

# STRUCTURE AND INTONATION IN SPOKEN LANGUAGE UNDERSTANDING\*

Mark Steedman

Computer and Information Science, University of Pennsylvania  
200 South 33rd Street  
Philadelphia PA 19104-6389  
([steedman@cis.upenn.edu](mailto:steedman@cis.upenn.edu))

## ABSTRACT

The structure imposed upon spoken sentences by intonation seems frequently to be orthogonal to their traditional surface-syntactic structure. However, the notion of “intonational structure” as formulated by Pierrehumbert, Selkirk, and others, can be subsumed under a rather different notion of syntactic surface structure that emerges from a theory of grammar based on a “Combinatory” extension to Categorical Grammar. Interpretations of constituents at this level are in turn directly related to “information structure”, or discourse-related notions of “theme”, “rheme”, “focus” and “presupposition”. Some simplifications appear to follow for the problem of integrating syntax and other high-level modules in spoken language systems.

One quite normal prosody (b, below) for an answer to the following question (a) intuitively imposes the intonational structure indicated by the brackets (stress, marked in this case by raised pitch, is indicated by capitals):

- (1) a. I know that Alice prefers velvet.  
But what does MARY prefer?  
b. (MA-ry prefers) (CORduroy).

Such a grouping is orthogonal to the traditional syntactic structure of the sentence.

Intonational structure nevertheless remains strongly constrained by meaning. For example, contours imposing bracketings like the following are not allowed:

- (2) #(Three cats)(in ten prefer corduroy)

\*I am grateful to Steven Bird, Julia Hirschberg, Aravind Joshi, Mitch Marcus, Janet Pierrehumbert, and Bonnie Lynn Webber for comments and advice. They are not to blame for any errors in the translation of their advice into the present form. The research was supported by DARPA grant no. N0014-85-K0018, and ARO grant no. DAAL03-89-C0031.

Halliday [6] observed that this constraint, which Selkirk [14] has called the “Sense Unit Condition”, seems to follow from the *function* of phrasal intonation, which is to convey what will here be called “information structure” – that is, distinctions of focus, presupposition, and propositional attitude towards entities in the discourse model. These discourse entities are more diverse than mere nounphrase or propositional referents, but they do not include such non-concepts as “in ten prefer corduroy.”

Among the categories that they *do* include are what Wilson and Sperber and E. Prince [13] have termed “open propositions”. One way of introducing an open proposition into the discourse context is by asking a Wh-question. For example, the question in (1), *What does Mary prefer?* introduces an open proposition. As Jackendoff [7] pointed out, it is natural to think of this open proposition as a functional *abstraction*, and to express it as follows, using the notation of the  $\lambda$ -calculus:

- (3)  $\lambda x [(prefer' x) mary']$

(Primes indicate semantic interpretations whose detailed nature is of no direct concern here.) When this function or concept is supplied with an argument *corduroy'*, it *reduces* to give a proposition, with the same function argument relations as the canonical sentence:

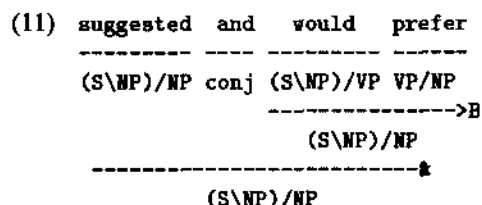
- (4)  $(prefer' corduroy') mary'$

It is the presence of the above open proposition rather than some other that makes the intonation contour in (1)b felicitous. (That is not to say that its presence uniquely *determines* this response, nor that its explicit mention is necessary for interpreting the response.)

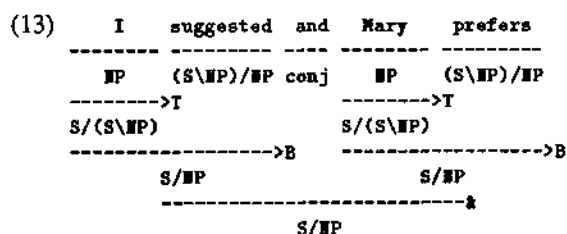
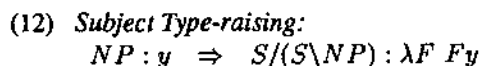
These observations have led linguists such as Selkirk to postulate a level of “intonational structure”, independent of syntactic structure and re-



and *would prefer*, *corduroy* can be accepted, via the following composition of two verbs (indexed as B, following Curry's nomenclature) to yield a composite of the same category as a transitive verb. Crucially, composition also yields the appropriate interpretation for the composite verb *would prefer*:



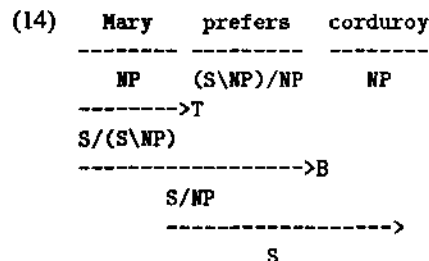
Combinatory grammars also include type-raising rules, which turn arguments into functions over functions-over-such-arguments. These rules allow arguments to compose, and thereby take part in coordinations like *I suggested, and Mary prefers, corduroy*. They too have an invariant compositional semantics which ensures that the result has an appropriate interpretation. For example, the following rule allows the conjuncts to form as below (again, the remainder of the derivation is omitted):



This apparatus has been applied to a wide variety of coordination phenomena (cf. [4], [15]).

## INTONATION AND CONTEXT

Examples like the above show that combinatory grammars embody a view of surface structure according to which strings like *Mary prefers* are constituents. It follows, according to this view, that they must also be possible constituents of non-coordinate sentences like *Mary prefers corduroy*, as in the following derivation:



(See [9], [18] and [19] for a discussion of the obvious problems for parsing written text that the presence of such "spurious" (i.e. semantically equivalent) derivations engenders, and for some ways they might be overcome.) An entirely unconstrained combinatory grammar would in fact allow any bracketing on a sentence, although the grammars we actually write for configurational languages like English are heavily constrained by local conditions. (An example might be a condition on the composition rule that is tacitly assumed below, forbidding the variable  $Y$  in the composition rule to be instantiated as NP, thus excluding constituents like  $*[ate\ the]_{VP/N}$ ).

The claim of the present paper is simply that particular surface structures that are induced by the specific combinatory grammar that are postulated to explain coordination in English subsume the intonational structures that are postulated by Pierrehumbert *et al.* to explain the possible intonation contours for sentences of English. More specifically, the claim is that that in spoken utterance, intonation helps to determine *which* of the many possible bracketings permitted by the combinatory syntax of English is intended, and that the interpretations of the constituents that arise from these derivations, far from being "spurious", are related to distinctions of discourse focus among the concepts and open propositions that the speaker has in mind.

The proof of this claim lies in showing that the rules of combinatory grammar can be made sensitive to intonation contour, which limit their application in spoken discourse. We must also show that the major constituents of intonated utterances like (1)b, under the analyses that are permitted by any given intonation, correspond to the information structure of the context to which the intonation is appropriate, as in (a) in the example (1) with which the paper begins. This demonstration will be quite simple, once we have established the following notation for intonation contours.

I shall use a notation which is based on the theory of Pierrehumbert [10], as modified in more recent work by Selkirk [14], Beckman and Pierrehumbert [1], [11], and Pierrehumbert and Hirschberg [12]. I

have tried as far as possible to take my examples and the associated intonational annotations from those authors. The theory proposed below is in principle compatible with any of the standard descriptive accounts of phrasal intonation. However, a crucial feature of Pierrehumbert's theory for present purposes is that it distinguishes two subcomponents of the prosodic phrase, the *pitch accent* and the *boundary*.<sup>2</sup> The first of these tones or tone-sequences coincides with the perceived major stress or stresses of the prosodic phrase, while the second marks the righthand boundary of the phrase. These two components are essentially invariant, and all other parts of the intonational tune are interpolated. Pierrehumbert's theory thus captures in a very natural way the intuition that the same tune can be spread over longer or shorter strings, in order to mark the corresponding constituents for the particular distinction of focus and propositional attitude that the melody denotes. It will help the exposition to augment Pierrehumbert's notation with explicit prosodic phrase boundaries, using brackets. These do not change her theory in any way: all the information is implicit in the original notation.

Consider for example the prosody of the sentence *Fred ate the beans* in the following pair of discourse settings, which are adapted from Jackendoff [7, pp. 260]:

- (15) Q: Well, what about the BEAns?  
Who ate THEM?  
A: FRED ate the BEA-ns.  
( H\* L ) ( L+H\* LH% )

- (16) Q: Well, what about FRED?  
What did HE eat?  
A: FRED ate the BEAns.  
( L+H\* LH% ) ( H\* LL% )

In these contexts, the main stressed syllables on both *Fred* and *the beans* receive a pitch accent, but a different one. In the former example, (15), there is a prosodic phrase on *Fred* made up of the pitch accent which Pierrehumbert calls H\*, immediately followed by an L boundary. There is another prosodic phrase having the pitch accent called L+H\* on *beans*, preceded by null or interpolated tone on the words *ate the*, and immediately followed by a boundary which is written LH%. (I base these annotations on Pierrehumbert and Hirschberg's [12, ex. 33] discussion of this example.)<sup>3</sup> In the second example (16) above, the

<sup>2</sup>For the purposes of this abstract, I am ignoring the distinction between the intonational phrase proper, and what Pierrehumbert and her colleagues call the "intermediate" phrase, which differ in respect of boundary tone-sequences.

<sup>3</sup>I continue to gloss over Pierrehumbert's distinction between "intermediate" and "intonational" phrases.

two tunes are reversed: this time the tune with pitch accent L+H\* and boundary LH% is spread across a prosodic phrase *Fred ate*, while the other tune with pitch accent H\* and boundary LL% is carried by the prosodic phrase *the beans* (again starting with an interpolated or null tone).<sup>4</sup>

The meaning that these tunes convey is intuitively very obvious. As Pierrehumbert and Hirschberg point out, the latter tune seems to be used to mark some or all of that part of the sentence expressing information that the speaker believes to be *novel to the hearer*. In traditional terms, it marks the "comment" – more precisely, what Halliday called the "rheme". In contrast, the L+H\* LH% tune seems to be used to mark some or all of that part of the sentence which expresses information which in traditional terms is the "topic" – in Halliday's terms, the "theme".<sup>5</sup> For present purposes, a theme can be thought of as conveying *what the speaker assumes to be the subject of mutual interest*, and this particular tune marks a theme as *novel to the conversation as a whole*, and as standing in a contrastive relation to the previous one. (If the theme is not novel in this sense, it receives *no* tone in Pierrehumbert's terms, and may even be left out altogether.)<sup>6</sup> Thus in (16), the L+H\* LH% phrase including this accent is spread across the phrase *Fred ate*.<sup>7</sup> Similarly, in (15), the same tune is confined to the object of the open proposition *ate the beans*, because the intonation of the original question indicates that eating beans *as opposed to some other comestible* is the new topic.<sup>8</sup>

## COMBINATORY PROSODY

The L+H\* LH% intonational melody in example (16) belongs to a phrase *Fred ate ...* which corresponds under the combinatory theory of grammar to a gram-

<sup>4</sup>The reason for notating the latter boundary as LL%, rather than L is again to do with the distinction between intonational and intermediate phrases.

<sup>5</sup>The concepts of theme and rheme are closely related to Grosz et al's [5] concepts of "backward looking center" and "forward looking center".

<sup>6</sup>Here I depart slightly from Halliday's definition. The present paper also follows Lyons [8] in rejecting Halliday's claim that the theme must necessarily be sentence-initial.

<sup>7</sup>An alternative prosody, in which the contrastive tune is confined to *Fred*, seems equally coherent, and may be the one intended by Jackendoff. I believe that this alternative is informationally distinct, and arises from an ambiguity as to whether the topic of this discourse is *Fred* or *What Fred ate*. It too is accepted by the rules below.

<sup>8</sup>Note that the position of the pitch accent in the phrase has to do with a further dimension of information structure within both theme and rheme, which one might identify as "focus". I ignore this dimension here.

matical constituent, complete with a translation equivalent to the open proposition  $\lambda x[(ate' x) fred']$ . The combinatory theory thus offers a way to derive such intonational phrases, using only the independently motivated rules of combinatory grammar, entirely under the control of appropriate intonation contours like  $L+H^* LH\%$ .<sup>9</sup>

It is extremely simple to make the existing combinatory grammar do this. We interpret the two pitch accents as functions over boundaries, of the following types:<sup>10</sup>

- (17)  $L+H^* := Theme/Bh$   
 $H^* := Rheme/Bl$

– that is, as functions over boundary tones into the two major informational types, the Hallidean “theme” and “rheme”. The reader may wonder at this point why we do not replace the category Theme by a functional category, say Utterance/Rheme, corresponding to its semantic type. The answer is that we do not want this category to combine with anything but a *complete* rheme. In particular, it must not combine with a function into the category Rheme by functional composition. Accordingly we give it a non-functional category, and supply the following special purpose prosodic combinatory rules:

- (18) Theme Rheme  $\Rightarrow$  Utterance  
 Rheme Theme  $\Rightarrow$  Utterance

We next define the various boundary tones as arguments to these functions, as follows:

- (19)  $LH\% := Bh$   
 $LL\% := Bl$   
 $L := Bl$

(As usual, we ignore for present purposes the distinction between intermediate- and intonational- phrase boundaries.) Finally, we accomplish the effect of interpolation of other parts of the tune by assigning the following polymorphic category to all elements bearing no tone specification, which we will represent as the tone 0:

- (20)  $0 := X/X$

<sup>9</sup>I am grateful to Steven Bird for discussions on the following proposal.

<sup>10</sup>An alternative (which would actually be closer to Pierrehumbert and Hirschberg’s own proposal to compositionally assemble discourse meanings from more primitive elements of meaning carried by each individual tone) would be to make the boundary tone the function and the pitch accent an argument.

Syntactic combination can then be made subject to the following simple restriction:

- (21) *The Prosodic Constituent Condition:* Combination of two syntactic categories via a syntactic combinatory rule is only allowed if their prosodic categories can *also* combine.

(The prosodic and syntactic combinatory rules need not be the same).

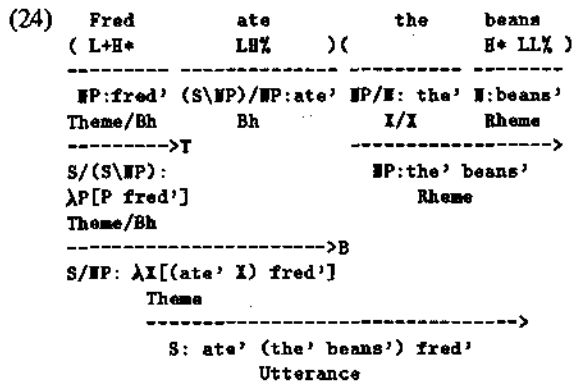
This principle has the sole effect of excluding certain derivations for spoken utterances that would be allowed for the equivalent written sentences. For example, consider the derivations that it permits for example (16) above. The rule of forward composition is allowed to apply to the words *Fred* and *ate*, because the prosodic categories can combine (by functional application):

- (22)
- |  |                      |  |
|--|----------------------|--|
| Fred<br>( L+H* )                           | ate ...<br>( LH% )   |  |
| NP:fred'<br>Theme/Bh                       | (S\NP)/NP:ate'<br>Bh |  |
| ----->T                                    |                      |  |
| S/(S\NP): $\lambda P[P fred']$<br>Theme/Bh |                      |  |
| ----->B                                    |                      |  |
| S/NP: $\lambda X[(ate' X) fred']$<br>Theme |                      |  |

The category  $X/X$  of the null tone allows intonational phrasal tunes like  $L+H^* LH\%$  tune to spread across any sequence that forms a grammatical constituent according to the combinatory grammar. For example, if the reply to the same question *What did Fred eat?* is *FRED must have eaten the BEANS*, then the tune will typically be spread over *Fred must have eaten ...*, as in the following (incomplete) derivation, in which much of the syntactic and semantic detail has been omitted in the interests of brevity:

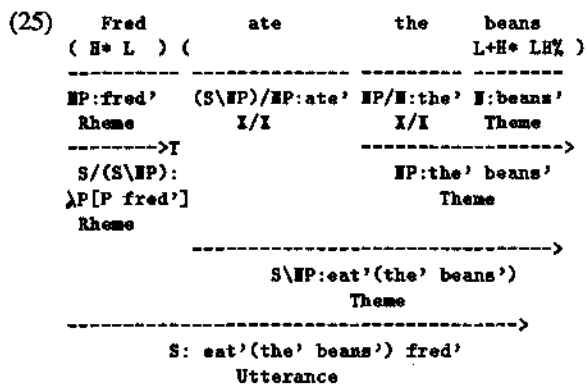
- (23)
- |          |           |         |           |  |
|----------|-----------|---------|-----------|--|
| Fred     | must      | have    | eaten ... |  |
| ( L+H* ) |           |         | ( LH% )   |  |
|          |           |         |           |  |
| NP       | (S\NP)/VP | VP/VPen | VPen/NP   |  |
| Theme/Bh | X/X       | X/X     | Bh        |  |
| ----->T  |           |         |           |  |
| Theme/Bh |           |         |           |  |
| ----->B  |           |         |           |  |
| Theme/Bh |           |         |           |  |
| ----->B  |           |         |           |  |
| Theme/Bh |           |         |           |  |
| ----->B  |           |         |           |  |
| Theme    |           |         |           |  |

The rest of the derivation of (16) is completed as follows, using the first rule in ex. (18):



The division of the utterance into an open proposition constituting the theme and an argument constituting the rheme is appropriate to the context established in (16). Moreover, the theory permits no *other* derivation for this intonation contour. Of course, repeated application of the composition rule, as in (23), would allow the L+H\* LH% contour to spread further, as in (*FRED must have eaten*)(*the BEANS*).

In contrast, the parallel derivation is forbidden by the prosodic constituent condition for the alternative intonation contour on (15). Instead, the following derivation, excluded for the previous example, is now allowed:

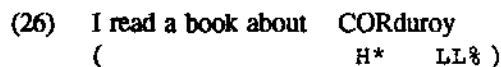


No other analysis is allowed for (25). Again, the derivation divides the sentence into new and given information consistent with the context given in the example. The effect of the derivation is to annotate the entire predicate as an L+H\* LH%. It is emphasised that this does *not* mean that the *tone* is spread, but that the whole constituent is marked for the corresponding discourse function — roughly, as contrastive given, or theme. The finer grain information that it is the object that is contrasted, while the verb is given, resides in the tree itself. Similarly, the fact that boundary sequences are associated with words at the lowest level of the derivation does not mean that they are *part of* the word, or specified in the lexicon, nor that the word is the entity that they are a boundary *of*. It is

prosodic phrases that they bound, and these also are defined by the tree.

All the other possibilities for combining these two contours on this sentence are shown elsewhere [17] to yield similarly unique and contextually appropriate interpretations.

Sentences like the above, including marked theme and rheme expressed as two distinct intonational/intermediate phrases are by that token unambiguous as to their information structure. However, sentences like the following, which in Pierrehumberts' terms bear a single intonational phrase, are much more ambiguous as to the division that they convey between theme and rheme:



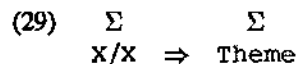
Such a sentence is notoriously ambiguous as to the open proposition it presupposes, for it seems equally appropriate as a response to any of the following questions:

- (27) a. What did you read a book about?
- b. What did you read?
- c. What did you do?

Such questions could in suitably contrastive contexts give rise to themes marked by the L+H\* LH% tune, bracketing the sentence as follows:

- (28) a. (I read a book about)(CORduroy)
- b. (I read)(a book about CORduroy)
- c. (I)(read a book about CORduroy)

It seems that we shall miss a generalisation concerning the relation of intonation to discourse information unless we extend Pierrehumberts theory very slightly, to allow *null* intermediate phrases, without pitch accents, expressing unmarked themes. Since the boundaries of such intermediate phrases are not explicitly marked, we shall immediately allow all of the above analyses for (26). Such a modification to the theory can be introduced by the following rule, which non-deterministically allows certain constituents bearing the null tone to become a theme:



The symbol Σ is a variable ranging over syntactic categories that are (leftward- or rightward- looking)

functions into  $S$ .<sup>11</sup> The rule is nondeterministic, so it correctly continues to allow a further analysis of the entire sentence as a single Intonational Phrase conveying the Rheme. Such an utterance is the appropriate response to yet another open-proposition establishing question, *What happened?*.)

With this generalisation, we are in a position to make the following claim:

- (30) The structures demanded by the theory of intonation and its relation to contextual information are the same as the surface syntactic structures permitted by the combinatory grammar.

A number of corollaries follow, such as the following:

- (31) Anything which can coordinate can be an intonational constituent, and *vice versa*.

## CONCLUSION

The pathway between phonological form and interpretation can now be viewed as in Figure 2:

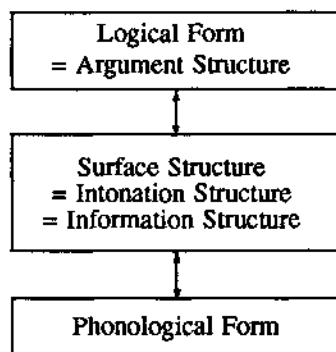


Figure 2: Architecture of a CCG-based Prosody

Such an architecture is considerably simpler than the one shown earlier in Figure 1. Phonological form maps via the rules of combinatory grammar directly onto a surface structure, whose highest level constituents correspond to intonational constituents, annotated as to their discourse function. Surface structure therefore subsumes intonational structure. It also subsumes information structure, since the translations of those surface constituents correspond to the entities and open propositions which constitute the topic or theme (if any) and the comment or rheme. These in

<sup>11</sup>The inclusion in the full grammar of further rules of type-raising in addition to the subject rule discussed above means that the set of categories over which  $\Sigma$  ranges is larger than it is possible to reveal in the present paper. (For example, it includes object complements). See the earlier papers and [17] for discussion.

turn reduce via functional application to yield canonical function-argument structure, or “logical form”.

There may be significant advantages for automatic spoken language understanding in such a theory. Most obviously, where in the past parsing and phonological processing have tended to deliver conflicting structural analyses, and have had to be pursued independently, they now are seen to be in concert. That is not to say that intonational cues remove all local structural ambiguity. Nor should the problem of recognising cues like boundary tones be underestimated, for the acoustic realisation in the fundamental frequency  $F_0$  of the intonational tunes discussed above is entirely dependent upon the rest of the phonology – that is, upon the phonemes and words that bear the tune. It therefore seems most unlikely that intonational contour can be identified in isolation from word recognition.<sup>12</sup>

What the isomorphism between syntactic structure and intonational structure *does* mean is that simply structured modular processors which use both sources of information at once can be more easily devised. Such an architecture may reasonably be expected to simplify the problem of resolving local structural ambiguity in both domains. For example, a syntactic analysis that is so closely related to the structure of the signal should be easier to use to “filter” the ambiguities arising from lexical recognition.

However, it is probably more important that the constituents that arise under this analysis are also semantically interpreted. The interpretations are directly related to the concepts, referents and themes that have been established in the context of discourse, say as the result of a question. These discourse entities are in turn directly reducible to the structures involved in knowledge-representation and inference. The direct path from speech to these higher levels of analysis offered by the present theory should therefore make it possible to use more effectively the much more powerful resources of semantics and domain-specific knowledge, including knowledge of the discourse, to filter low-level ambiguities, using larger grammars of a more expressive class than is currently possible. While vast improvements in purely bottom-up word recognition can be expected to continue, such filtering is likely to remain crucial to successful speech processing by machine, and appears to be characteristic of all levels of human processing, for both spoken and written language.

<sup>12</sup>This is no bad thing. The converse also applies: intonation contour effects the acoustic realisation of words, particularly with respect to timing. It is therefore likely that the benefits of combining intonational recognition and word recognition will be mutual.

## REFERENCES

- [1] Beckman, Mary and Janet Pierrehumbert: 1986, 'Intonational Structure in Japanese and English', *Phonology Yearbook*, 3, 255-310.
- [2] Chomsky, Noam: 1970, 'Deep Structure, Surface Structure, and Semantic Interpretation', in D. Steinberg and L. Jakobovits, *Semantics*, CUP, Cambridge, 1971, 183-216.
- [3] Curry, Haskell and Robert Feys: 1958, *Combinatory Logic*, North Holland, Amsterdam.
- [4] Dowty, David: 1988, Type raising, functional composition, and non-constituent coordination, in Richard T. Oehrle, E. Bach and D. Wheeler, (eds), *Categorial Grammars and Natural Language Structures*, Reidel, Dordrecht, 153-198.
- [5] Grosz, Barbara, Aravind Joshi, and Scott Weinstein: 1983, 'Providing a Unified Account of Definite Noun Phrases in Discourse', *Proceedings of the 21st Annual Conference of the ACL*, Cambridge MA, July 1983, 44-50.
- [6] Halliday, Michael: 1967, *Intonation and Grammar in British English*, Mouton, The Hague.
- [7] Jackendoff, Ray: 1972, *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge MA.
- [8] Lyons, John: 1977. *Semantics, vol. II*, Cambridge University Press.
- [9] Pareschi, Remo, and Mark Steedman. 1987. A lazy way to chart parse with categorial grammars, *Proceedings of the 25th Annual Conference of the ACL*, Stanford, July 1987, 81-88.
- [10] Pierrehumbert, Janet: 1980, *The Phonology and Phonetics of English Intonation*, Ph.D dissertation, MIT. (Dist. by Indiana University Linguistics Club, Bloomington, IN.)
- [11] Pierrehumbert, Janet, and Mary Beckman: 1989, *Japanese Tone Structure*, MIT Press, Cambridge MA.
- [12] Pierrehumbert, Janet, and Julia Hirschberg, 1987, 'The Meaning of Intonational Contours in the Interpretation of Discourse', ms. Bell Labs.
- [13] Prince, Ellen F. 1986. On the syntactic marking of presupposed open propositions. Papers from the Parasession on Pragmatics and Grammatical Theory at the 22nd Regional Meeting of the Chicago Linguistic Society, 208-222.
- [14] Selkirk, Elisabeth: *Phonology and Syntax*, MIT Press, Cambridge MA.
- [15] Steedman, Mark: 1985a. Dependency and Coordination in the Grammar of Dutch and English, *Language* 61.523-568.
- [16] Steedman, Mark: 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5, 403-439.
- [17] Steedman, Mark: 1989, Structure and Intonation, ms. U. Penn.
- [18] Vijay-Shankar, K and David Weir: 1990, 'Polynomial Time Parsing of Combinatory Categorical Grammars', *Proceedings of the 28th Annual Conference of the ACL*, Pittsburgh, June 1990.
- [19] Wittenburg, Kent: 1987, 'Predictive Combinators: a Method for Efficient Processing of Combinatory Grammars', *Proceedings of the 25th Annual Conference of the ACL*, Stanford, July 1987, 73-80.