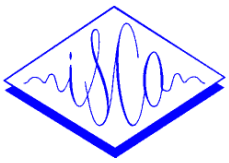


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Intonation and the Request/Question Distinction

Elizabeth A. Hinkelman[1]

Deutsches Forschungszentrum fuer Kuenstliche Intelligenz
Stuhlsatzenhausweg 3, D-W-6600 Saarbruecken, Germany

Abstract

Both linguistic intuitions and psychological evidence argue that intonation plays a role in the interpretation of spoken English utterances. The work presented here is a pilot study for exploration of this role. We suggest that intonation is one of several **extrapositional** linguistic features of an utterance which, along with extralinguistic information, signal that the content of an utterance is a question, assertion, request, greeting, or some other speech act. We present our model for recognition of speaker intentions. We show, using a forced-choice paradigm, that the request/question distinction can be intonationally disambiguated. We discuss the intonational features that may be responsible, and indicate the steps necessary to making use of these results in automated speech act recognition.

1. Intonation and Speaker Intentions

Consider the sentence

(1) Do you know the time?

Under appropriate circumstances, a speaker may utter this sentence to ask what time it is, as a yes/no question about the hearer's information, as an offer to tell someone the time, or even as a reminder that it is late. Each one of these different uses can be viewed as a different action, or speech act [Austin1963a]. Syntax-based models of speech act recognition [Gazdar1979a, Sadock1974a] have been popular in linguistics, but do not address the problem of dependence on circumstances. Artificial intelligence models [Allen1983a, Cohen1990a] address the problem of dependence on circumstances, but until recently [Hinkelman1989a] they have ignored surface properties of utterances. None have incorporated intonation systematically.

Bolinger [Bolinger1982a] refutes syntax-minded attempts to link intonation and speech acts, cautioning against the assumption that there is a one-to-one mapping from intonational contours to "anything". He does this by picking celebrated contours, then providing several precise contexts which each yield a different interpretation for the contour. Psycholinguistic success has also been limited. [Geluykens1988a] failed to find a simple mapping in either direction between final rise contour and polar questions, and concluded from this that intonation plays no role in polar questions. [Beun1989a] found intonation irrelevant for identifying declarative-syntax questions in Dutch. Nonetheless, [Nusbaum1991a] have been able to show that subjects who hear a list of unrelated declarative-syntax sentences will reliably identify those having final rises as questions. We hypothesize that when syntactic and contextual factors are held equal, intonation can be shown to have a role in disambiguating a wider range of speech acts.

2. A Model of Intention Recognition

We model the speech act recognition process in three stages [Hinkelman1989b] [Hinkelman1989a] [Hinkelman1991a]. The first stage makes use of *extrapositional* linguistic information, that is, information sources such as sentence mood, sentential adverbs, and modal auxiliaries, which are not incorporated in any systematic way into most current semantic theories, and which are often viewed as external to the sentence proper. This extrapositional information is used to generate a set of possible speech act interpretations. In the second stage, these interpretations are verified against hearer's information about the situation, eliminating interpretations inconsistent with previously stated information and inferences about the speaker's goals and intentions. For example, a question is eliminated in the case where the speaker is already believed to know the answer. The third stage, when necessary, involves extended reasoning about the speaker's plans. All of these stages have been implemented as artificial intelligence programs; the first two for both English and for German.

If intonational features were indeed a factor in the recognition of speaker intentions, they would be incorporated into extrapositional information used in the first stage of this model. The first stage employs rules that associate fragments of linguistic feature structure with partial speech act interpretations. If an utterance structure unifies with the fragment, a partial interpretation is built. Thus, an utterance like

(2) Can you flip the omelet?

could be recognized as a possible request because it fits the rule

surface:
  category: sentence
  mood: ynQuestion
  voice: active
  subject:
    surface:
      category: nounPhrase
      head: "you"
  auxilliaries: {"can" "could" "will" "would" "might"}
  verb: +action
speechAct: **requestAct** *action:* **V**(*reference.action*)
  V speechAct

This rule interprets *Can you...?* questions as requests, looking for the subject "you" and any of these modal verbs. Note that the semantic interpretation of main verb is used as the action being requested, and must therefore be one marked as an action verb. The rule also allows for a weaker interpretation that unifies with the output of other rules. In the absence of intonation, the sentence would at the end of the first stage still be ambiguous between a request and a question interpretation, hopefully to be disambiguated by context in the second stage[2].

[1]This research was supported by the United States National Science Foundation grant IRI-9109914, by the University of Chicago laboratory for Speech, Division of Social Sciences (Howard Nusbaum, Director), and Language Laboratory and Archives, Division of Humanities (Director Karen Landahl). Karen Deaton performed most of the experimental work. The author would also like to thank Gerald Sadock, Boaz Keyser, David MacNeill, and Karl-Eric McCullough for valuable discussions.

10.21437/ICSLP.1992-351

The general question rule is:

surface:
category: sentence
mood: ynQuestion
speechAct: **askAct** *proposition:* **V**(*reference*)
V speechAct

To incorporate intonation into this stage would be to treat it as partial rather than total evidence for the speaker's intention, susceptible to override by other linguistic features and context. Any particular intonational structure could also be evidence for several different intentions. We thereby avoid Bolinger's one-to-one fallacy.

3. Intonation and Stress

If there are intonational correlates of speaker intentions, the next problem is to identify and represent them. The ideal situation for our theory would be that intonation perception yields discrete feature values attached at the top of, or to relevant parts of, the parse tree, or some compositional companion structure[3].

One pitch theory which has this property is that of [Pierrehumbert1990a], in which pitch accents (either high and low) signal salience of the corresponding semantic units, and phrase accents and boundary tones signal the relationship of the unit to adjacent ones. Loudness and duration are also factors [LeHiste1970a], but we discuss only pitch here.

Recall that [Nusbaum1991a] have shown that subjects who hear a list of unrelated declarative-syntax sentences will reliably identify those having final rises as questions. They found that processing was much slower for these utterances than simple declarative utterances. Our model would account for Broiher's results with rules that say that declarative mood is evidence for a statement, and that rising intonation is evidence for a question, falling for a statement. Using Pierrehumbert's pitch notation, we have

surface:
category: sentence
mood: declarative
speechAct: **informAct** *proposition:* **V**(*reference*)
V speechAct

surface:
category: sentence
phrase_accent: H
boundary_tone: H%
speechAct: **askAct** *proposition:* **V**(*reference*)

surface:
category: sentence
phrase_accent: L
boundary_tone: L%
speechAct: **informAct** *proposition:* **V**(*reference*)

The statements match the first and third rule; the questions the first and second. The delayed response time for the questions could be explained by a garden path effect: the mood rule matches early in the sentence, but the intonation rule does not get triggered until the sentence is complete. An alternative explanation for the delayed response time would be that since the first rule prefers an inform interpretation, it partially conflicts with the second rule. Nusbaum views the delay as due to conflict of evidence.

4. Two Perception Studies

We now examine the request/question distinction, for which intonational cues are less obvious. We present two pilot studies investigating sentences having the syntax of yes-no questions, but whose content permits both a request interpretation and a question interpretation. We refer to this class of sentences as "pivot sentences". The hypothesis tested in these studies is that, other factors equal, intonational features have a measureable effect on hearer judgements of the request/question distinction. We refer to the property of being influenced by intonation as "pivoting".

The experimentors first chose ten specific situations that would be familiar to university subjects. They constructed a pivot sentence appropriate to each situation, and revised it to reduce the number of unvoiced consonants for a clear f0 contour. Resulting sentences include, for example, "Could we have nine bagels?" and "Has the dog been let out?" The sentences were recorded twice by experimentors. In one recording, each sentence was preceeded by a sentence setting the context for a question, and uttered with intent to produce a question. Care was taken to avoid using contrastive stresses, since this would introduce a dimension of variability: the questioned entity ceases to be the entire proposition, and instead is the item receiving the stress. In the other recording, an initial sentence set a context appropriate to a request, and the intention was to produce a request. The two recordings are designated as pivot question (PQ) and pivot request (PR) in the table of stimulus classes below.

Table 1: Stimulus Classes, Study 1

count	reps	class	syntax	text	intention
5	12	I	imperative	request	request
5	12	R	yes/no	request	request
10	6	PR	yes/no	pivot	request
10	6	PQ	yes/no	pivot	question
5	12	Q	yes/no	question	question
5	12	QC	declarative	question	question

The task was a forced choice in which subjects were asked whether they would "DO" or "SAY" their response to an utterance. In order to provide subjects with clear instances of these response categories, the imperative syntax class (I) and simple question (Q) were used. Simple questions were sentences with yes/no syntax and texts that are much more plausible as questions than as requests ("Can those snakes really fly?"). This category was then balanced by a category of syntactic questions which would typically be used as requests (R). The imperative category was balanced with the category having declarative syntax but question intent (QC) and high-rise contour. Because it is known to be recognized reliably when contrasted with statements, this category serves as a reference for the overall reliability of the task.

The six categories of stimuli are shown in Table 1, along with the number of stimuli in that category and the number of repetitions with which they were heard. Subjects heard each pivot stimulus six times and each nonpivot stimulus twelve times, so that all texts occurred with the same frequency. The stimuli were presented to subjects by a laboratory computer in twelve blocks of thirty, selected randomly but balanced within the block. Subjects were instructed to listen to the entire sentence and to press a labelled key according to whether the speaker would expect the hearer to DO or to SAY something in response. The response keys were balanced for handedness. Both subject responses and response times were recorded.

4.1. Results

The prediction of principal interest was that the pivot sentences would be interpreted as reqests or questions on the basis of intonation.

Nine subjects categorized the non-pivot stimuli with great accuracy, including both so-called "indirect speech act" classes R and QC. It is worth noting that R and QC differed significantly from each other in response time: R grouped with the pivots, faster even than Q and I, whereas QC were by far the slowest. This is unsurprising, because QC inverts the sentence type's normal function with a strong final-position rise, effectively a garden path.

Subjects categorized classes PQ and PR with significantly ($p<.01$, see plot 1.) less accuracy overall, and (correct responses only) did so with the greatest speed (see plot 2). There was no overall bias for or against the "indirect" reading, but most individual stimuli proved biased. They had correct PQ and PR responses in complementary distribution, with a striking exception. PQ and PR 5 both got correct responses with a frequency well above chance, pivoting as predicted. (see plot 3). The text was "Will you read those books?"

4.2. A Second Perception Study

In order to establish the reproduceability of the pivot pattern, we tested fifteen additional sentences. These sentences were chosen to be lexically as neutral as possible while still including an explicit action description. Five of the original pivot sentences were also included, and the same presentation method used.

Table 2: Stimulus Classes, Study 2

count	reps	class	syntax	text	intention
5	12	R	yes/no	request	request
20	6	PR	yes/no	pivot	request
20	6	PQ	yes/no	pivot	question
5	12	QC	declarative	question	question

Of the fifteen new sentences, three (*) showed pivoting effects comparable to that of sentence 5, that is, averaged four or more correct responses out of six for both PQ and PR stimuli over all twelve subjects. Seven (+) more averaged above 50% correct in both directions. This argues for hearer use of intonation in disambiguation for a certain range of cases.

Table 3: Per Cent Correct Responses, Study 2

Stim	4+	5*	6	9	10	11*	12*	13+	14+	15+
PQ	65	67	26	22	78	67	72	57	50	76
PR	69	86	89	89	28	85	76	86	88	54

Stim	16	17	18+	19	20+	21+	22	23+	24	25*
PQ	15	40	60	39	75	57	51	57	29	68
PR	94	85	75	82	60	86	60	68	68	85

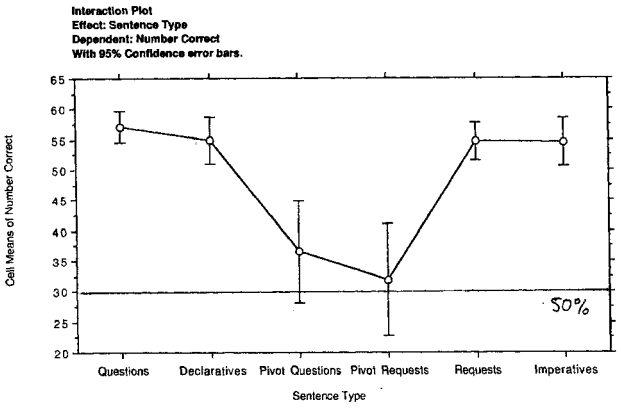
We also note that there appears to be strong individual variation in subject sensitivity to intonation; approximately one subject in four performed nearly perfectly.

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
Subject	8	164.000	20.500		
Sentence Type	5	5471.722	1094.344	17.294	.0001
Sentence Type * Subject	40	2531.111	63.278		

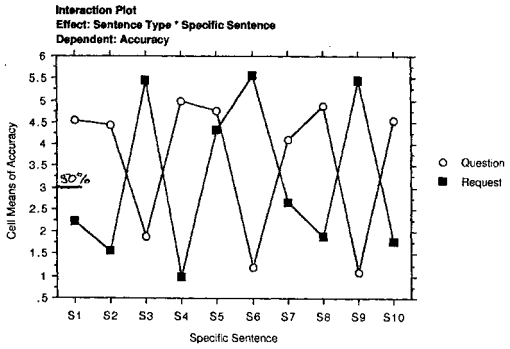
Dependent: Number Correct

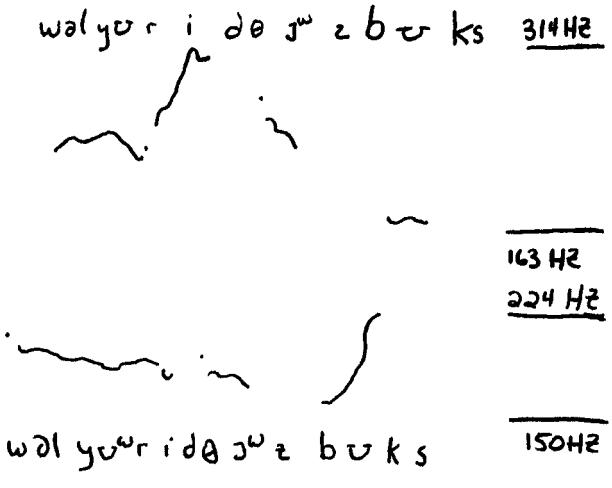
WARNING: P-Values reported for this repeated measures model have not been corrected for possible violations of the assumption of no correlation between observations.



Means Table
Effect: Sentence Type
Dependent: RT

	Count	Mean	Std. Dev.	Std. Error
Declarative	9	1009.434	436.804	145.601
Imperative	9	732.974	305.703	101.901
Pivot-Question	9	494.254	299.566	99.855
Pivot-Request	9	474.002	324.373	108.124
Question	9	629.733	325.718	108.573
Request	9	528.241	336.367	112.122





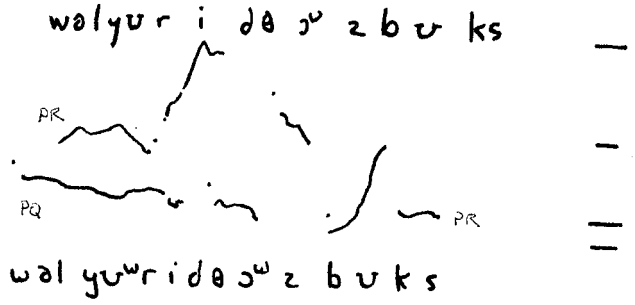
5. A Clue, and Generalization

5.1. Contour Example

Consider a contour pair that pivots successfully. The f0 extraction was performed using the Kay Elemetrics ILS system. The first is the request contour, which would be analyzed in Pierrehumbert's scheme as having an H* on "read" followed by a series of lows, including a low phrase accent and boundary tone. Hirschberg and Pierrehumbert would thus regard "read" as new information to be instantiated and asserted into the hearer's belief space. This is plausible[4], because in performing a request, the speaker is more interested in getting across the notion of doing the action than in the hearer's prior view of it. Therefore the speaker does not highlight the modal verb, but rather, the desired action itself. A yes-no question version of the same text can reverse this, stressing the auxiliary as the predicate being questioned relative to the rest of the proposition.

Two perceptual observations go unlabelled. First, the "low" phrase accent and boundary tone are not low relative to the speaker's range, and give a feeling of incompleteness. Mapping it onto a non-aux-inverted sentence does not yield a clear declarative. Second, the voice quality is somewhat nasal and gives a feeling of complaint. So it seems that something is missing from this picture: the slope of the contour, perhaps, or the fact that it must be interpreted relative to the aux-inverted syntax. A comparison with a contrastive-stress assertion would not be amiss.

The question contour is exactly the familiar high rise, central to traditional analyses. The contrast between the two contours is distinct.



6. Conclusion

Intonation has a role to play in the interpretation of speakers' intentions, when properly relativized to other information sources. We include it along with extrapositional linguistic information which is later relativized to context. The request/question distinction in particular is (in the absence of context) sometimes recognized by all hearers, and consistently recognized by some. Our model avoids the pitfalls of one-to-one mapping, and may well be able to incorporate intonation in a straightforward way. What is next required is a further investigation of intonational features, and an extension of the model to handle varying strengths of evidence.

---

[2]It is possible that context has a priming effect, which would allow this "second" stage to have effects even before the entire sentence is heard.

[3]Another possibility is a more static tune library. This would explain some child language learning data [Petersa], which shows production of stereotypical tunes.

[4]predicted, even, in [Hinkelman1991b]

References

Allen1983a.
James Allen, "Recognizing Intentions From Natural Language Utterances," pp. 107-166 in *Computational Models of Discourse*, ed. B. Berwick,MIT Press, Cambridge, MA (1983).

Austin1963a.
J. L. Austin, *How to Do Things with Words*, Harvard University Press, Cambridge, MA (1963).

Beun1989a.
Robbert-Jan Beun, *The Recognition of Declarative Questions in Information Dialogues*, Catholic University Brabant, Brabant, the Netherlands (October 1989).

Bolinger1982a.
Dwight Bolinger, "Nondeclaratives from and Intonational Standpoint," *Papers from the Parasession on Nondeclaratives*, pp. 1-22 Chicago Linguistic Society, (April 1982).

Cohen1990a.
Philip R. Cohen and Hector J. Levesque, "Rational Interaction as the Basis for Communication," in *Intentions and Communication*, ed. M. E. Pollack,MIT Press, Cambridge, MA (1990).

Gazdar1979a.
Gerald Gazdar, *Pragmatics: Implicature, Presupposition and Logical Form*, Academic Press, New York (1979).

Geluykens1988a.
Ronald Geluykens, "On the Myth of Rising Intonation in Polar Questions," *Journal of Pragmatics* **12** pp. 467-485 Elsevier, (1988).

Hinkelman1991b.
Elizabeth A. Hinkelman, "Acoustic Cues of Speaker Intentions," Proposal for NSF grant nr. IRI-9109914, University of Chicago (January 1991).

Hinkelman1991a.
Elizabeth A. Hinkelman and James F. Allen, "Speech Act Interpretation Without Literal Meaning," *sub. Computational Linguistics*, (January 1991).

Hinkelman1989b.
Elizabeth A. Hinkelman and James F. Allen, "Two Constraints on Speech Act Ambiguity," *Proc. Association for Computational Linguistics*, (1989).

Hinkelman1989a.
Elizabeth A. Hinkelman, "Linguistic and Pragmatic Constraints on Utterance Interpretation," TR 288, Computer Science Department, University of Rochester, Rochester, NY (1989).

LeHiste1970a.
Ilse LeHiste, *Suprasegmentals*, MIT Press, Cambridge, MA (1970).

Nusbaum1991a.
Howard C. Nusbaum, Kevin J. Broiher, and Judith C. Goodman, "Listening to the Sound of Sentences," *Journal of the Acoustical Society of America* **89** p. 2011 (1991).

Petersa.
Ann M. Peters, "Language Learning Strategies: Does the Whole Equal the Sum of the Parts?," *Language* **53** pp. 560-73 (1977).

Pierrehumbert1990a.
Janet Pierrehumbert and Julia Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions and Communication*, ed. M. E. Pollack,MIT Press, Cambridge, MA (1990).

Sadock1974a.
Jerrold M. Sadock, *Toward a Linguistic Theory of Speech Acts*, Academic Press, New York (1974).