Intonational Cues to Student Questions in Tutoring Dialogs

Jennifer J. Venditti, Julia Hirschberg, Jackson Liscombe

Department of Computer Science
Columbia University, New York, USA
{jjv,julia,jaxin}@cs.columbia.edu

Abstract
Successful Intelligent Tutoring Systems (ITSs) must be able to recognize when their students are asking a question. They must identify question form as well as question function in order to respond appropriately. Our study examines whether intonational features, specifically, F0 height and rise range, are useful cues to student question type in a corpus of 643 American English questions. Results show a quantitative effect of both form and function. In addition, among clarification-seeking questions, we observed differences based on the type of clarification being sought.1

Index Terms: intonation, questions, tutoring, dialog acts.

1. Introduction

1.1. Motivation
Student questions are common in one-on-one tutorial interactions. For example, [1] found that a student will ask an average of 26.5 questions per hour during a tutoring session, in contrast to .11 questions per hour in a classroom setting. In building effective Intelligent Tutoring Systems (ITSs), it is essential to be able to detect student questions and respond appropriately. To accomplish this, ITSs need first to recognize question form: for example, polar questions seek a yes-no answer, while wh-questions seek different information. In addition, however, tutoring systems need to recognize question function: for example, the tutor’s response will be different when a wh-question seeks information (1a) from when a wh-question seeks clarification (1b).

(1) a. (S has just submitted an essay to the tutor.)
S: Oh, what do you think about that?
T: Uh, well that uh you have uh there are too many parameters here which uh need definition ...
b. T: So if there is if the only force on an object in earth’s gravity then what is its motion called?
S: What was the motion called?
T: Yes, what’s the name for this motion?

Similarly, yes-no questions seeking information (e.g. Do they move in the same direction?) may cue the tutor to provide more than a simple yes or no. Yes-no questions seeking clarification, on the other hand, are likely to trigger a clarification subdialog. Still another class of yes-no questions, those seeking confirmation, may trigger some reinforcement strategy on the part of the tutor.

So, if automated tutors are to make distinctions among questions similar to human tutors, question form and function must be detected. A common strategy in text-processing systems is to look for subject-aux inversion to identify yes-no questions and for wh-words to identify wh-questions. However, Shriberg et al. [2] found that questions are often misclassified as statements when the classification relies upon words alone, due to the presence of declarative questions, which may be distinguished from statements only in terms of their prosodic characteristics. They showed that integration of a prosodic tree model with their language model based on (true) words yields the best performance accuracy in question detection in a corpus of Switchboard conversations. While prosodic information has been applied to the detection of question form, the prosodic characteristics of question function is less well understood. In this paper we examine the intonation of student questions in tutorial dialogs, including an analysis of question function, to determine the intonational markings of various types of dialog acts that student questions can perform.

1.2. Previous studies of question prosody
Intonation is widely believed to provide the most useful acoustic-prosodic cue to question identification in spoken corpora; there is a huge body of literature describing the intonational contrast of statements vs. questions in terms of falling vs. rising fundamental frequency (F0) contours (e.g. [3, 4, 5, inter alia]). Most studies have pointed out a systematic effect of syntactic form on question intonation — the common wisdom is that yes-no questions and declarative questions tend to rise, and wh-questions tend to fall. Additional studies have refined this view, presenting distributions of rising and falling contours for these question types [6, 7] and providing details on the variation within a given question class. Much less work has been done to identify intonational cues to the function of questions in discourse, perhaps because functional categories are themselves more difficult to specify. Some descriptive work has investigated the meanings that questions uttered with different intonational contours can convey in various contexts [8, 9, 10, 11]. Corpus-based studies have examined question function in Map-Task corpora, to determine whether rising/falling contour type or pitch accent type can distinguish questions fulfilling different dialog acts (Glasgow [12], German, Italian, Bulgarian [13]). In addition, [14] showed that the type of clarification request affected the distribution of rises and falls in another corpus of German task-oriented spoken dialogues. Laboratory studies have found that peak location of accents in Swedish could be varied to differentiate between questions seeking clarification of perception (Did you say X?) vs. those clarifying understanding (Did you really mean X?) [15]. However, a comprehensive quantitative analysis of the prosodic features of question function in English is still lacking. In this paper we present an analysis of question intonation in a corpus of human-human tutoring dialogues, with the goal of identifying features that might be useful for question function

1This research was funded in part by NSF grant IIS-0328295.
identification in an ITS system.

2. The corpus

We examine a corpus of human-human tutoring dialogs collected by [16] for the development of ITSpoke, a speech-enabled ITS designed to teach physics. The corpus consists of one-on-one sessions between undergraduate students (all American English speakers) and a professional tutor. We have tagged 1030 student questions, and have observed a rate of 25.2 questions per hour, an average of 13.3% of total student speaking time. This paper examines only a subset of the entire tagged corpus: 643 tokens from the 5 students who asked the most questions of all students.

3. Tagging questions

3.1. Coding question type

We coded student questions along two dimensions: form and function. Coding of question form was based on surface syntactic structure. We distinguish the following 6 form categories:

- **Declarative question** ($dQ$): It’s a vector? or A vector?
- **Yes-no question** ($ynQ$): Is it a vector?
- **Wh-question** ($whQ$): What is a vector?
- **Tag question** ($ynTAG$): It’s a vector, isn’t it?
- **Alternative question** ($altQ$): Is it a vector or a scalar?
- **Particle (part): Huh?

The coding of question function is less straightforward. After considering a number of dialog act annotation schemes including [17, 18], we adopt a simplification of Stenström’s categorization of question acts [19]. We collapse her 10 distinctions into 4 which we feel are most critical for ITSs to distinguish:

- **Confirmation-seeking check question** (chk), see also [17, 20].
- **Clarification-seeking question** (clar), see also [14].
- **Information-seeking question** (info), see also [17, 18].
- **Other (oth)

3.2. Segmentation, categorization, and F0 measures

The portion of the question from the nuclear accent to the rightmost edge of the phrase was marked, based on the waveform and spectrographic records. Contours were not given a full phonological (ToBI) transcription, but were classified into two groups: falling (e.g. H^*L-L%) vs. non-falling (e.g. H^*H-H%, L^*H-H%, H^*L-H%). For this study, we examined the following acoustic measures: speaker-normalized ($z$-score) F0 of (i) the nuclear accent (nucF0)\(^3\), (ii) rightmost edge of question (i.e. the boundary tone location) (btF0), and (iii) the difference between (i) and (ii) (riserange).

\(^3\) Defining what is a ‘question’ can be tricky (e.g. [1, 6]). Bolinger notes that “a Q[uestion] is fundamentally an attitude, which might be called a ‘claiming’ — it is an utterance that ‘claims’ a ... response.” [3, p. 4]. We follow in the spirit of this rather lay characterization: questions are those student utterances judged as seeking some kind of response from the tutor.

4. Analysis and results

4.1. Student question types

Table 1 shows the distribution of question form and function in our corpus, pooled across all subjects. Both $dQ$s and $ynQ$s occurred with every discourse function, $whQ$s functioned as either clarification- or information-seeking, $ynTAG$s functioned as either confirmation- or clarification-seeking, and particles were solely clarification-seeking.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>chk</th>
<th>clar</th>
<th>info</th>
<th>oth</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dQ$</td>
<td>257</td>
<td>81</td>
<td>2</td>
<td>4</td>
<td>444 (53.5)</td>
</tr>
<tr>
<td>$ynQ$</td>
<td>53</td>
<td>80</td>
<td>27</td>
<td>5</td>
<td>165 (25.7)</td>
</tr>
<tr>
<td>$whQ$</td>
<td>47</td>
<td>47</td>
<td>21</td>
<td>4</td>
<td>68 (10.6)</td>
</tr>
<tr>
<td>$ynTAG$</td>
<td>41</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>46 (7.2)</td>
</tr>
<tr>
<td>$altQ$</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>12 (1.9)</td>
</tr>
<tr>
<td>part</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8 (1.2)</td>
</tr>
<tr>
<td>N</td>
<td>357</td>
<td>226</td>
<td>51</td>
<td>9</td>
<td>643</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>chk</th>
<th>clar</th>
<th>info</th>
<th>oth</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>550</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2. Rises vs. falls

The distribution of falling intonation (L-L%) across question types is shown in Table 2. With the exception of particles, each form category contains some occurrences of falling contours, as has been reported in the literature (e.g. [3, 4, 6, 7, 21]). Falling contours were found in each function category as well. Both $whQ$s and $altQ$s show high percentages of falling contours (42.6% and 66.7%, respectively), and information-seeking $ynQ$s exhibit more terminal falls (81%) than those $whQ$s seeking clarification (25.5%).

<table>
<thead>
<tr>
<th>Fall Region</th>
<th>chk</th>
<th>clar</th>
<th>info</th>
<th>oth</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>550</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3. F0 measures

We conducted a quantitative analysis of F0 height in 573 non-falling (i.e. rising or plateau) contours.\(^4\) Figures 1 and 2 plot normalized F0 means on the nuclear accent (nucF0) and the boundary tone (btF0), respectively. Points with N<5 are not plotted, and will not be discussed here.

\(^4\) 568 tokens in our corpus were non-falling. Of these, 12 were removed because the final portion of the utterance was cut off (e.g. due to interlocutor interruption), and 1 more was removed due to insufficient F0 data.

4.3.1. Question form

Our main interest regarding question form is whether there are any F0 height differences between $dQ$s and $ynQ$s. Both categories are said to display similar ‘rising’ intonation (cf. [4, 6, 22, 23]). However, the nature of the rise may be distinct. Declarative questions are thought to be high-rising (e.g. $H^*H-H%$): the student
asserts information (perhaps answering a previous tutor question) thereby adding it to the mutual belief space, while at the same time questioning whether the tutor can relate this information to the contents of the tutor’s own (unshared) beliefs, namely, the correct answer [11, p. 407]. In contrast, ynTAGs are often thought to be low-rising (e.g. L*H-H%;) the L* cues information that is salient but is not to be added (yet) to the mutual belief space [8].

Our utterances have not been given a detailed phonological transcription, so we are not able to distinguish cases of H* from L* accents. In the absence of ToBI labels, we examine nucF0 values, with the prediction that nucF0 should be higher for dQs than for ynTAGs, as H* should be higher than L*. A two-way ANOVA on question form x function shows a main effect of question form for nucF0 (F(5)=19.34,p=0). However, planned comparisons using the Tukey method (alpha=.01) show that the nucF0 height of dQs was not significantly different from that of ynTAGs (see filled/solid circles in Figure 1). Rather, the main effect is due to ynTAGs having significantly lower F0, and particles having significantly higher F0. Aside from these ynTAG/particle effects, the only other comparison which was significant was whQ>dQ. A main effect of form was also observed for the dependent measures btF0 (F(5)=10.71,p=0, see Figure 2) and riserange (F(5)=3.60,p<.01, figure not shown). Again, the significant contrasts involved only ynTAGs and particles. The fact that both nucF0 and btF0 are lower for ynTAGs (and higher for particles) indicates that these questions are realized in a lower (or higher) overall register than other form types.

4.3.2. Question function

Our main interest regarding question function is simply whether or not the type of dialog act (DA) performed by the question affects the realization of question F0. We have already observed that DA type can affect distribution of rises and falls. However, there are no English studies that we are aware of which examine more quantitative effects of DA type on F0 height.

A two-way ANOVA on form x function shows a main effect of question function for each dependent measure (nucF0: F(3)=16.60,p=0; btF0: F(3)=8.56,p<.001; riserange: F(3)=3.94,p<.01). Planned comparisons using the Tukey method (alpha=.01) indicate that this is due to the nucF0 and btF0 of clarification-seeking clarQs being significantly higher than confirmation-seeking chkQs, and the nucF0 of information-seeking infoQs being significantly higher than both chkQs or clarQs (see Figures 1 and 2). The fact that both nucF0 and btF0 are higher for clarQs than for chkQs indicates that clarification-seeking questions are realized in a higher register than confirmation-seeking questions. Note that this effect is net of the form effect, and removing ynTAG and part data still yields a significant clarQ>chkQ contrast. It appears that information-seeking infoQs are realized in an even higher register, though the F0 height at the boundary in comparison with the other types did not reach significance. Planned comparisons of riserange were not significant, indicating that the extent of the rise for each type does not differ (net of the form effect). There were no significant interactions between form and function for any dependent measure (nucF0: F(5)=1.71,p=13; btF0: F(5)=.40,p=.22; riserange: F(5)=.66,p=.65).

4.3.3. Types of clarification

This section takes a closer look at clarification requests (clarQs, aka. ‘CRs’) in our corpus. A student seeks clarification when communication has somehow broken down, and hence s/he is unable to ground the information the tutor has attempted to add to the mutual belief space. Several authors have adopted Clark’s [24] four levels of coordination (see 1-4 in Table 3) for classifying the source of the communication problem. Rodríguez and Schlangen [14] found that the type of clarification affected the distribution of falling vs. rising contours in German: CRs clarifying reference had significantly more rising tunes, while those clarifying perceptual understanding had significantly more rising tunes. Edlund et al. [15] found that peak alignment affected interpretation of CRs in Swedish. Our interest is whether there are F0 height cues to clarification question type in our corpus.

Each clarification-seeking question in our corpus was tagged with one of the 5 categories shown in Table 3. In addition to Clark’s 4 levels, we added a ‘non-interlocutor-related’ (NIR) category, to describe CRs which were not targeted at the tutor’s utterance, but rather at the task/examination question at hand. For the current analysis, we combined Clark’s categories 1 and 2 into a single ‘acoustic/perceptual’ category. A one-way ANOVA shows a main effect of clarification type on both nucF0 (F(5)=5.41,p<.001) and btF0 (F(3)=6.6,p<.001), and a marginal effect on riserange (F(3)=2.59,p=.08). For each dependent measure, the ranking of the categories, from highest F0 to low-
Table 3: Sources of communication problems.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Understanding</th>
<th>Perception</th>
<th>Intention</th>
<th>Non-interlocutor-related (NIR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of communication problem</td>
<td>Problem with reference resolution (This up here?, What did I imply or what does the statement imply?), or with general understanding (Is that the same thing or is that different?, What do you mean?)</td>
<td>Problem hearing what the tutor said (‘G’ as in god?, Did you say a word or a letter?, reprise/echo questions like A what?)</td>
<td>Problem determining what the tutor intended by his utterance (You want an exact number?, Uh are you asking me another characteristic of freefall?)</td>
<td>Problem with reference resolution (This up here?, What did I imply or what does the statement imply?), or with general understanding (Is that the same thing or is that different?, What do you mean?)</td>
</tr>
</tbody>
</table>

5. Discussion

Successful Intelligent Tutoring Systems (ITSs) must be able to recognize when their students are asking a question. Systems need to identify question form as well as function in order to respond appropriately. Our study examined whether intonational features, specifically, F0 height and rise range, are useful cues to student understanding. Our study examined whether intonational features, specifically, F0 height and rise range, are useful cues to student understanding. In our corpus, clarification-seeking questions had higher F0, is: acoustic/perceptual clarification are realized with a higher F0 than those in the F0 measures we examined, contrary to our hypothesis. (Rising) whQs also did not have distinct F0, so ITSs may have to rely on lexical information to identify these. Tags were realized in a significantly lower register, and particles were significantly higher. We suspect that lexical cues would also aid in identifying these question types.

ITSs may make better use of F0 to identify question function. In our corpus, clarification-seeking questions had higher F0 than confirmation-seeking questions, and information-seeking questions had even higher F0 (particularly on the nuclear accent). These function distinctions may not be readily identifiable using lexical/syntactic features: Table 1 showed that there is no one-to-one form/function mapping (except for particles). Finally, we observed that the F0 of clarification questions differs depending on the type of clarification sought. clarification seeking acoustic/perceptual clarification are realized with a higher F0 than those seeking clarification of the tutor’s intention.

6. References


552