Intonation and Syntax in Spoken Language Systems

*Preliminary Draft**
Mark Steedman
Computer and Information Science, U.Penn.

Abstract

The present paper argues that the notion of "intonational structure" as formulated by Pierrehumbert, Selkirk, and others, can be subsumed under the generalised notion of syntactic surface structure that emerges from a theory of grammar based on a "Combinatory" extension to Categorial Grammar. According to this theory, the syntactic structures and the intonation structures of English are identical, and have the same grammar. Some simplifications appear to follow for the problem of integrating syntax and other high-level modules in spoken language systems.

Phrasal intonation is notorious for structuring the words of spoken utterances into groups which frequently violate orthodox notions of constituency. For example, the normal prosody for the answer (b) to the following question (a) imposes the intonational constituency indicated by the brackets (stress is indicated by capitals, and an indication of the perceived contour is given):

(1) a. I know that brassicas are a good source of minerals, but what are LEGumes a good source of?
b. (LEGumes are a good source of) VITamins.

Such a grouping cuts right across the traditional syntactic structure of the sentence. The presence 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.

Nevertheless, intonational structure is strongly constrained by meaning. Contours imposing bracketings like the following are not allowed:

(2) # Three doctors (in ten prefer cats)

Halliday [5] seems to have been the first to identify this phenomenon, which Selkirk [12] has called the "Sense Unit Condition", and to observe that this constraint seems to follow from the *function* of phrasal intonation, which is to convey distinctions of focus, information, and propositional attitude towards entities in the discourse. These entities are more diverse than mere nounphrase or propositional referents, but they do not include such non-concepts as "in ten prefer cats."

One discourse category that they *do* include is what E. Prince [11] calls "open propositions". Open propositions are most easily understood as being that which is introduced into the discourse context by a Wh-question. So for example the question in (1), *What are legumes a good source of?* introduces an open proposition which it is most natural to think of as a functional *abstrac-*

*I am grateful to Julia Hirschberg, Aravind Joshi, Mitch Marcus, Janet Pierrehumbert, and Bonnie Lynn Webber for comments and advice. They should not be held responsible for any errors in the translation of their advice into the present framework. The present paper is a revised version of one that appears in the preliminary papers of the March 1989 AAAI symposium on Spoken Language Systems, Stanford CA. An expanded version of the present paper appears as CIS Technical Report No. 89-20, University of Pennsylvania. The research was supported by DARPA grant no. N0014-85-K0018.

tion, which would be written as follows in the notation of the λ-calculus:

(3) $\lambda x[good'(source'\ x)\ legumes']$

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

(4) $good'(source'\ vitamins')legumes'$

It is the presence of the above open proposition that makes the intonation contour in (1) felicitous. (I am not claiming that its presence ***determines*** this response, nor that its presence is necessary for interpreting the response.)

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. But another and much more problematic class are the fragments that result 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 cats" are not conjoinable:

(5) *Three doctors in ten prefer cats,
and in twenty eat carrots.

While coordinate constructions have constituted another major source of complexity for natural language understanding by machine, 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, [14]) 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) *eats* :- `(S\NP)/NP: eat'`

The category can be regarded as encoding the semantic type of their translation. Such functions can combine with arguments of the appropriate type and position by functional application:

```
(7) Harry      eats      apples
    ------  ---------  ------
      NP    (S\NP)/NP    NP
            -------------->
                 S\NP
      ---------------<
               S
```

Because the syntactic functional type is identical to the semantic type, apart from directionality, this derivation also builds a compositional interpretation, *eats'apples'harry'*, and of course such a "pure" categorial grammar is context free. Coordination might be included in CG via the following rule, allowing any constituents of like type, including functions, to form a single constituent of the same type:

(8) $X \quad conj \quad X \quad \Rightarrow \quad X$

```
(9) I    cooked    and     ate     a frog
    -- --------- ---- ---------- ------
    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 *compose*, as well as apply, under the following rule

(10) Forward Composition:
$X/Y : F \quad Y/Z : G \quad \Rightarrow \quad X/Z : \lambda x\ F(Gx)$

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 cooked, and might eat, the beans* 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, assuming that a semantics is also provided for the coordination rule.

```
(11)    cooked   and  might     eat
      --------- ---- --------- -----
      (S\NP)/NP conj (S\NP)/VP VP/NP
                     --------------->B
                        (S\NP)/NP
       --------------------------&
                (S\NP)/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 cooked, and you ate, the legumes.* 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) Subject Type-raising:

$$NP : y \;\Rightarrow\; S/(S\backslash NP) : \lambda F\ Fy$$

```
(13)      I      cooked   and   you       ate
      -------- --------- ---- -------- ---------
         NP    (S\NP)/NP conj    NP    (S\NP)/NP
      -------->T              -------->T
      S/(S\NP)                S/(S\NP)
      ------------------>B    ------------------>B
              S/NP                   S/NP
              ---------------------------&
                          S/NP
```

This theory has been explicitly addressed to a wide range of coordination phenomena in a number of languages [4], [13].

Intonation in a CCG.

Inspection of the above examples shows that Combinatory grammars embody an unusual view of surface structure, according to which strings like *Betty might eat* are constituents. In fact, according to this view, surface structure is a much more ambiguous affair than is generally realised, for they must also be possible constituents of non-coordinate sentences like *Betty might eat the mushrooms*, as well. (See [7] and [15] for a discussion of the obvious problems that this fact engenders for parsing written text.) An entirely unconstrained combinatory grammar would in fact allow more or less any bracketing on a sentence. However, the actual grammars we write for configurational languages like English are heavily constrained by local conditions. (An example would be a condition on the composition rule that is tacitly assumed here, forbidding the variable Y to be instantiated as NP, thus excluding constituents like $*$[*eat the*]$_{VP/N}$).

The claim of the present paper is simply that particular surface structures that are induced by the specific combinatory grammar that was introduced to explain coordination in English are identical to the intonational structures that are required to specify the possible intonation contours for those same sentences of English. More specifically, the claim is that that in spoken utterance, intonation largely determines *which* of the many possible bracketings permitted by the combinatory syntax of English is intended, and that the interpretations of the constituents are related to distinctions of focus among the concepts and open propositions that the speaker has in mind. Thus, whatever problems for parsing written text arise from the profusion of equivalent alternative surface structures engendered by this theory, these "spurious" ambiguities seem to be to a great extent resolved by prosody in spoken language. The theory therefore offers the possibility that phonology and parsing can be merged into a single unitary process.

The proof of this claim lies in showing that the rules of combinatory grammar can be annotated with intonation contour schemata, which limit their application in spoken discourse, and to

[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.

showing that the major constituents of intonated utterances like (1)b, under the analyses that these rules permit correspond to the focus structure of the context to which they are appropriate, such as (1)a.

I shall use a notation which is based on the theory of Pierrehumbert [8], as modified in more recent work by Selkirk [12], Beckman and Pierrehumbert [1], [9], and Pierrehumbert and Hirschberg [10]. I have tried as far as possible to take my examples and the associated intonational annotations from those authors.

I follow Pierrehumbert in assuming two abstract pitch levels, and three types of tones, as follows. There are two phrasal tones, written H and L, denoting high or low "simple" tones — that is, level functions of pitch against time. There are also two boundary tones, written H% and L%, denoting an intonational phrase-final rise or fall. Of Pierrhumberts six pitch accent tones, I shall only be concerned with two, the H* accent and the L+H*. The phonetic or acoustic realisation of pitch accents is a complex matter. Roughly speaking, the L+H* pitch accent that is extensively discussed below in the context of the L+H* LH% melody generally appears as a maximum which is preceded by a distinctive low level, and peaks *later* than the corresponding H* pitch accent when the same sequence is spoken with the H* L melody that goes with "new" information, and which is the other melody considered below.

In the more recent versions of the theory, Pierrehumbert and her colleagues distinguish *two* levels of prosodic phrase that include a pitch accent tone. They are the intonational phrase proper, and the "intermediate phrase". Both end in a phrasal tone, but only intonational phrases have additional boundary tones H% and L%. Intermediate phrases are bounded on the right by their phrasal tone alone, and do not appear to be characterised in F_0 by the same kind of final rise or fall that is characteristic of true intonational phrases. The distinction does not play an active role in the prosent account, but I shall follow the more recent notation of prosodic phrase boundaries in the examples, without further comment on the distinction.

There may also be parts of prosodic phrases where the fundamental frequency is merely interpolated between tones, notably the region between pitch accent and phrasal tone, and the region before a pitch accent. In Pierrehumbert's notation, such substrings bear no indication of abstract tone whatsoever.

A crucial feature of this theory for present purposes is that the position and shape of a given pitch accent in a prosodic phrase, and of its phrase accent and the associated right-hand boundary, are essentially invariant. If the constituent is very short – say, a monosyllabic nounphrase – then the whole intonational contour may be squeezed onto that one syllable. If the constituent is longer, then the pitch accent will appear at its left edge, the phrasal tone and boundary tone if any will appear at its right edge, and the intervening pitch contour will merely be interpolated. In this way, the 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.

Consider for example the prosody of the sentence *Fred ate the beans* in the following pair of discourse settings, which are adapted from Jackendoff [6, pp. 260]:

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

(15) 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 (14), the pitch accent contour on *Fred* is H*, while that on *beans* is L+H*. (I base these annotations on Pierrehumbert and Hirschberg's [10, ex. 33] discussion of this example.)

In the second example (15) above, the pitch accents are reversed: this time *Fred* is L+H* and *beans* is H*. The assignment of these tones seem to reflect the fact that (as Pierrehumbert and Hirschberg point out) H* is used to mark information that the speaker believes to be *new to the* *hearer.* In contrast, L+H* seems to be used to mark information which the current speaker knows to be given to the hearer (because the current hearer asked the original question), but which constitutes a novel topic of conversation for the speaker, standing in a contrastive relation to some *other* given information, constituting the previous topic. (If the information were merely given, it would receive *no* tone in Pierrehumbert's terms — or be left out altogether.) Thus in (15), the L+H* LH% phrase including this accent is spread across the phrase *Fred ate.*[2] Similarly, in (14), 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.

Syntax-driven Prosody.

The L+H* LH% intonational melody in example (15) belongs to a phrase *Fred ate ...* which corresponds under the combinatory theory of grammar to a grammatical constituent, complete with a translation equivalent to the open proposition $\lambda x[(ate'\ x)\ fred']$. The combinatory theory thus offers a way to assign intonation contours entirely under the control of independently motivated rules of grammar. In particular, the forward composition rule (10) offers the possibility of limiting the construction of non-standard constituents according to the intonation contours on the composed elements.

I show elsewhere that this effect can be achieved using a simple annotation of the composition rule capturing the injuction "Don't compose across an intonational phrase or intermediate phrase boundary". Application is not constrained by intonation, and all rules mark their result with the concatenation of the intonation contour on their inputs. The annotated composition rule correctly allows the derivation of the non-standard constituent *Fred ate* in example (15), where it is marked with L+H* LH%, because this string does not include an internal phrase boundary. It will also accept strings in which the same contour is spread over more lengthy open propositions, such as *Fred must have eaten ...*, as in *(FRED must have eaten)(the BEAns)*. However, the same rule correctly *forbids* the derivation of such a constituent in example (14), because *Fred* is marked as H*L, ending in an intermediate phrase boundary, and thus cannot compose with the material to its right. Other examples considered by Jackendoff are also accepted by these rules, to yield only the contextually appropriate interpretations.

Conclusion.

According to the present theory, the pathway between phonological form and interpretation is much simpler than has been thought up till now. Phonological Form maps directly onto Surface Structure, via rules of combinatory grammar annotated with abstract intonation contours. Surface Structure is identical to intonational structure, and maps directly onto Focus Structure, in which focussed and backgrounded entities and open propositions are represented by functional abstractions and arguments. Such structures reduce to yield canonical Function-Argument Structures. The proposal thus represents a return to the architecture proposed by Chomsky [2] and Jackendoff [6]. The difference is that the concept of surface structure has changed. It now really is *only* surface structure, supplemented by "annotations" which do nothing more than indicate the information structural status and intonational tune of *constituents* at that level.

While many problems remain, both in parsing written text with grammars that include associative operations, and at the signal-processing end, the benefits for automatic spoken language understanding are likely to be significant. Most obviously, where in the past parsing and phonological processing have delivered conflicting structural analyses, and have had to be pursued independently, they now are seen to be in concert. Processors can therefore be devised which use both sources of information at once, thus simplifying both problems. (For example, intonation

[2] 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 is accepted by the present rules.

may largely determine syntactic structure in the present sense. And 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.) What is likely to be more important in the long run, however, is that the constituents that arise under this analysis are also semantically interpreted. The paper has argued that these 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. The shortening and simplification of the path from speech to these higher levels of analysis offers the possibility of using those probably more effective resources to filter the proliferation of low level analyses as well.

References

[1] Beckman, Mary and Janet Pierrehumbert: 1986, 'Intonational Structure in Japanese and English', *Phonology Yearbook*, 3, 255-310.

[2] Chomsky, Noam: 1970, 'Remarks on nominalisation', in R. Jacobs and P. Rosenbaum, *Readings in English Transformational Grammar*, Ginn, Waltham, MA, pp. 184-221.

[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] Halliday, Michael: 1967, *Intonation and Grammar in British English*, Mouton, The Hague.

[6] Jackendoff, Ray: 1972, *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge MA.

[7] 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.

[8] Pierrehumbert, Janet: 1980, *The Phonology and Phonetics of English Intonation*, Ph.D dissertation, MIT. (Distributed by Indiana University Linguistics Club, Bloomington, IN.)

[9] Pierrehumbert, Janet, and Mary Beckman: 1989, *Japanese Tone Structure*, MIT Press, Cambridge MA.

[10] Pierrehumbert, Janet, and Julia Hirschberg, 1987, 'The Meaning of Intonational Contours in the Interpretation of Discourse', ms. Bell Labs.

[11] 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.

[12] Selkirk, Elisabeth: *Phonology and Syntax*, MIT Press, Cambridge MA.

[13] Steedman, Mark: 1985a. Dependency and Coordination ... Language 61.523-568.

[14] Steedman, Mark: 1987. Combinatory grammars and parasitic gaps. NL&LT, 5, 403-439.

[15] 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.