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.

11	1	710	1	2	A	11	[s:t] - - De'larey's the Can\adian . st/ucnt
11	1	710	1	1	A	11	(re_m/embe#)#
11	1	720	1	1	A	11	-last y/ear#
11	1	730	1	1	B	11	~[mh/m]#
11	1	740	1	1	A	11	[s:] he 'should have had his . dissertatior \iir#
11	1	750	1	1	A	11	((at the)) be'ginnin of M\ay# .
11	1	760	1	1	A	11	((but)) the 'cann thing ((hasr't)) c/cne# -
11	1	770	1	1	A	11	[s:] I 'did get a \p\ostcard fr/om him# -
11	1	780	1	1	A	11	'saying that [s:m] the !thing is now :r/eeey# .
11	1	790	1	2	A	11	and that he will 'send it by the :end . of
11	1	790	1	1	A	11	:\unc# .
11	1	800	1	1	A	11	'that's what he !slays# .
11	1	810	1	2	A	11	'now . !A he may not . sero it . quite as socr as .
11	1	810	1	1	A	11	:th/at#
11	1	820	1	1	A	11	and '==#
11	1	830	1	1	A	11	it 'may take a hell of a long time to !c/cne# .
11	1	840	1	1	A	11	'if he !puts it into the :diplomatic t\ag#
11	1	850	1	1	A	11	'as [s:m] - !h\at's his _name# .
11	1	860	1	1	A	11	Mickey 'C\chr _cio# .
11	1	870	1	1	A	11	'then ((it's)) rct so l\ach -
11	1	880	1	1	A	11	'but [s:] !how are y\cu going to be pl/acec#
11	1	890	1	1	A	11	'for *((ih\aving#))*
11	1	900	1	2	E	11	*[s:] -- I 'couldn't want it before the :erc of
11	1	900	1	1	E	11	June :\anyho# k/eynard#
11	1	910	1	1	E	11	be'cause I'm !ying to Macr\ich# .
11	1	920	1	1	E	11	on the 't\/ent#
11	1	930	1	1	E	11	and 'coming back on the twenty-n/inth# -
21	1	940	2	1	E	11	*[s:]* . I +shall+ "not
11	1	950	1	1	A	11	*I s/ee#
11	1	960	1	1	A	11	+y/es#
11	1	940	1	1	E	11	be#
11	1	970	1	1	E	11	a 'way fr/om here :th\en#
11	1	980	1	1	E	11	ur't\il#
11	1	990	1	1	E	11	at 'any rate
12	1	1000	1	1	E	11	the 'lend of +-- a'tout the erc of \August# - -
11	1	1010	1	1	A	11	*[s:]#
20	1	1020	1	1	A	11	[s:]

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

- A 71 [ə:m] - - Dellaney's the CÁNADIAN · STÚDENT {RE}MÉMBER■ 72 ||last
YÉAR■
- B 73 ||[mh̃m]■
- A 74 [ə:] he ||should have had his · dissertation YN■ 75 «at the» be||ginning of
MÁY■ · 76 «but» the ||damn thing «hasn't» CÔME■ - 77 [ə:] I ||did get a
ΔPÔSTCARD FRÔM him■ - - 78 ||saying that [ə:m] the Δthing is now ΔRÉADY■
· 79 and that he will ||send it by the Δend · of ΔJÛNE■ · 80 ||that's what he
ΔSAYS■ · 81 ||now · ΔA he may not · send it · quite as soon as · ΔTHÁT■
82 and ||[h̃]■ 83 it ||may take a hell of a long time to ΔCÔME■ · 84 ||if he Δputs
it into the Δdiplomatic BAG■ 85 ||as [ə:m] - ΔWHÁT'S his Δ-name■ ·
86 Mickey ||CÔHN Δdid■ · 87 ||then «it's» not so BÂD■ - 88 ||but [ə:] Δhow
are YÔU going to be PLÁCED■ 89 ||for ★«ΔHÁVING»★
- B 90 ★[ə:] - ★ I ||wouldn't want it before the Δend of June ΔÁNYHOW RÉYNARD■
91 be||cause I'm Δgoing to MADRID■ · 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 BÊ■ 97 a||way from home ΔTHÉN■ 98 UN||TÍL■ 99 at ||ÁNY rate■ 100 the
||ÈND of ★-★ a||bout the end of ÁUGUST■ - -
- A 101 ★||[m̃]■★ 102 [ə:]
- B 103 so ★||any time in JULÝ■ 104 ||and★ ÁUGUST■ 105 ||but [ə:] + · +
- A 106 ★(- - a hiss-whistle)★ +||YÉS■+
- > B 105 Δnot too 'far into 'August if ★PÔSSIBLE■★ - 107 ||ÔTHERWISE■ 108 I'll be
||stuck until about [d̃i:]
- A 109 ★||NÔ■★
- > B 108 Δtwentieth · [ə:] I'm ||HÓPING■ 110 to ||get into SPÁIN■ · 111 from a||bout
the Δtwenty- · ΔEIGHTH of ÁUGUST■ 112 «to» un||til about the Δtwentieth or
Δsomething of that kind of SEPTÈMBER■ ★ · ★ 113 but
- A 114 ★||YÉAH■★
- > B 113 ||[Δəw] a||part from ΔTHÁT■ · 115 I'll be at ||HÔME■ 116 and a||though
I'll be doing CSC Δstuff■ 117 and ||that kind of THING■ 118 ||I can always
'put it on one ★SIDE■★ 119 and ||get on with the PÁPER■
- A 120 ★||YÉAH■★ 121 [ə:] you ||see the ΔÔTHER Δ-man■ 122 ||CHÔMLEY■
123 ||ought · ||ought · ||ought ΔÁLSO■ 124 to have · ||got his in on TÍME■
125 and I SUS||PÈCTED■ 126 ||ÁLWAYS■ 127 that Dellaney would be LÁTE■ ·
128 that ||Chomley would be on TÍME■ 129 and that ||this would · produce a
nice ΔSTÁGGERING■ 130 of · of their arrival on your ΔDÈSK■ ★-★ 131 [ə:m]
||now it looks as if they they both
- B 132 ★[m]||[hm̃]■★
- > A 131 AR·RÍVE■ 133 [ə:] I ||think that we Δmustn't worry too Δmuch ΔBÔUT THÍS■
134 ||we we ||make it Δperfectly clear that Δpapers must be in on the Δfirst of
ΔMÁY■ ★-★ 135 [ə:m]
- B 136 ★[m]||[hm̃]■★