TOWARD TREATING ENGLISH NOMINALS CORRECTLY

Richard W. Sproat, Mark Y. Liberman
Linguistics Department
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Abstract

We describe a program for assigning correct stress contours to nominals in English. It makes use of idiosyncratic knowledge about the stress behavior of various nominal types and general knowledge about English stress rules. We have also investigated the related issue of parsing complex nominals in English. The importance of this work and related research to the problem of text-to-speech is discussed.

1. Introduction

We will discuss the analysis of English expressions consisting of a head noun preceded by one or more open-class specifiers: *rising prices*, *horse blanket*, *mushroom omelet*, *banana bread*, *parish priest*, *gurgle detector*, *quarterback sneak*, *blind spot*, *red herring*, *bachelor's degree*, *Planck's constant*, *Madison Avenue*, *Wall Street*, *Washington's birthday sale*, *error correction code logic*, *steel industry collective bargaining agreement*, *expensive toxic waste cleanup*, *windshield wiper blade replacement*, *computer communications network performance analysis primer*, and so forth. For brevity, we will call such expressions 'nominals.' Our main aim is an algorithm for assigning stress patterns to such nominal expressions; we will also discuss methods for parsing them.

Nominals are hard to parse, since their pre-terminal string is usually consistent with all possible constituent structures, so that we seem to need an analysis of the relative plausibility of the various meanings (Marcus, 1980; Finin, 1980). Even when the constituent structure is known (as trivially in the case of binary nominals), nominal stress patterns are hard to predict, and also seem to depend on meaning (Bolinger, 1972; Fudge, 1984; Selkirk, 1984). This is a serious problem for text-to-speech algorithms, since nominal expressions are common at the ends of phrases, and the location of a phrase's last accent has a large effect on its sound. Complex nominals are common in most kinds of text; for example, in the million words of the Brown Corpus (Francis and Kučera, 1982), there are over 75,000 nominals containing more than two words.

However, we have been able to make some progress on the problems of parsing and stress assignment for nominals in unrestricted text. This paper concentrates on the representation and use of knowledge relevant to the problem of assigning stress; this same knowledge turns out to be useful in parsing.

For the purposes of this paper, we will be dealing with nominals in contexts where the default stress pattern is not shifted by phenomena such as as intonational focus or contrastive stress, exemplified below:

(1) a. We're only interested in solvable problems. (words like *only* depend on stress to set their scope — otherwise, this nominal's main stress would be on its final word.)
b. He's a lion-**tamer**, not a lion-**hunter**. (in a non-contrastive context, these nominals' main stresses would be on their penultimate words.)

These interesting phenomena rarely[1] shift main phrase stress in expository text, and are

1. In our samples, only a fraction of a percent of complex nominals in phrase-final position have their main stress shifted by focus or contrast.

best seen as a modulation of the null-hypothesis stress patterns.

We have argued elsewhere (Liberman and Sproat, 1987) for the following positions: (i) the syntax of modification is quite free — various modifiers of nominal heads (including adjectives, nouns, and possessives) may occur as sisters of any X-bar projection of the nominal head; (ii) modification at different X-bar levels expresses different types of meaning relations (see also Jackendoff, 1973); (iii) the English nominal system includes many special constructions that do not conform to the usual specifier-head patterns, such as complex names, time and date expressions, and so forth; (iv) the default stress pattern depends on the syntactic structure.

Points (ii) and (iv) are common opinions in the linguistic literature. In particular, we support generative phonology's traditional view of phrasal stress rules, which is that structures of category N^0 have the pattern assigned by the **compound stress rule**, which makes left-hand subconstituents stress-dominant unless their right-hand sisters are lexically complex.[2] In simple binary cases, this amounts to left-hand stress. All other structures are (recursively) right stressed, according to what is called the **nuclear stress rule**.[3]

Points (i) and (iii) are less commonplace. They make it impossible to predict stress from the preterminal string of a binary nominal, since the left-hand element may be attached at any bar level, or may be involved in some special construction. We do not have space to argue here for this point of view, but some illustrative examples may help make our position clearer.

Examples of adjectives and possessives within N^0 include *sticky bun, black belt, safe house, straight edge, sick room, medical supplies, cashier's check, user's manual, chef's knife, Melzer's solution,* etc. We can see that this is not simply a matter of non-compositional semantics by contrasting the stress pattern of *red herring, blue moon, Irish stew, hard liquor, musical chairs, dealer's choice, Avogadro's number, cat's pajamas.* The N^0 status of e.g. *user's manual* can be seen by its stress pattern as well as its willingness to occur inside quantifiers and adjectives: *three new user's manuals*, but **three new John's books*. In addition, there are several classes of possessive phrases that take right-hand stress but pattern distributionally like adjectives, i.e. occur at N^1 level, as in *three Kirtland's Warblers*. Examples of nouns at N^1 level include the common 'material-made-of' modifiers (such as *steel bar, rubber boots, paper plate, beef burrito,*), as well as most time and place modifiers (*garage door, attic roof, village street, summer palace, spring cleaning, holiday cheer, weekend news*), some types of modification by proper names (*India ink, Tiffany lamp, Miami vice, Ming vase*), and so on.

Thus a stress-assignment algorithm must depend on meaning relationships between members of the nominal, as well as the collocational propensities of the words involved.

We have written a program that performs fairly well at the task of assigning stress to nominals in unrestricted text. The input is a constituent structure for the nominal, and the output is a representation of its stress contour. Some examples of nominals to which the program assigns stress correctly are given in (2), where primary stress is marked by boldface and secondary stress by italics:

---

2. Various authors (e.g. Liberman & Prince 1977, Hayes 1980) have suggested that the behavior of the *compound stress rule*, which in fact applies to compound nouns but not to compound adjectives or verbs, is related to the tendency of non-compound English nouns to have their main stress one syllable farther back than equivalent verbs or adjectives. This generalization strengthens the argument that [N N] constituents with left-hand stress are of parent category N^0.

3. See Chomsky and Halle (1968), Liberman and Prince (1977), Hayes (1980) for various versions of these rules.

(2)
[[*Boston* University] [**Psychology** Department]]
[[[*Tom* Paine] Avenue] **Blues**]
[**corn** flakes]
[rice **pudding**]
[apricot **jam**]
[wood **floor**]
[cotton **shirt**]
[kitchen **towel**]
[Philadelphia **lawyer**]
[city **employee**]
[valley **floor**]
[afternoon **sun**]
[evening **primrose**]
[**Easter** bunny]
[**morning** sickness]
[[*Staten* Island] **Ferry**]
[**South** street]
[**baggage** claim]
[Mississippi **Valley**]
[Buckingham **Palace**]
[Surprise **Lake**]
[Murray **Hill**]

There are two main components to the program, the first of which deals almost exclusively with binary nominals and the second which takes n-ary nominals and figures out the stress pattern of those. We deal with each in turn.

2. Binary Nominals

Much of the work in assigning stress to nominals in English involves figuring out what to do in the binary cases, and this section will discuss how various classes of binary (and *some* n-ary nominals, n>2) are handled. For example, to stress [[*Boston* University] [Psychology Department]] correctly it is necessary to know that *Psychology Department* is stressed on the left-hand member. Once that is known, the stress contour of the whole four-member nominal follows from general principles, which will be outlined in the subsequent section of this paper.

To determine the stress pattern of a binary nominal, the following procedure is followed:

1. First of all, check to see if the nominal is listed as being one of those which is exceptionally stressed. For instance, our list of some 7000 left-stressed nominals includes [**morning** sickness], which will thus get left stress despite the general preference for right stress in nominals where the left-hand member is construed as as describing a location or time for the right-hand member. [Morning **prayers**], which follows the regular pattern, is stressed correctly by the program. Similarly, [**Easter** Bunny] is listed as taking left stress whereas [Easter **feast**] is correctly stressed on the right. There is a common misconception to the effect that all and only the lexicalized (i.e. listed) nominal expressions are left-stressed. This is false: lexicalization is neither a necessary nor a sufficient condition for left stress. *Dog annihilator* is left-stressed although not a member of the phrasal lexicon, and *red herring* is right-stressed although it must be lexically listed. Such examples abound (see, also, section 1).

2. If the nominal is not listed, check through all of the heuristic patterns that might fit it. A few examples of these patterns are given below — some of them are semantic or pragmatic in character, others are syntactic, and others are simply lexical. Note that there is not an easy boundary (for such an algorithm) between a pattern based on meaning and one based on word identity, since semantic classes correspond roughly to lists of words.

MEASURE-PHRASE: the left-hand member describes a unit of measure in terms of which the right-hand member is valued. Examples: *dollar bill*, *pint jug*, *5 gallon tank*... These normally take right stress.

LOCATION-TIME-OR-SUBSTANCE: the left-hand member describes the location or time of the right-hand member, or else a substance out of which the right-hand member is made. Location examples: *kitchen towel*, *downstairs bedroom*, *city hall*... Time examples: *Monday morning*, *Christmas Day*, *summer vacation*... Substance examples: *wood floor*, *china doll*, *iron maiden*. These normally take right stress.

ING-NOMINAL, AGENT-NOMINAL, DERIVED-NOMINAL: All of these are cases

where the right-hand member is a noun derived from a verb, either by affix *-ing* (*sewing*), *-er* (*catcher*) or some other affix (*destruction*). Nominals with these typically have left-hand stress if the left-hand member can be construed as a grammatical object of the verb contained in the right-hand member: *dog catcher, baby sitting, automobile demolition*. On the other hand if the left-hand member is a subject of the verb in the right-hand member then stress is usually right-hand: *woman swimmer, child dancing, student demonstration*.

NOUN-NOUN: If both elements are nouns, and no other considerations intervene, left-hand stress occurs a majority of the time. Therefore a sort of default rule votes for left-hand stress when this pattern is matched. Examples of correct application include: *dog house, opera buff, memory cache*. Not much weight is given to this possibility, since something which is simply *possibly* a left-stressed noun-noun compound may be many other things as well. Complex typologies of the meaning relations in noun-noun compounds can be found in Lees (1960), Quirk et al. (1972), Levi (1978). These typologies cross-cut the stress regularities in odd ways, and are semantically rather inhomogeneous as well, so their usefulness is questionable.

SELF: The left-hand member is the word *self* (e.g., *self promotion, self analysis...*). Right-hand stress is invariably assigned, since *self* is anaphoric, hence destressed following the normal pattern for anaphors.

PLACE-NAME: The right-hand member is a word like *pond, mountain, avenue* etc., and the left-hand member is plausibly a name. These cases get right-hand stress. Obviously, names ending in the word *Street* are an exception ([Madison Avenue] vs. [Wall Street]).

All of the applicable patterns for a given nominal are collected. Each pattern has a weight. For instance, as noted above, little weight is given to the observation that a particular nominal may be a noun-noun compound, since the preterminal string [N N] often belongs to categories that yield right-hand stress. On the other hand, if the analysis and its stress pattern are almost certain, as it is for sequences of the form [self N], then much weight is given to this pattern. The weights are tallied up as 'votes' for assigning to one member or the other. The pattern with the most votes wins. Currently the weights are assigned in an ad hoc manner by hand; we plan to replace the manual weight assignment with the results of a statistical survey of nominal types in various forms of English.

3. Assigning Stress to N-Ary Nominals

Given the stress pattern of binary cases, assigning stress to the general n-ary case is straightforward. The algorithm implemented is a version of one developed over the years by various researchers, including Chomsky and Halle (1968), Liberman and Prince (1977), Hayes (1980), Prince (1983) and others. Main stress is assigned to each level of constituent structure recursively, with relative stress values normally preserved as larger pieces of structure are considered. A convenient representation for tallying stress is the so-called 'metrical grid'; each word is associated with a set of marks or ticks on a grid whose higher, sparser levels correspond to metrically more important positions. For example, *dog catcher* would be represented as:

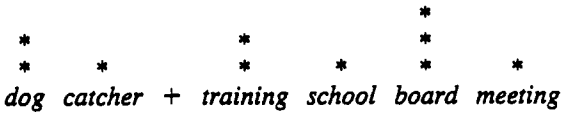
(3)

```
*
*     *
dog  catcher
```

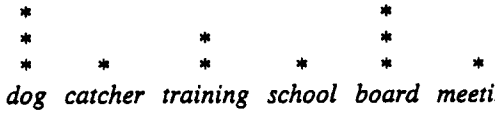
The fact that *dog* has two ticks as opposed to the one tick assigned to *catcher* is indicative of the stress prominence of *dog*.

When we combine two constituents together we upgrade the ticks of the highest tick-column of the weakest member to be the same as the highest column of the strongest member. For instance if we combine *dog catcher* with *training school board meeting* we will proceed by the following method:

(4)

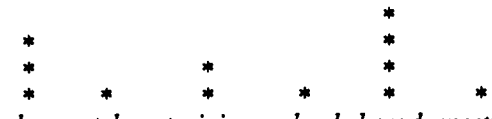


→



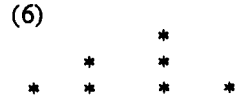
As a result, the most stressed element in each subunit starts out at 'tick parity' with the most stressed element in the other subunit. We then increment one of these main stresses to make it the main stress of the entire nominal.[4]

(5)



Finally the program tests for the applicability of the so-called Rhythm Rule. Given the rules so far, for a nominal such as *City Hall parking lot* we would expect the following stress contour:

(6)



However, the actual stress contour is:

(7)

The Rhythm Rule removes clashes between strong stresses by moving the left-hand stress back to the most prominent previous stress within the domain of the left-hand primary stress.

4. Performance of the Heuristic on 200 Binary Nominals.

To get a rough idea of how well our program is doing, we took 200 [N N] nominals from the *Bell Labs News*, and compared the performance of the current heuristic with two other procedures: (1) assigning stress uniformly to the right (which is what all current text-to-speech systems would do in such cases) and (2) assigning stress to the left if and only if the binary nominal can be analyzed as consisting of a noun followed by a noun. We had made no previous effort to develop heuristics appropriate for the content of this source material. The results were as follows:

(8) (i) Assigning uniform rightward stress: 45% correct.
(ii) Assigning leftward stress if N-N: 66%.
(iii) Current program: 80%.

Of our program's 40-odd failures, the cause was insufficient information in roughly 30 cases; only 10 were due to misanalysis. We classified the failure as being due to insufficient information when the program could say nothing about the categorization of either member of the compound, or could only ascertain that it might be dealing with a noun-noun sequence (which, the reader will recall, is given very little weight in making a decision). For instance, the program knows nothing about the stress properties of chemical terms, which invariably have right-hand stress, and therefore failed on *gallium arsenide* and several similar expressions. If the program had some information about at least one of the words, but still came up with the wrong

4. As pointed out in Liberman (1975), such bottom-up recursive stress assignment algorithms can simply be thought of as the definition of a relation of relative prominence on all the sets of sister nodes in the tree.

answer, then we classified the error as a case of misanalysis. The fact that most of the errors were due to insufficient information suggests that the program can be improved substantially by increasing its set of heuristic patterns and its knowledge of word classes. We guess that 90-95% correct stress is a plausible goal for [N N] nominals, even in technical writing, where our experience suggests that readers will assign left-hand and right-hand stress to such constituents with about equal frequency.

5. The Parsing Issue.

Our stress assignment program assumes a parsed input, not a reasonable option for a working text-to-speech system. There is some practical value in correct stress assignment to binary nominals only, since they are commoner than longer ones in most kinds of text; in the Tagged Brown Corpus (Francis and Kučera, 1982) we found that roughly 80% of the complex nominals were binary, 15% were ternary, and that therefore only about 5% had more than three members. Still, a count of 15% for ternary nominals is significant. Furthermore, higher percentages for complex nominals with more than two members are expected for technical writing than are exhibited in the Brown Corpus. We have therefore also investigated the use of the stress-assignment heuristics in parsing nominal expressions of higher complexity than binary. How would such patterns be useful? Consider an expression like *water supply control*, to which we would want to assign the structure *[[water supply] control]*. Given that we assume binary branching, we have two options, namely *[water [supply control]]* and *[[water supply] control]*. While the first analysis is not impossible, the second analysis would be favored since one of our patterns references the word *supply*, and lists substances such as water among the types of things that can have supplies. In effect, *supply* has a slot to its left which can optionally be filled by a noun referring to a substance or commodity of some kind, among which *water* is a prominent example. The word *supply* is not nearly so close to the core examples of likely arguments for *control*. Of course, listed complex nominals straightforwardly aid in parsing: a nominal such as *City Hall parking lot* is fairly easy to analyze given that in any case *City Hall* and *parking lot* are in our phrasal lexicon.

It seems clear that substantial amounts of lexical knowledge are necessary to parse complex nominals. This comes as no surprise, in light of much recent linguistic work suggesting that a substantial portion of linguistic knowledge resides 'in the lexicon.'

References

Bolinger, D. 1972. Accent is Predictable (if you're a mind-reader). Language 48, 633-45.

Chomsky, N. and M. Halle 1968. *The Sound Pattern of English*. New York: Harper and Row.

Finin, T. 1980. *The Semantic Interpretation of Compound Nominals*. Doctoral dissertation, University of Illinois.

Francis, W. N. and Kučera, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company.

Fudge, E. 1984. *English Word-Stress*. London and Boston: Allen and Unwin.

Hayes, B. 1980. *A Metrical Theory of Stress Rules*. Doctoral dissertation, MIT, distributed by Indiana University Linguistics Club.

Jackendoff, R. *X-bar Syntax: A Study of Phrase-Structure*. Cambridge and London: MIT Press.

Lees, R. 1960. *The Grammar of English Nominalizations*. Bloomington: Indiana University Press.

Levi, J. 1978. *The Syntax and Semantics of Complex Nominals*. New York and London: Academic Press.

Liberman, M. 1975. *The Intonational System of English*. Doctoral dissertation, MIT, reprinted 1979 by Garland, New York and London.

Liberman, M. and A. Prince 1977. On Stress and Linguistic Rhythm. *Linguistic Inquiry* 8, 249-336.

Liberman, M. and R. Sproat 1986. Stress Patterns in English Noun Phrases. Ms., AT&T Bell Labs.

Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge and London: MIT Press.

Prince, A. 1983. Relating to the Grid. *Linguistic Inquiry*, 14, 19-100.

Quirk, R., S. Greenbaum and G. Leech 1972. *A Grammar of Contemporary English*. London: Longman.

Selkirk, E. 1984. *Phonology and Syntax*. Cambridge and London: MIT Press.