ 'not

A [95] ☆‖I SÉE■☆ [96] +‖YÉS■+

>B [94] BÉ■ [97] al‖way from home ▵THÈN■ [98] UN‖TÍL■ [99] at ‖ÀNY rate■ [100] the ‖ÈND of ☆-☆ a‖bout the end of ÀUGUST■ - -

A [101] ☆‖[m̀]■☆ [102] [ə:]

B [103] so ☆‖any time in JULÝ■ [104] ‖and☆ ÀUGUST■ [105] ‖but [ə:] + · +

A [106] ☆(- - a hiss-whistle)☆ +‖YÉS■+

>B [105] Δnot too 'far into 'August if ☆PÒSSIBLE■☆ - [107] ‖ÓTHERWISE■ [108] I'll be ‖stuck until about [ðə́:]

A [109] ☆‖NÓ■☆

>B [108] Δtwentieth · [ə] I'm ‖HÓPING■ [110] to ‖get into SPÂIN■ · [111] from a‖bout the Δtwenty- · ▵ÈIGHTH of ÁUGUST■ [112] «to» un‖til about the Δtwentieth or ▵something of that kind of SEPTÈMBER■ ☆·☆ [113] but

A [114] ☆‖YÈAH■☆

>B [113] ‖[ʌðəw] a‖part from ▵THÀT■ · [115] I'll be at ‖HÔME■ [116] and al‖though I'll be doing CSC ▻stuff■ [117] and ‖that kind of THÍNG■ [118] ‖I can always 'put it on one ☆SÍDE■☆ [119] and ‖get on with the PÀPER■

A [120] ☆‖YÈAH■☆ [121] [ə:] you ‖see the ΔÓTHER ▻man■ [122] ‖CHÒMLEY■ [123] ‖ought · ‖ought · ‖ought ΔÀLSO■ [124] to have · ‖got his in on TÍME■ [125] and I SUS‖PÈCTED■ [126] ‖ÀLWAYS■ [127] that De‖laney would be LÁTE■ · [128] that ‖Chomley would be on TÍME■ [129] and that ‖this would · produce a nice 'ΔSTÀGGERING■ [130] of · of their ar‖rival on your ΔDÈSK■ ☆-☆ [131] [ə:m] ‖now it looks as if they they both

B [132] ☆[m]‖[hm̀]■☆

>A [131] AR▵RÌVE■ [133] [ə] I ‖think that we ▵mustn't worry too ▵much A▵BÒUT THÍS■ [134] ‖we we ‖make it Δperfectly ▵clear that ▵papers must be in on the ▵first of ▵MÁŶ■ ☆-☆ [135] [ə:m]

B [136] ☆[m]‖[hm̀]■☆