FINDING CLAUSES IN UNRESTRICTED TEXT BY FINITARY AND STOCHASTIC METHODS

Eva I. Ejerhed
AT&T Bell Laboratories & University of Umea
Department of Linguistics
University of Umea
S-90187 Umea, Sweden

ABSTRACT

The paper presents and compares two different methods of parsing, a regular expression method and a stochastic method, with respect to their success in identifying basic clauses in unrestricted English text. These methods of parsing were developed in order to be applied to the task of improving the detection of large prosodic units in the Bell Labs text-to-speech system, and were so applied experimentally. The paper also discusses the notion of basic clause that was defined as the parsing target. The result of a comparison of the error rates of the two parsing methods in the recognition of basic clauses showed that there was a 13% error rate for the regular expression method and a 6.5% error rate for the stochastic method.

1. Introduction

The present paper describes the procedure that was followed in an extended experiment to reliably find basic surface clauses in unrestricted English text, using various combinations of finitary and stochastic methods. The purpose was to make some improvements in the detection and treatment of large prosodic units above the level of fgroups in the Bell Labs text-to-speech system. This system currently relies exclusively on punctuation (commas and periods) for the detection of such units, i.e. tonal minor and major phrases. Commas are correlated with tonal minor phrases, and sentence final periods with tonal major phrases. The notion of fgroup (one or more function words followed by one or more content words), and its implementation in the Bell Labs text-to-speech system is described in Liberman & Buchsbaum (1985).

Correct automatic detection of major syntactic boundaries, in particular clause boundaries, is a prerequisite for automatic insertion of final lengthening, boundary tones and pauses at such boundaries within sentences (cf. Allen, Hunnicutt & Klatt 1987,and Altenberg 1987). These prosodic phenomena make significant contribution to the naturalness and intelligibility of synthetic speech. Unfortunately, the task of parsing unrestricted text correctly, in order to find the relevant sentence internal syntactic boundaries has turned out to be very difficult. This paper is a report of an attempt to provide a better foundation for parsing text by the use of simple finitary and stochastic computational methods. These simple methods have not figured prominently in the theory and practice of natural langauge parsing, with some exceptions (Langendoen 1975, Church 1982, Ejerhed & Church 1983). For an experimental, and more complicated method to derive all prosodic units in the text-to-speech system, i.e. not just tonal minor and major phrases but every type of prosodic unit, from the syntactic structure and length of constituents, see Wright, Bachenko & Fitzpatrick (1986).

The first purpose of the experiment was to test the performance of a finite state parser, when the parser was given the rather difficult and substantive tasks of finding basic, non-recursive clauses in continuous text, in which each word had been tagged with a part of speech label. Parts of the tagged Brown corpus were used, representing the genres of both informative and imaginative prose. The clause grammar, consisting of a regular expression for clauses of different kinds, was constructed by the author and it was first applied to text that was guaranteed to have correct parts of speech assigned to the words, so that problems in constructing the grammar could be isolated from problems in assigning correct parts of speech. The finite state parser that used the clause grammar consisted of a program that matched regular expressions for clauses against the longest substrings of tagged words that fit them, and it was constructed and implemented by K. Church.

The second purpose was to see whether basic clauses could also be recognized by stochastic programs, after these had been trained on suitable training material. The training material was prepared by hand-correcting the output of the program that processed the regular expressions for clauses. A stochastic program for assigning unique part of speech tags to words in unrestricted text had been created by K. Church, and trained on the tagged Brown corpus (see Church 1987). The resultant program is 95-99% correct in its performance, depending on the criteria of correctness used, and it can be used as a lexical front end to any kind of parser, i.e. not necessarily stochastic or finite state parsers. However, the question presented itself whether the stochastic procedure that was so successful in recognizing parts of speech could also be applied to more advanced tasks such as recognizing noun phrases and clauses. The present paper concentrates on the parsing of basic clauses. The parsing of noun phrases by the same two methods is compared in Ejerhed (1987), and the stochastic parsing of noun phrases is described in detail in Church (1987).

The structure of the paper is as follows. Section 2 defines the target of a basic clause, and reports on the outcome of the search for such units by the two methods. Section 3 discusses the correlations between clause units as defined by this paper, and the prosodic units of tonal minor and major phrases in the Bell Labs text-to-speech system.

2. *Finding Clauses*

2.1 *Why Clauses?*

Syntactic surface clauses are interesting units of language processing for a variety of reasons. In the surface clause, criteria of form and meaning converge to guarantee both that it can be recognized solely by surface syntactic properties and that it constitutes a meaningful unit (ideally a proposition) in a semantic representation.

Clauses have been investigated in psycholinguistic research. Jarvella (1971) found effects of both sentence boundaries and clause boundaries in recall of spoken complex sentences and took them, along with previous results of Jarvella & Pisoni (1970), to support a clause-by-clause view of within-sentence processing.

Later research on reading comprehension has found effects on gaze duration not only of word length and word frequency, but also of syntactic local ambiguity (garden paths) and of ends of sentences (Just & Carpenter 1984). However, the study of clause units as distinct from sentence units has not been carried out systematically in psycholinguistic experiments so far, and a lot of basic facts remain to be found out about the role of clause units of different kinds in the processes whereby spoken and written language is comprehended.

2.2 *The Definition of A Basic Clause*

Finding basic noun phrases is important as a stepping stone to finding clauses, on the assumption that an important subset of them have an initial sequence consisting of a noun phrase followed by a tensed verb as a defining characteristic. The result of scoring the respective success of the two methods of parsing basic noun phrases in sample text portions, reported in Ejerhed (1987), was the following. The regular expression output had 6 errors in 185 noun phrases, i.e. a 3.3% error rate. The stochastic output had 3 errors in 218 noun phrases, i.e. a 1.4% error rate. Both results must be considered good in the absolute sense of an automatic analysis of unrestricted text, but the stochastic method has a clear advantage over the regular expression method. Basic noun phrases can be found, which is of important for clause recognition.

The definition of basic clause that was used in this study has the following characteristics: a) it concentrates on certain defining characteristics present at the beginnings of clauses; b) it follows from a particular hypothesis about syntactic working memory: that it is limited to processing one clause at the time; and c) it assumes that the recognition of any beginning of a clause automatically leads to the syntactic closure of the previous clause.

It should be clear from the above, that the theoretical reasons for pursuing a recursion-free definition of a basic clause have to do with a theory of linguistic performance, rather than with a theory of linguistic competence, in which memory limitations play no part. It is a hypothesis of the author's current clause-by-clause processing theory, that a unit corresponding to the basic clause is a stable and easily recognizable surface unit, and that it is also an important partial result and building block in the construction of a richer linguistic representation that encompasses syntax as well as semantics and discourse structure.

2.3 *A Regular Expression for Basic Clauses*

Several versions of a regular expression for basic clauses were written by the author and preceded the one presented in Appendix 1, which was, applied to 60 files of Brown corpus tagged text of 2000 words each, newspaper texts A01-A20, scientific texts J01-J20 and fiction texts K01-K20.

The first half of the definition of **clause** introduces a few auxiliary definitions: *comp* for a set of complementizers, *punct* for a set of punctuation marks, and *tense* for a set of verb forms that are either certainly tensed ("BED" "BEDZ" "BEM" "BER" "BEZ" "DOD" "DOZ" "HVD" "HVZ" "MD" "VBD" "VBZ") or possibly tensed ("BE" "DO" "HV" "VB"). The definition of clause also uses the previously defined **brown-np-regex**. The second and larger part of the definition of **clause** consists of a union of six concatenations.

The first defines complete main clauses as consisting of a sequence of an optional coordinating conjunction CC followed by an obligatory basic noun phrase followed by optional non-clausal complements and an optional adverb followed by an obligatory tensed verb followed by anything expcept the punctuations or complementizers indicated in the list after (not ..., followed by optional punctuation.

The second defines clauses introduced by an obligatory CC followed by an optional adverb followed by an obligatory element which is either a tensed or participial verb form, followed by the same clause ending as in the first definition.

The third concatenation defines a subordinate clause as starting with an optional coordinating conjunction followed by an obligatory complementizer followed by the same clause ending as in the first and second definitions.

The remaining three definitions are of clause fragments rather than full clauses. Consider the following sentence: *The man [who liked ice cream,] ate too much.*

In it, the relative clause makes a basic clause unit that breaks up the main clause into two clause fragments. The third concatenation defines noun phrase fragments that begin with a basic noun phrase followed optionally by one or more prepositional phrases, or sequences of CC np or $ np, followed by the same clause ending as in the other definitions. In the example above, *[the man]* would be a noun phrase fragment.

The fifth concatenation defines verb phrase fragments, e.g. *[ate too much].*

The sixth concatenation defines clause fragments that are adjuncts, i.e. adverbial phrases, prepositional phrases and adjective phrases. The typical case in which such a fragment is recognized is when it precedes another clause: *[On a clear day,] [you can see forever].*

2.4 Output of Regular Expression for Clauses

The regular expression in Appendix 1 was automatically expanded into a deterministic fsa for clause recognition by Church's program. This rule compilation will not be described here. An excerpt from the result of applying it to the 60 files mentioned in the introduction to this section is presented in Appendix 2, where the location and nature of hand-corrections have been highlighted. The hand-correction was guided by the following principles.

1) There should be at most one tensed verb per clause. This inserts a clause boundary after a tensed clause and before a tensed verb in the following kind of case, which the current regular expression matcher does not capture: *[The announcement] [that the President was late] [was made late in the afternoon].*

2) There should be a clause boundary after a sentence initial prepositional or adverbial phrase and before the sequence np tensed verb, whether or not they are separated by a comma:*[At the summit in Iceland] [Gorbachev insisted ...].*

3) There should be a clause boundary before CC followed by a tensed verb. Although the second concatenation in the clause regex aimes at capturing such clauses, it is not always successful in doing so because there is no way, given the current implementation of negation in the regular expression program, to state that a clause should end before a concatenation of items, i.e. before (* CC tense). Only single items can be negated at present. Example: *[The Purchasing Departments are well operated] [and follow generally accepted practices].*

4) There should be a clause boundary before a preposition (IN) followed by a wh-word, i.e. before (* IN (+ WDT WPO WP$ WRB WQL)). For the same reason given under 3), there is no way currently to state that a clause should end before such a sequence. Example: *[The City Executive Committee deserves praise for the manner] [in which the election was conducted].*

Several interesting observation were made in the course of doing these hand-corrections. For one, there were errors in the Brown corpus assignment of tags, in particular several errors confusing VBD and VBN, and there were errors where the sequence TO VB was tagged IN NN. More seriously, it turned out that the words *as* and *like*, which have the property of functioning either like prepostions IN or subordinating conjunctions CS were always tagged CS, thus leading to incorrect recognition of clauses in many cases. Another problem for recognizing clauses on the basis of identifying tensed verbs was that the tag VB is applied to forms that are either infinitival or present tensed (or subjunctive), depending on context. It would have been better if such forms had been considered lexically ambiguous and given distinct tags. However, by and large the tagged Brown corpus is a very good and useful product, both in the choice of tags, and in the consistency with which they have been applied. Doing the hand-correction also forced the realization that the clause recognition program, like the noun phrase recognition program, dependes crucially on accurate assignment of parts of speech to all words, on order to work well. For this task, Church's stochastic parts program is admirably suited, since it gives correct assignments in a very large number of cases, and it holds the potential of further improvement in its performance with further training.

2.5 Stochastic Recognition of Clauses

As stated before, the regex **clause** was applied to sixty texts in the Brown corpus, and the output was hand-corrected. The hand-corrected files, containing an estimated total of at least 20,000 basic clauses, including clause fragments, were then used as training material for a stochastic recognition program. The training consisted of observing the location of clause opens and clause closes, and a special training specifically in locating tensed verbs. After training, the stochastic parts program and thereafter the stochastic clause recognizer was applied by K. Church to a large amount of Associated Press newswire text from May 26, 1987 (526 blocks, 2381353(8) bytes). An excerpt of the result is presented in Appendix 3. The result, again, is strikingly good.

A comparison of the nature and amount of errors in recognizing basic clauses in a sample of uncorrected regex output, and a sample of output from the stochastic clause program, can be made on the basis of Tables 1 and 2 at the end.

It appears that the stochastic program is more successful than the current regular expression method. However, certain improvements in the regex program could change that. What is needed is the facility to process *generalized regular expressions*, which admit the operations of complement and intersection, in addition to the operations of concatenation, union and Kleene star that characterize *regular expressions*. In any case there are some interesting differences in the kinds of errors made by the current regex program and the stochastic one for recognizing clauses. The regex program systematically errs by underrecognizing, never by overrecognizing, and in the selected portions that were scored, it only puts a few clause boundaries in the wrong place. It misses lots of clause boundaries, but the ones it gets are mostly correct.

The stochastic program, on the other hand, is able to get many clause boundaries correctly that elude the regular expression matcher, e.g. clauses not introduced by complementizers. The stochastic program errs both by overrecognizing and underrecognizing clauses, and sometimes it also places the clause open or clause close in the wrong place. Some cases of incorrect clause recognition are due to incorrect assignments of parts of speech to words. However, the total number of errors with the stochastic method (21) is smaller than the total number of errors with the regex method (40), for approximately the same number of clauses to be recognized, 304 versus 308. This is a very surprising outcome indeed, and if taken literally, without any further weighting of the different types of errors, it means that the error rate for the stochastic method for recognizing clauses is 6.5%, as compared with 13% for the regex method.

3. On the Relation between Clauses and Intonation Units

Finding basic clause units in arbitrary text is necessary in order to locate tonal minor phrases, which, in addition to a phrase accent, also have a boundary tone, and, particularly at slow rates of speech, a pause at the end of the phrase. The current experiment in text analysis has been concerned primarily with informative rather than imaginative prose, and envisages applications of the text-to-speech system to the reading of informative prose like newspaper text.

In the current Bell Labs text-to-speech system, tonal minor phrase boundaries are identified on the basis of commas, and tonal major phrase boundaries are identified on the basis of periods. Finding more tonal minor phrase boundaries by using syntactic structure, in addition to punctuation, is the problem we are trying to address with the methods described in this paper. In order to know where tonal minor phrase boundaries actually occur in the reading of informative texts, which typically have very long sentences (an average of 21 words compared with 14 words in general fiction based on Brown corpus data), it would be necessary to make recordings of several persons reading both authentic and prepared texts in a rhetorically explicit way, to borrow a phrase from Beckman & Pierrehumbert (1986), and then make extensive speech analyses of them, particularly of fundamental frequency movements and pauses. In the absence of such data for American English, the following kinds of boundaries between clauses and clause fragments were hypothesized to constitute intonation breaks with the status of tonal minor phrase boundaries. They are marked with # in the examples below.

a) After sentence initial adverbials and before np tense: *[At the summit in Iceland] # [Gorbachev insisted ...]*

b) After a relative clause and before a tensed verb: *[A House Committee] [which heard his local option proposal] # [is expected] [to give it a favorable report.]*

c) After other noun phrases with clausal complements and before a tensed verb: *[The announcement] [that the President was late] # [was made by the Press Secretary to the waiting journalists.]*

d) Before a set of complementizers categorized CS in the Brown corpus, it is frequently the case that there is an intonation break: *[that/CS ...], [whether/CS ...], [if/CS ...], [since/CS ...], [because/CS ...], [as/CS ...].*

However, there are some exceptions to this , in particular:

(i) Comparatives: *[This is not as/QL fast/JJ] [as/CS I would like ...]* or *[The theorem is more/RBR general/JJ] [than/CS what we have described]*

(ii) The words *as/CS* and *like/CS* when used as prepositions, i.e. followed by noun phrases that are not subjects of clauses: *[Jenkins left the White House in 1984,] [and joined Wedtech] [as/CS its director of marketing two years later.]*

For testing purposes, short passages of seven consecutive sentences each from the Brown files, and four sentences each from the AP newswire stories were synthesized by the author, using the Bell Labs text-to-speech system. Those boundaries between clauses and clause fragments that are identified above were implemented in the same way that commas are, i.e. with a phrase accent belonging to the tonal minor phrase, final lengthening, a boundary tone, and a short pause of 200 ms. The results have not yet been subjected to perceptual tests.

There are some studies of the relation between clause units and intonation units that provide relevant data for future work. Garding (1967) studied prosodic features in spontaneous and read Swedish speech. She found that in the spontaneous speech, pauses were equally divided between syntactic pauses and hesitation pauses, a syntactic pause being defined as one that coincides with a syntactic boundary. In the read speech, all pauses were syntactic pauses: "They appear between main clause and subordinate clause, before adverbial modifiers and between the different parts of an enumeration. The pause length is shortest in enumerations and before relative clauses (4-10 cs) and longest before adverbial modifiers and between complete sentences." (p. 48).

In a study of the intonational properties of relative clauses in British English, Taglicht (1977) compared the speech of a news broadcast with impromptu speech, and found that both genres separated nonrestrictive relative clauses prosodically. The news broadcast also separated a large proportion (71%) of the restrictive relative clauses prosodically.

A recent and very extensive study of the grammtical properties of intonation units, or tone units (TU) is Altenberg (1987). He studied a monologue of 48 minutes duration from the London-Lund Corpus of spoken English, and his results concerning the correlation of clause boundaries and tone unit boundaries are presented in Table 3 at the end.

4. Conclusion

The study reported above shows that basic clauses, including basic noun phrases, are stable and surface recognizable units in the definitions they were given here, and that both finitary and stochastic methods can be used to find them in unrestricted text with a high degree of success. The comparison between the error rate of these two methods showed that the stochastic method performed better both in the recognition of basic noun phrases and basic clauses, which is an unexpected result.

REFERENCES

[1] Allen, Jonathan., Hunnicutt, M.Sharon. & Klatt, Dennis, 1987, *From text to speech. The MITalk system*, Cambridge, Cambridge University Press.

[2] Altenberg, Bengt, 1987, *Prosodic patterns in spoken English. Studies in the correlation between prosody and grammar for text-to-speech conversion.* Lund Studies in English 76, Lund, Lund University Press.

[3] Beckman, Mary & Pierrehumbert, Janet, 1986, Intonational structure in Japanese and English, *Phonology Yearbook* 3(1986), 255-309.

[4] Church, Kenneth W., 1982, *On memory limitations in natural language processing*, Bloomington, Indiana, Indiana University Linguistics Club.

[5] Church, Kenneth W., 1987, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, AT&T Bell Labs, (in this volume).

[6] Ejerhed, Eva, 1987, Finding Noun Phrases and Clauses in Unrestricted Text: On the Use of Stochastic and Finitary Methods in Text Analysis, (ms), AT& T Bell Labs.

[7] Ejerhed, Eva & Church, Kenneth W., 1983, Finite State Parsing, in F. Karlsson (ed.), *Papers from the Seventh Scandinavian Conference of Linguistics*, Helsinki. University of Helsinki, Department of General Linguistics.

[8] Francis, Nelson & Kucera, Henry, 1982, *Frequency Analysis of English Usage, Lexicon and Grammar*, Boston, Houghton Mifflin Company.

[9] Garding, Eva, 1967, Prosodiska drag i spontant och uppl{st tal, in G. Holm (ed.), *Svenskt talsprak*, Stockholm, Almqvist & Wiksell, 40-85.

[10] Jarvella, Robert, 1971, Syntactic Processing of Connected Speech, *JVLVB* 10, 409-416(1971).

[11] Jarvella, Robert & Pisoni, D.B., 1970, The Relation between Syntactic and Perceptual Units in Speech Processing, *JASA*, 1970, 48, 84 (A).

[12] Just, Marcel & Carpenter, Patricia, 1984, Using Eye Fixations to Study Reading Comprehension, in D. Kieras & M. Just (eds), *New Methods in Reading Comprehension Research*, Lawrence Earlbaum Associates, Hillsdale, New Jersey, 151-182.

[13] Langendoen, D. Terrence, 1975, Finite-State Parsing of Phrase-Structure Languages and the Status of Readjustment Rules in Grammar, *Linguistic Inquiry*, Vol VI (1975), Number 4.

[14] Liberman, Mark & Buchsbaum, Adam, 1985, Structure and Usage of Current Bell Labs Text to Speech Program, (ms), AT&T Bell Labs.

[15] Taglicht, J., 1977, Relative clauses as postmodifiers: meaning syntax and intonation, in W.-D. Bald & R. Ilson (eds.), *Studies in English usage*, Frankfurt/M, Peter Lang, 73-107.

[16] Wright, Charles, Bachenko, Joan & Fitzpatrick, Eileen, 1986, The contribution of parsing to prosodic phrasing in an experimental text-to-speech system, *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University.

APPENDIX 1 Regular expression for basic clauses.

```
(defvar *clause*
 (let* ((comp '(+ "CS" "TO" "WDT" "WRB"
                 "WPS" "WPO" "WP$" "WQL"))
   (punct '(+ "," "." "--" ":"))
   (tense '(+ "BE" "BED" "BEDZ" "BEM"
             "BER" "BEZ" "DO" "DOD" "DOZ"
             "HV" "HVD" "HVZ" "MD" "VB"
             "VBD" "VBZ")))
      ;; main clause: (CC) np tense ...
   '(+ (* (cl-user::opt "CC")
      ,*brown-np-regex*
      (cl-user::opt (++ (* (+ "CC" "IN" "$")
            ,*brown-np-regex*)))
      (cl-user::opt (+ "RB" "RBR"))
            ,tense
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
            "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct))
       (* "CC"    ; main clause: CC tense ...
      (cl-user::opt (+ "RB" "RBR"))
      (+ ,tense "VBG" "VBN" "BEG" "HVG")
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
            "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct))
       (* (cl-user::opt "CC")    ; sub clause
      (++ ,comp)
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
            "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct))
       (* (cl-user::opt "CC")    ; np clause fragment
      ,*brown-np-regex*
      (cl-user::opt (++ (* (+ "CC" "IN" "$")
                     ,*brown-np-regex*)))
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
            "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct))
        ;; vp clause fragment
       (* (+ ,tense "VBG" "VBN" "BEG" "HVG")
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
             "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct))
        ;; adjunt clause fragment
       (* (cl-user::opt "CC")
      (++ (* (+ "RB" "RBR" "RP" "QL"
             "*" "NR" "JJ" "JJR"
             "IN" ,*brown-np-regex* )))
      (cl-user::opt (++ (not "," "." "--" ":"
            "CS" "TO" "WDT" "WRB"
            "WPS" "WPO" "WP$" "WQL")))
      (cl-user::opt ,punct)))))
```

APPENDIX 2

Sample of output of applying the regular expression *clause* as defined in Appendix 1, to Brown newspaper story A01. Hand-corrections are marked by double asterisks for underrecognized, and single asterisks for overrecognized clause boundaries.

[the/AT Fulton/NP-TL County/NN-TL Grand/JJ-TL Jury/NN-TL said/VBD Friday/NR ** an/AT investigation/NN of/IN Atlanta/NP 's/$ recent/JJ primary/NN election/NN produced/VBD no/AT evidence/NN] [that/CS any/DTI irregularities/NNS took/VBD place/NN./.]

[the/AT jury/NN further/RBR said/VBD in/IN term-end/NN presentments/NNS] [that/CS the/AT City/NN-TL Executive/JJ-TL Committee/NN-TL ,/,] [which/WDT had/HVD over-all/JJ charge/NN of/IN the/AT election/NN ,/,] [deserves/VBZ the/AT praise/NN and/CC thanks/NNS of/IN the/AT City/NN-TL of/IN-TL Atlanta/NP-TL for/IN the/AT manner/NN ** in/IN] * [which/WDT the/AT election/NN was/BEDZ conducted/VBN ./.]

[the/AT September-October/NP term/NN jury/NN had/HVD been/BEN charged/VBN by/IN Fulton/NN-TL Superior/JJ-TL Court/NN-TL Judge/NN-TL Durwood/NP Pye/NP] [to/TO investigate/VB reports/NNS of/IN possible/JJ irregularities/NNS in/IN the/AT hard-fought/JJ primary/NN] [which/WDT was/BEDZ won/VBN by/IN Mayor-nominate/NN-TL Ivan/NP Allen/NP Jr./NP ./.]

[only/RB a/AT relative/JJ handful/NN of/IN such/JJ reports/NNS was/BEDZ received/VBN ,/,] [the/AT jury/NN said/VBD ,/,]N [considering/IN the/AT widespread/JJ interest/JJ in/IN the/AT election/NN ,/,] [the/AT number/NN of/IN voters/NNS and/CC the/AT size/NN of/IN this/DT city/NN ./.]

[the/AT jury/NN said/VBD ** it/PPS did/DOD find/VB] [that/CS many/AP of/IN Georgia/NP 's/$ registration/NN and/CC election/NN laws/NNS are/BER outmoded/JJ or/CC inadequate/JJ and/CC often/RB ambiguous/JJ ./.]

[it/PPS recommended/VBD] [that/CS Fulton/NP legislators/NNS act/VB] [to/TO have/HV these/DTS laws/NNS studied/VBN and/CC revised/VBN to/IN the/AT end/NN of/IN modernizing/VBG and/CC improving/VBG them/PPO ./.]

[the/AT grand/JJ jury/NN commented/VBD on/IN a/AT number/NN of/IN other/AP topics/NNS ,/,] [among/IN them/PPO the/AT Atlanta/NP and/CC Fulton/NP-TL County/NN-TL purchasing/VBG departments/NNS ** which/WDT it/PPS said/VBD ** are/BER well/QL operated/VBN ** and/CC follow/VB generally/RB accepted/VBN practices/NNS] [which/WDT inure/VB to/IN the/AT best/JJT interest/NN of/IN both/ABX governments/NNS ./.]

APPENDIX 3

Sample of output of stochastic procedure for finding clause boundaries. Tensed verbs should be in bold face. In the recognition of these clauses, the constraint was enforced that there be at most one tensed verb per clause. Hand-corrections marked as in Appendix 2.

[former/AP U.S./NP Attorney/NN General/NN Ramsey/NP Clark/NP **said/VBD!** Monday/NR] [he/PPS **believed/VBD!**] [he/PPS **had/HVD!** found/VBN evidence/NN of/IN a/AT growing/VBG CIA/NP role/NN in/IN the/AT Philippines/NPS '/$ war/NN against/IN communist/NN rebels/NNS ./.]

[Clark/NP ,/,] [who/WPS **arrived/VBD!** last/AP week/NN] * [as/CS the/AT head/NN of/IN a/AT private/JJ ,/,] [human/JJ rights/NNS team/NN ,/,] [**said/VBD!**] [he/PPS **hopes/VBZ!**] [**to/TO!** document/VB the/AT evidence/NN] [and/CC present/VB it/PPO to/IN U.S./NP Secretary/NN of/IN State/NN George/NP P./NP Shultz/NP ./.]

[our/PP$ concern/NN **is/VBZ!** the/AT role/NN of/IN the/AT United/VBN States/NNS ,/,] [Clark/NP **told/VBD!** a/AT news/NN conference/NN ./.[

[we/PPSS **believe/VB!**] [we/PPSS **can/MD!** see/VB ,/,] [and/CC we **hope/VB!**] [**to/TO!** be/BE able/JJ] [**to/TO!** document/VB] * [before/CS we/PPSS **are/BER!**] * [through/RP in/IN our/PP$ report/NN ,/,] [evidence/NN clearly/RB establishing/VBG the/AT implementation/NN of/IN a/AT low-intensity/JJ campaign/NN here/RB ,/,] [with/IN violence/NN ,/,] [**to/TO!** kill/VB off/RP all/ABN opposition/NN ,/,] [every/AT opposition/NN to/IN authority/NN ,/,] [to/IN militarism/NN ./.]

[Ralph/NP McGehee/NP ,/,] [a/AT former/AP Central/JJ Intelligence/NN Agency/NN employee/NN ,/,] [**said/VBD!**] [he/PPS **recognized/VBD!** indications/NNS of/IN CIA/NP influence/NN in/IN the/AT Philippine/JJ military/NN 's/$ operations/NNS against/IN the/AT communist/JJ New/JJ People/NNS 's/$ Army/NN ./.]

[he/PPS **cited/VBD!** military/JJ search-and-destroy/JJ missions/NNS ,/,] [forced/VBN evacuation/NN of/IN civilians/NNS from/IN rebel-held/JJ areas/NNS and/CC the/AT increase/NN in/IN the/AT strength/NN of/IN civilian/JJ anti-communist/JJ vigilante/JJ groups/NNS ./.]

[the/AT allegations/NNS of/IN growing/VBG U.S./NP involvement/NN in/IN the/AT support/NN of/IN president/NN Corazon/NP Aquino/NP 's/$ government/NN **came/VBD!** with/IN claims/NNS by/IN Philippine/JJ leftists/NNS] [that/CS right-wing/JJ death/NN squads/NNS **are/BER!** operating/VBG freely/RB against/IN suspected/VBN leftists/NNS ./.]

TABLES

Table 1. Errors in regex recognition of clauses.

Regex Output					
Story	Sentences	Clauses (before)	Clauses (after)	Under	Wrong-place
a01	28	86	104	18	1
j01	28	98	107	9	1
k01	28	87	97	10	1
Total	84	271	308	37	3

Table 2. Errors in stochastic recognition of clauses.

Stochastic Output						
Story	Sentences	Clauses (before)	Clauses (after)	Under	Over	Wrong-place
STORY-1	15	64	64	0	1	1
STORY-2	15	52	51	0	1	0
STORY-3	15	45	46	1	0	2
STORY-4	30	141	143	8	4	3
Total	75	302	304	9	6	6

Table 3. The cooccurrence of clause boundaries and tone unit boundaries (from Altenberg 1987:57 Table 4:3).

Clause boundaries cooccurring with a TU boundary			
Clause boundary	Total	TU	%
After initial clauses	29	29	100
Around medial clauses	15	15	100
Before finite adverbial clauses	46	46	100
Before adverbial ing-clauses	14	14	100
Before nonrestrictive relative clauses	26	26	100
Before asyndetic clause coordination	15	15	100
Around nonrestrictive appositive clauses	3	3	100
After postmodifying clauses	67	66	99
Before syndetic clause coordination	153	150	98
Before nonfinite postmodifying clauses	25	19	76
Before restrictive relative clauses	26	18	69
After comment clauses	13	9	69
Before adverbial infinitive clauses	12	8	67
Before comment clauses	13	8	62
Before nominal that-clauses	32	19	59
Before direct speech	7	4	57
Before nominal relative/interrogative clauses	16	7	44
Before nonfinite nominal clauses	21	7	33
Before clauses as prepositional complement	21	1	5