April 1990

Structure and Intonation in Spoken Language Understanding

Mark Steedman
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/cis_reports

Recommended Citation


This paper is posted at ScholarlyCommons. https://repository.upenn.edu/cis_reports/515
For more information, please contact repository@pobox.upenn.edu.
Structure and Intonation in Spoken Language Understanding

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 Categorial Grammar. Interpretations of constituents at this level are in turn directly related to "information structures", or discourse-related notions of "theme", "rhemé", "focus" and "presupposition". Some simplifications appear to follow for the problem of integrating syntax and other high-level modules in spoken language systems.

Comments
Structure and Intonation In
Spoken Language Understanding

MS-CIS-90-23
LINC LAB 169

Mark Steedman

Department of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104

April 1990

To appear in: Proceedings of the 28th Annual Meeting of the Association for
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)

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 Categorial 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)

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-
lated to information structure. The theory that results can be viewed as in Figure 1:

![Figure 1: Architecture of Standard Metrical Phonology](image)

The involvement of two apparently uncoupled levels of structure in natural language grammar appears to complicate the path from speech to interpretation unreasonably, and to thereby threaten a number of computational applications in speech recognition and speech synthesis.

It is therefore interesting to observe that all natural languages include syntactic constructions whose semantics is also reminiscent of functional abstraction. The most obvious and tractable class are Wh-constructions themselves, in which exactly the same fragments that can be delineated by a single intonation contour appear as the residue of the subordinate clause. Another and much more problematic class of fragments results from coordinate constructions. It is striking that the residues of wh-movement and conjunction reduction are also subject to something like a “sense unit condition”. For example, strings like “in ten prefer corduroy” are not conjoinable:

(5) *Three cats in twenty like velvet, and in ten prefer corduroy.

Since coordinate constructions have constituted another major source of complexity for theories of natural language grammar, and also offer serious obstacles to computational applications, it is tempting to think that this conspiracy between syntax and prosody might point to a unified notion of structure that is somewhat different from traditional surface constituency.

COMBINATORY GRAMMARS.

Combinatory Categorial Grammar (CCG, [16]) is an extension of Categorial Grammar (CG). Elements like verbs are associated with a syntactic “category” which identifies them as functions, and specifies the type and directionality of their arguments and the type of their result:

(6) prefers := (S\NP)/NP : prefer'

The category can be regarded as encoding the semantic type of their translation, which in the notation used here is identified by the expression to the right of the colon. Such functions can combine with arguments of the appropriate type and position by functional application:

(7) Mary prefers corduroy
    NP (S\NP)/NP NP
    --------------<
    S\NP
    -------------
    S

Because the syntactic types are identical to the semantic types, apart form directionality, the derivation also builds a compositional interpretation, (prefer' corduroy') mary', and of course such a “pure” categorial grammar is context free. Coordination might be included in CG via the following rule, allowing constituents of like type to conjoin to yield a single constituent of the same type:

(8) X conj X ⇒ X

(9) I loath and detest velvet
    -- -------------- ----
    NP (S\NP)/NP conj (S\NP)/NP NP
    -------------------------------
    (S\NP)/NP

(The rest of the derivation is omitted, being the same as in (7).) In order to allow coordination of contiguous strings that do not constitute constituents, CCG generalises the grammar to allow certain operations on functions related to Curry’s combinators [3]. For example, functions may nondeterministically compose, as well as apply, under the following rule:

(10) Forward Composition:
    X/Y : F Y/Z : G ⇒ X/Z : λx F(Gx)

The most important single property of combinatory rules like this is that they have an invariant semantics. This one composes the interpretations of the functions that it applies to, as is apparent from the right hand side of the rule.\(^1\) Thus sentences like I suggested,

\(^1\)The rule uses the notation of the λ-calculus in the semantics, for clarity. This should not obscure the fact that it is functional composition itself that is the primitive, not the λ operator.
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:

\[(11)\text{ suggested and would prefer} \]
\[\begin{array}{c}
(S\text{\ NP})/\text{ NP} \\
\text{ conj} \\
(S\text{\ NP})/\text{ VP} \\
\text{ VP/NP} \\
\text{ \longrightarrow \ } \text{B} \\
\end{array}
\]

\[(S\text{\ NP})/\text{ NP}\]

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):

\[(12)\text{ Subject Type-raising:}\]
\[\begin{array}{c}
NP : y \Rightarrow S/(S\text{\ NP}) : \lambda F \mathit{Fy}
\end{array}
\]

\[(13)\text{ I suggested and Mary prefers}\]
\[\begin{array}{c}
\text{ NP} \\
(S\text{\ NP})/\text{ NP} \\
\text{ conj} \\
(S\text{\ NP})/\text{ NP} \\
\text{ \longrightarrow \ } \text{T} \\
S/(S\text{\ NP}) \\
\text{ \longrightarrow \ } \text{T} \\
S/(S\text{\ NP}) \\
\end{array}
\]

\[\begin{array}{c}
\text{ \longrightarrow \ } \text{B} \\
S/\text{ NP} \\
\end{array}
\]

\[(S\text{\ NP})/\text{ NP}\]

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:

\[(14)\text{ Mary prefers corduroy}\]
\[\begin{array}{c}
\text{ NP} \\
(S\text{\ NP})/\text{ NP} \\
\text{ \longrightarrow \ } \text{T} \\
S/(S\text{\ NP}) \\
\text{ \longrightarrow \ } \text{B} \\
S/\text{ NP} \\
\end{array}
\]

\[(S\text{\ NP})/\text{ NP}\]

(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*.)

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 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 intoned 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].
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.* 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 right-hand 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 BEAns.
   (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.) In the second example (16) above, the 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).4

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".5 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.)6 Thus in (16), the L+H* LH% phrase including this accent is spread across the phrase *Fred ate*. 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.8

**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-

---

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

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

6Here 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.

7An 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.

8Note 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 me 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 \( L+H^* \ LH^\% \). 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:1°

\[
\begin{align*}
L+H^* & := \text{Theme/Bh} \\
H^* & := \text{Rheme/Bl}
\end{align*}
\]

— that is, as functions over boundary tones into the two major informational types, the Hallidayan “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 theme. 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:

\[
\text{(18)} \quad \text{Theme} \quad \text{Rheme} \Rightarrow \text{Utterance} \\
\text{Rheme} \quad \text{Theme} \Rightarrow \text{Utterance}
\]

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

\[
\begin{align*}
LH^\% & := \text{Bh} \\
LI^\% & := \text{Bl} \\
L & := \text{Bl}
\end{align*}
\]

(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 \( 0 \):

\[
\text{(20)} \quad 0 := X/X
\]

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

\[
\text{(21)} \quad \text{The Prosodic Constituent Condition: Combination of two syntactic categories via a syntactic combinatory rule is only allowed if their prosodic categories can also combine.}
\]

(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):

\[
\text{(22)} \quad \text{Fred} \quad \text{ate} \ldots \quad (L+H^* \quad LH^\%) \\
\text{NP:fred'} \quad \text{S/} (S\NP)\NP: \text{ate'} \\
\text{Theme/Bh} \quad \text{Bh} \\
\text{-----------------------------}\Rightarrow T \\
\text{S/} (S\NP): \lambda \text{X[(ate' \ X) fred']} \\
\text{Theme/Bh} \quad \text{---------------}\Rightarrow B \\
\text{S/} \text{NP:} \text{X}[(\text{ate' \ X) fred'}] \\
\text{Theme}
\]

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:

\[
\text{(23)} \quad \text{Fred} \quad \text{must} \quad \text{have} \quad \text{eaten} \ldots \quad (L+H^* \quad LH^\% ) \\
\text{NP} \quad \text{S/} (S\NP)/VP \quad \text{VP/VPen} \quad \text{VPen/MP} \\
\text{Theme/Bh} \quad \text{X/X} \quad \text{X/X} \quad \text{Bh} \\
\text{---------------}\Rightarrow T \\
\text{Theme/Bh} \quad \text{---------------}\Rightarrow B \\
\text{Theme/Bh} \quad \text{---------------}\Rightarrow B \\
\text{Theme}
\]

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 r... that 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

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
functions into S. The rule is nondeterministic, so it
correctly continues to allow a further analysis of the
entire sentence as a single Intonational Phrase convey-
ing 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 in-
tonation and its relation to contextual infor-
mation 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 inter-
pretation can now be viewed as in Figure 2:

```
Logical Form
= Argument Structure

Surface Structure
= Intonation Structure
= Information Structure

Phonological Form
```

*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 con-
stituents correspond to intonational constituents, an-
notated as to their discourse function. Surface struc-
ture therefore subsumes intonational structure. It also
subsumes information structure, since the transla-
tions of those surface constituents correspond to the en-
tities and open propositions which constitute the topic
or theme (if any) and the comment or rhyme. These in-
turn reduce via functional application to yield canoni-
cal 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 phono-
logical processing have tended to deliver conflicting
structural analyses, and have had to be pursued inde-
pendently, they now are seen to be in concert. That is
not to say that intonational cues remove all local struc-
tural ambiguity. Nor should the problem of recognis-
ing cues like boundary tones be underestimated, for
the acoustic realisation in the fundamental frequency
$F_0$ of the intonational tunes discussed above is en-
tirely dependent upon the rest of the phonology –
that is, upon the phonemes and words that bear the
tune. It therefore seems most unlikely that intona-
tional contour can be identified in isolation from word
recognition.12

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 am-
biguity 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 am-
biguities 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 di-
rectly 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 en-
tities 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 dis-
course, to filter low-level ambiguities, using larger
grammars of a more expressive class than is cur-
rently possible. While vast improvements in purely
bottom-up word recognition can be expected to con-
tinue, such filtering is likely to remain crucial to suc-
cessful speech processing by machine, and appears to
be characteristic of all levels of human processing,
for both spoken and written language.

---

11The 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.

12This 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 combing
intonational recognition and word recognition will be mutual.
REFERENCES


