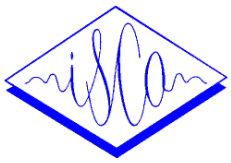

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

STRESS ASSIGNMENT IN COMPLEX NOMINALS FOR ENGLISH TEXT-TO-SPEECH

Richard Sproat

Linguistics Research Department
AT&T Bell Laboratories, Room 2d-451
Murray Hill, NJ 07974, USA

ABSTRACT

A difficult problem in English syntactic analysis is the treatment of complex nominals. In the context of text-to-speech one needs not only to syntactically analyze such constructions, but also to assign appropriate stress. I describe NP, a complex nominal analyzer currently under development in the context of the Bell Labs Text-to-Speech system. NP makes use of syntactic, semantic, lexical and statistical information to decide on the most appropriate structure for complex nominals; phonological rules are then applied to produce stress appropriate to that structure. An evaluation of the performance of the current program is presented.

1. INTRODUCTION

One difficult problem in the syntactic analysis of English is the treatment of COMPLEX NOMINALS (*CN*) such as the following:
(1) *several very large ornamental ducks*
(2) *computer communications network performance analysis primer*
(3) *New York Avenue*
(4) *former Attorney General Edwin Meese III*
Generally covered by the term CN are cases where: a sequence of one or more adjectives (or adjective phrases) modifies a noun (1); a head noun is preceded by some number of other nouns (2); the construction is a complex proper name (3); or the construction is some combination of the above, as in (4) where the noun-noun sequence *Attorney General* is modified by the adjective *former* and the resulting construction modifies the complex proper name *Edwin Meese III*.
For text-to-speech (*TTS*), we are interested in assigning appropriate stress (and intonation) to various kinds of phrases, including CNs. Traditional linguistic description [Chomsky and Halle, 1968] (and see also [Liberman and Sproat, 1990]) argues that one needs to know two things in order to algorithmically assign appropriate stress to a CN in a 'discourse neutral' context (see [Hirschberg, 1990] for methods which model the effects of discourse on the stressing of CNs). First, one must know the structure of the CN in terms of a (binary branching) tree. Secondly, one must know the node labels of the tree; it is commonly assumed that compound *words* are stressed on the left, whereas *phrases* are stressed on the right:

Type	Syntactic Node Label	Stress Pattern	Example
Nominal compound	N^0	LEFT	*DOG house*
Nominal phrase	$\overline{N}$	RIGHT	*nice HOUSE*

However, both stress *and* structure are hard to compute in general, since the decision is often not constrained by syntactic considerations. So, knowing the parts of speech (*POS*) for the words in a CN is only a partial predictor of the stress pattern; for example, N-N sequences are

stressed on the left between 75% and 95% of the time depending upon the type of text (see [Liberman and Sproat, 1990]), but there are many (systematic) exceptions (*lobster RAGOUT*). And POS information can be useless in computing the structure of a very long CN; for (2), the most plausible structure seems to be
[[*computer* [*communications network*]] [[*performance analysis*] *primer*]],
yet POS information only tells us that we have a sequence of nouns, which does not constrain the possible structures. Often the best structure in such cases must be picked on the basis of semantic plausibility. Since such semantic analysis is difficult, most CN analyzers have been designed for a restricted domain, and have been built on top of a fairly complete knowledge representation for that domain [Finin, 1980]. Yet CNs are common in English text—over 70,000 tokens per million words of text as estimated from the Brown Corpus [Francis and Kucera, 1982], and since in the TTS domain one cannot typically restrict the input, one would like to provide a treatment of CNs which is both general and accurate.
In this paper, I shall suggest that while one cannot as yet hope to do thorough semantic and structural analyses for CNs in unrestricted English text, one can usefully do fairly substantial analyses using simple and easily obtained lexical and statistical information. I shall describe NP, a CN analyzer which I am developing in the context of Bell Labs' TTS system, which makes use of syntactic, lexical and semantic information, as well as statistical information derived from large text corpora, to attempt to produce correct structures for CNs in unrestricted text. Having produced the structure, NP computes an appropriate stress pattern, and inserts minor intonational phrase boundaries in appropriate places in sufficiently long CNs. I shall also discuss NP's current performance.

2. A DESCRIPTION OF NP

NP takes as input, text which has been labeled for POS, where the left and right edges of noun phrases have been delimited; this labeling is accomplished using a stochastic POS assigner [Church, 1988]. For each noun phrase, NP uses a CKY (context-free) recognition algorithm (see [Harrison, 1978, section 12.4]) to build a chart of possible constituents. The recognizer uses the following kinds of context-free rules:

- Phrase structure rules—e.g., $NP \rightarrow DET + \overline{N}$. There are currently about 40 such rules.
- Context-free semantic/lexical schemata which predict the category label of the mother node. For example:
  - General schemata: a term for FURNITURE combines with a term for ROOM to form an $\overline{N}$ (i.e., righthand stress—*kitchen TABLE*).
  - Schemata with particular head nouns: street names ending in the word *Street* are N^0 (i.e., lefthand stress—*PARK Street*). Currently there are about 600 of these first two kinds of rules, with about 95% of those being specific to food terms.
  - Lists of particular nominals with information about their phrasal status—e.g., *WHITE House* is an N^0 (taking lefthand stress). There are currently about 6500 entries of this kind.

In constructing the semantic rules, on-line thesauruses have proved useful. Having built the chart, NP picks one of the set of trees defined therein by doing a top-down walk and applying the following heuristics at each node:

- Given a choice of n possible expansions of a node, prefer the expansion which is derived by semantic schemata over the expansion which is derived purely by syntactic schemata.
- Ceteris paribus, given a choice of n expansions of a node, pick the more phrasal expansion (e.g., pick $\overline{N}$ over N^0).

- Ceteris paribus, given a choice of n expansions of a node, pick the expansion which has the more right branching structure. (In English, a right branching—or flat—structure has stress close to the right edge of the phrase, and if there is little evidence, it is better to err in the direction of putting stress too far to the right than to err in the opposite direction.)

Additionally, in the case of a node dominating three words, we make use of a mutual-information-like measure (IX) proposed in [Liberman and Sproat, 1990]. The scores $IX(w_1, w_2)$, $IX(w_2, w_3)$ for two bigrams w_1w_2, w_2w_3 in a trigram $w_1w_2w_3$, are computed according to the normal equation for mutual information [Fano, 1961], except that the instances where the bigrams (w_1w_2, w_2w_3) and unigrams (w_1, w_2, w_3) occur in the trigram ($w_1w_2w_3$) in the corpus are discounted from the computation. If $IX(w_1, w_2)$ is greater than $IX(w_2, w_3)$ then we favor a left branching parse (if that is syntactically possible) and if the reverse is true then we favor a right branching parse. The statistics for computing the IX measure were derived from a corpus of 10 million words of the *Associated Press* Newswire for 1988. As an example of the use of this statistic, consider that the trigram *Wall Street Journal* occurs 75 times in the corpus, the bigram *Wall Street* 711 times and the bigram *Street Journal* 80 times. Since *Wall Street* occurs 636 times outside the context of *Wall Street Journal*, but *Street Journal* only occurs 5 times outside this context, the measure IX strongly favors the (correct) structure [[*Wall Street*] *Journal*].

Finally, once a tree has been built over as much of the CN as is feasible, NP passes the tree to the phonology, where the position of the primary stress is determined, and the rhythm rule is applied (at present following the algorithm in [Monaghan, 1990]). Additionally, in CNs with sufficiently long subconstituents, minor intonational phrase boundaries are inserted.

As an example of the parsing algorithm, consider the sentence *I put it on the living room table.* Church's POS algorithm will assign POS labels to the words in the sentence and will bracket the CN *the living room table.* NP will then apply syntactic and semantic rules to the CN. For example, NP knows that *living room* is a ROOM (and also that it is an N^0), so it places a node [N^0,ROOM] in the chart, spanning *living room.* It also knows that *table* is a kind of FURNITURE, and that a ROOM term and FURNITURE term can combine into an $\overline{N}$, so it places the node [$\overline{N}$,ROOM&FURNITURE] in the chart, spanning the string *living room table.* Finally, the determiner *the* can combine with an $\overline{N}$ to make an NP. Having put these (and other) nodes into the chart, the algorithm starts at the top of the chart, first expanding the NP node spanning the string *the living room table* (because that is the only node at that level) and then expanding the node [$\overline{N}$,ROOM&FURNITURE] (spanning *living room table*), which wins over other candidates because it is of a high bar level ($\overline{N}$ rather than N^0—the second heuristic above) and because it has a semantic annotation (the first heuristic above). We therefore arrive at the correct structure: [$_{NP}$ *the* [$_{\overline{N}}$ [$_{N^0}$ *living room*] *table*]]. Phonological rules are then applied to yield the correct stress pattern: *the LIVING room TABLE.*

3. EVALUATION

The most significant improvement afforded by NP to the default stress assignment for CNs given by our synthesizer is in the placement of primary stress; other improvements, such as the correct placement of accents within the CN prior to the primary stress are noticeable, but are not as striking. I therefore concentrated on evaluating the *current* performance of NP at predicting primary stress placement. I chose at random 500 CNs containing two or more words (not counting determiners and pronominal possessives) from a day's worth of the *Associated Press* Newswire, and ran NP on those nominals. An independent judge (Jill Burstein) listened to the output and was asked to evaluate it as a discourse neutral way of saying the CN, using a three-point scale: GOOD (= "that's how I would say it"); MAYBE (= "not how I would say it, but it is still acceptable"); BAD (= "I couldn't imagine it being said that way"). The totals for each judgment were as follows:

- GOOD: 455 (91%) • MAYBE: 27 (5.4%) • BAD: 18 (3.6%)

So, the current program gets primary stress acceptably in 96.4% (= 91+5.4) of the cases. To put this in perspective, one needs to compare NP's performance at stress placement with two simpler algorithms: *Method A* always places stress on the last word of the CN (what our TTS system traditionally has done); *Method B* places stress on the penultimate word of the CN if and only if the CN ends in two nouns. The following table shows the performance of these methods (in percentage correct) on the 500 CN sample:

Method A	Method B	NP
77%	93.8%	96.4%

Clearly Method B works much better than Method A, meaning that—on this particular corpus—a fair amount can be achieved by just having part of speech information. However, the performance of NP shows that we can do better still. Furthermore, while purely syntactic algorithms like Method B can never achieve full coverage for reasons already discussed, it is possible to extend NP's coverage. For example, one of the errors made by NP on the test corpus involved failing to place final stress on *red snapper RAGOUT*. Although NP knows that FISH and STEW terms can combine to form $\overline{\text{N}}$s, it does not know that *red snapper* is a kind of fish; such information is easily added to NP's database however.

Of course, there are additional advantages of an approach like NP's over a simple stress assignment algorithm like that of Method B: NP can also assign structure to a CN, which is useful for assigning appropriate CN-internal intonational boundaries, as mentioned above. I conclude then that the general approach which NP incorporates is a useful step towards more natural sounding synthetic speech.

ACKNOWLEDGMENTS

I wish to thank Ken Church, Julia Hirschberg and David Talkin for useful comments.

REFERENCES

[Chomsky and Halle, 1968] Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* Harper and Row, New York.

[Church, 1988] Church, K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin. Association for Computational Linguistics.

[Fano, 1961] Fano, R. (1961). *Transmission of Information.* MIT Press, Cambridge MA.

[Finin, 1980] Finin, T. (1980). *The Semantic Interpretation of Compound Nominals.* PhD thesis, University of Illinois, Urbana-Champaign.

[Francis and Kucera, 1982] Francis, W. N. and Kucera, H. (1982). *Frequency analysis of English usage.* Houghton Mifflin, Boston.

[Harrison, 1978] Harrison, M. (1978). *Introduction to Formal Language theory.* Addison Wesley, Reading MA.

[Hirschberg, 1990] Hirschberg, J. (1990). Using discourse context to guide pitch accent decisions in synthetic speech. In *ESCA Workshop on Speech Synthesis*, Autrans, France. ESCA.

[Liberman and Sproat, 1990] Liberman, M. and Sproat, R. (1990). The stress and structure of modified noun phrases in English. In Sag, I., editor, *Lexical Matters.* University of Chicago Press. To Appear.

[Monaghan, 1990] Monaghan, A. (1990). Rhythm and stress-shift in speech synthesis. *Computer Speech and Language*, 4(1):71–78.