Prosodic measurements and question types in the Spontal corpus of Swedish dialogues

Sofia Strömbergsson, Jens Edlund, David House

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
sostr@csc.kth.se,{edlund,davidh}@speech.kth.se

Abstract

Studies of questions present strong evidence that there is no one-to-one relationship between intonation and interrogative mode. In this paper, we describe some aspects of prosodic variation in the Spontal corpus of 120 half-hour spontaneous dialogues in Swedish. The study is part of ongoing work aimed at extracting a database of 600 questions from the corpus, complete with categorization and prosodic descriptions. We report on coding and annotation of question typology and present results concerning some prosodic correlates related to question type for the 600 questions. A prosodically salient distinction was found between the two categories termed, in our typology, forward and backward looking questions.

Index Terms: speech prosody, spontaneous speech, question intonation, interrogative intonation

1. Introduction

Posing questions and the question-response sequence plays a central role in dialogue [1]. The analysis of such sequences has given rise to the theory of adjacency pairs forming the basic units for talk in interaction [2]. The signaling of interrogative mode in speech through intonation as a contrast to declarative statements is a topic which has long attracted interest from intonation researchers. It is readily assumed and often documented that intonation alone can transform a declarative into an interrogative, but a satisfactory analysis of question intonation has often eluded both descriptive phonetics and intonation models. Question intonation varies in different languages where also different types of questions (e.g. wh, yes/no or echo questions) can result in different kinds of question intonation [3].

In many languages, y/n questions are reported to have a final rise, while wh-questions typically are associated with a final low. Wh-questions are, however, often associated with a number of various contours [4]. In Dutch, a relationship has been documented between incidence of final rise and question type in conversational data in which wh-questions, y/n questions and declarative questions obtain increasing numbers of final rises in that order [5]. There are also languages with no morphosyntactic differences between y/n questions and statements and therefore make use of intonation to mark questions. In Neapolitan Italian [6], a late time alignment of a final accent plays a decisive role in the perception of interrogative mode.

Although much of the work done on question intonation has been confined to elicited speech an increasing number of studies are accessing large databases of conversational speech. In [7], around 200 questions were extracted from the Survey of English Usage, and rising intonation was found not to be very frequent in y/n questions. In a study of around 150 wh-questions in conversational question-answer sequences in German by Selting [8], intonation could not be systematically related to syntactic sentence structure type. She argues for prosody as an independent signaling system and describes prosody as an activity-type distinctive cue exemplified by "astonished questions" [9]. In a study of pitch patterns in nearly 300 German y/n questions and wh-questions taken from spontaneous speech, Kohler [10], tested the hypothesis that final rising and final falling intonation occur in both syntactic structures. He found that both pitch patterns occur in both structures, but that y/n questions had predominantly rising patterns (with more high-rising than low rising patterns) while wh-questions had mostly falling patterns (but a substantial number of low-rising patterns). By re-synthesizing complementary pitch patterns in the two structures, Kohler established that in "both syntactic structures, rising pitch expresses friendliness, interest and openness towards the addressee, while falling pitch focuses on routine, lack of interest and categoricalness" (p. 207). He also explained the difference in distribution between the structures by their different semantic and pragmatic functions. Wh-questions are information and fact oriented, while y/n questions ask for a decision from the addressee and are thus more addressee oriented.

In an investigation of 200 wh-questions extracted from a large corpus of computer-directed spontaneous speech in Swedish in [11], phrase-final rising intonation was seen as signaling dialogue acts and speaker attitude over and beyond an information question. Final rises occurred in 22 percent of the utterances, primarily in conjunction with final focal accent. Perception tests showed that high and late focal accent peaks in a wh-question are perceived as friendlier and more socially interested than low and early peaks.

Taken together, these studies present strong evidence that there is not a one-to-one relationship between intonation and interrogative mode.

The present study is part of a project that aims to investigate and describe intonational variation in questions in the Spontal corpus [12]. By investigating and describing variation within a subset of 600 questions taken from the corpus, the project will test the hypothesis that the concept of a standard type of question intonation such as a final pitch rise which contrasts to a final low of declarative intonation is not consistent with the pragmatic use of intonation in dialogue. We report on the extraction of questions from the Spontal corpus, on the coding, annotation and refining of a question typology, and on results concerning prosodic correlates to question type.

2. Method

2.1. Extraction of questions

The Spontal corpus contains in excess of 60 hours of dialogue: 120 nominal half-hour sessions, recorded in high-quality audio and video [12]. The subjects are all native speakers of Swedish, allowed to talk about anything they wanted at any point in the session, including meta-comments on the recording environment.

Orthographic transcriptions of the corpus have been made using a transcription tool which separates the two speakers into separate audio channels and divides the temporal progression of the dialogue into talkspurts based on pauses (i.e. talkspurts in the sense of [13]). Each dialogue was transcribed by one annotator, and then checked by another. For a subset of 24 dialogues, both primary and secondary annotators looked for questions while annotating and labeled these with a question tag. The definition of "question" was deliberately kept quite open: "Anything that resembles, structurally or functionally, in whole or in part, a question". In all, 908 talkspurts received the question label.

2.2. Question markup

Three independent annotators labeled all 908 instances with respect to four relatively simple queries, each of which could apply to any type of question. During the process, annotators could choose to *skip* talkspurts that they felt were in no sense a question, or that were otherwise impossible to judge. 168 talkspurts were skipped by at least one annotator, and therefore excluded from further analysis, leaving us with a set of 740 talkspurts that were labeled by all three annotators.

The queries **Q1-Q4** were kept simple in the hope that naïve annotation would help categorize the questions without relying heavily on preconceptions, and that certain clusters of **Q1-Q4** labels might map to certain question types as described in the literature. The labels could then be used to categorize questions, by form and function, in a reasonably objective and repeatable manner. Inspiration for the queries was taken from a coding scheme for question-response sequences developed by Stivers and Enfield [14]. An annotation tool was developed which enabled annotators to easily listen to a talkspurt and step through the queries. Response selection was executed by simple keyboard commands, mouse clicks, or by tapping a touchpad.

Q1 had to do with question type. Most, if not all, theories of questions agree on the existence of y/n and wh-questions. Following [14], we asked whether the talkspurt would best be described as a y/n question (**Y/N**), a wh-question (**WH**), an alternative question which include a restricted set of alternative answers (**ALT**), a multi-question which is defined as two or more questions posed in a single talk spurt (**MULTI**), or other (**OTHER**). Given our considerably wider scope of what constitutes a question, which includes questions seeking acknowledgement and questions contained in reported speech, results were not entirely predictable. **Q2** concerned whether a response was required (**REQUIRED**), possible (**OPTIONAL**), or prohibited (**PROHIBITED**). **Q3** was to be answered in the positive if the person producing the question-like talkspurt showed a clear attitude towards the previous dialogue such as surprise, distrust or uncertainty (**ATTITUDE**), and in the negative if not (**NOATTITUDE**). **Q4** was to be answered in the positive if the question-like talk was a case of reported speech (**REPORTED**), and in the negative if not (**DIRECT**).

2.3. Data cleaning

As reported in [15], annotators disagreed frequently on **Q3**. Annotators all agreed that they had, over time, begun interpreting **Q3** in a different manner. In an attempt to resolve this, the category description was changed to better fit the annotators' interpretation. The new options were **FORWARD** and **BACKWARD** accompanied with the question "Does the person asking the question ask for something that has not already been said (**FORWARD**) or is it more a question of verifying or showing attitude towards what has already been stated (**BACKWARD**)?". The talkspurts were then re-annotated for **Q3** with this new definition. Some other inconsistencies were discovered in the use of the label **MULTI**. Therefore, one of the annotators inspected all talkspurts annotated (by at least one annotator) as **MULTI**, and made a final decision on the label.

Table 1. *Distribution of question types as defined by clustering the annotations of **Q1-Q3**.*

Rank	Count	%	Label cluster
1	211	35%	Y/N_REQUIRED_FORWARD
2	181	30%	WH_REQUIRED_FORWARD
3	43	7%	Y/N_OPTIONAL_FORWARD
4	35	6%	WH_REQUIRED_BACKWARD
5	34	6%	Y/N_REQUIRED_BACKWARD
6	22	4%	ALT'S_REQUIRED_FORWARD
7	21	4%	WH_OPTIONAL_FORWARD
8	16	3%	Y/N_OPTIONAL_BACKWARD
9	11	2%	OTHER_REQUIRED_FORWARD
10	10	2%	OTHER_OPTIONAL_BACKWARD
11	6	1%	WH_OPTIONAL_BACKWARD
12	3	1%	Y/N_PROHIBITED_FORWARD
13	3	1%	WH_PROHIBITED_FORWARD
14	2	0%	OTHER_OPTIONAL_FORWARD
15	1	0%	OTHER_REQUIRED_BACKWARD
16	1	0%	ALT'S_REQUIRED_BACKWARD

REPORTED questions are typically embedded in talkspurts; they are often preceded by a lexical marking that signals that what follows will be a case of reported speech, e.g. "*och hon bara: "* (Eng. *"and she's like:"*). Talkspurts labeled as **MULTI** by definition also contain more speech than just one question. In order to enable automatic prosodic analysis of these types of questions, talkspurts with these labels were segmented manually. Finally, we excluded talkspurts where the annotators had all disagreed on at least one query label. (When at least two annotators agreed on a category for a query, this category was selected as the label for that query.) This left us with a set of 641 questions. The targeted 600 questions were selected from this set so that they were balanced for the interlocutors' gender and previous acquaintance, but otherwise at random. Table 1 displays the distribution of question types within the set of 600 questions.

2.4. Prosodic analysis: DUR, VAR and DIFF

Three different prosodic measures were extracted from the 600 questions: duration (**DUR**), pitch variation (**VAR**) and an estimate of intonation slope (**DIFF**). The calculation of pitch variation followed the description in [16]; pitch was tracked in semitones, and the standard deviation of the pitch was calculated per question as a measure of pitch variation for that particular question. These measures were then averaged within each question type, to find the average pitch variation within a question type (**VAR**).

As a rough estimate of the rising or falling intonation across a question, we used the difference between the average pitch of the first half of the question and the average pitch of the second half of the question over question types (**DIFF**). Negative values correspond to predominantly falling pitch, positive values correspond to predominantly rising pitch.

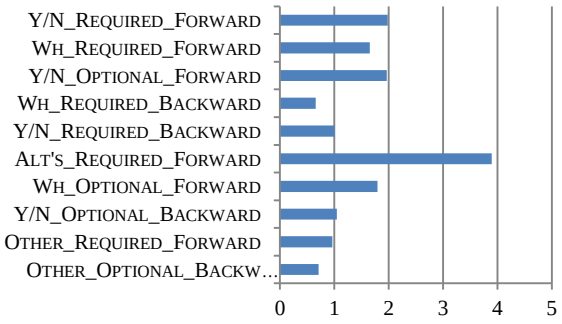

Figure 1. **Dur** *(seconds) of the 10 most frequent question types (question types with 10 or more occurrences).*

3. Results

In order to explore whether the three prosodic measures **Dur**, **Var** and **Diff** were dependent on the question type (its **Q1** label), a one-way MANOVA was performed. This determined that only **Dur** ($F(3,596) = 26.06$, $p < .001$) and **Diff** ($F(3, 596) = 4.41$, $p = .004$) were dependent on the **Q1** label. Analysis of more specific question types – as clustered also by **Q2** and **Q3** labels – is described in the following.

3.1. Duration

Figure 1 shows **Dur** of the questions within the 10 most common question types (rank 1-10 in Table 1).Y/N A trivial observation is that alternative questions are longer than questions of other types. In addition, backward looking questions are shorter than other types. A one-way ANOVA revealed that the differences between the question categories are significant: $F(9,574) = 14.61$, $p < 0.001$. A Bonferroni post-hoc analysis showed that **Alt's_Required_Forward** are significantly longer than questions of all other types. A one-way ANOVA comparing **Backward** ($N=103$; $M=.87$; $SD=.59$) to **Forward** ($N=497$; $M=1.92$; $SD=1.44$) revealed that **Backward** is significantly shorter than **Forward**; $F(1, 582) = 52.13$, $p < .000$.

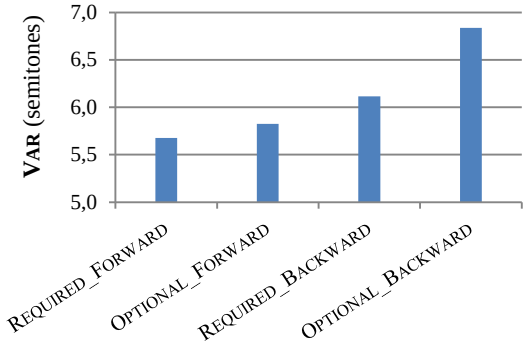

Figure 2. ***Var*** *(semitones) of question types clustered by* ***Q2*** *and* ***Q3*** *for categories with more than 10 instances.*

3.2. Pitch variation

For the analysis of **Var**, questions were clustered by **Q2** and **Q3**, i.e. by their answering elicitation degree (**Required**, **Optional**, **Prohibited**) and their directionality (**Backward**, **Forward**). This clustering yielded four categories with more than 10 instances in each: **Required_Forward** ($N = 425$), **Optional_Forward** ($N = 66$), **Required_Backward** ($N = 71$) and **Optional_Backward** ($N = 32$). Figure 2 shows their average pitch variation in semitones, and suggests that **Backward** contain more variation in pitch than **Forward**. A one-way ANOVA showed that the differences between the categories are significant: $F(3, 590) = 10.05$, $p < .001$. A Bonferroni post-hoc analysis revealed significant differences between **Optional_Backward** and the three other categories, and between **Required_Backward** and **Required_Forward**.

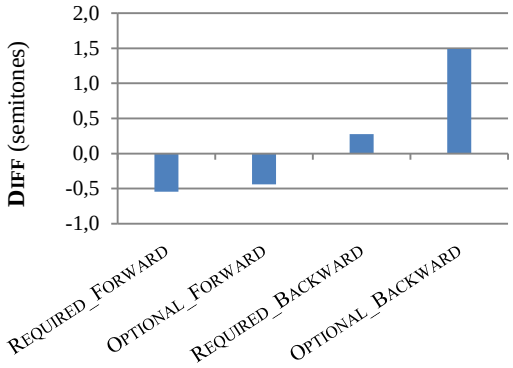

Figure 3. ***Diff*** *(semitones) of question types clustered by* ***Q2*** *and* ***Q3*** *for categories with more than 10 instances.*

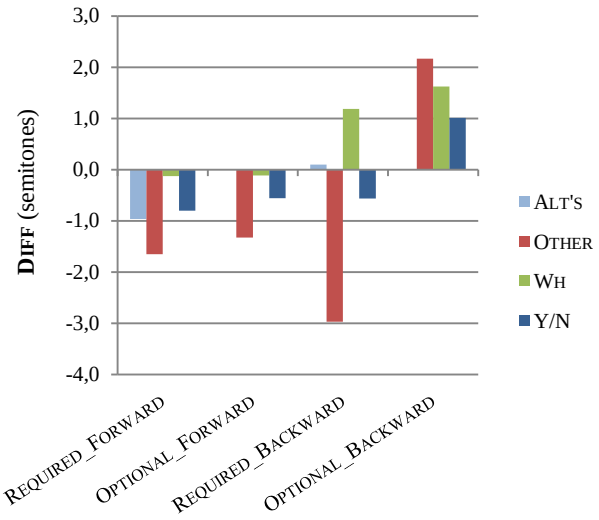

Figure 4. ***Diff*** *(semitones) of question types clustered by* ***Q2*** *and* ***Q3*** *labels, split by their* ***Q1*** *label.*

3.3. Rising/falling intonation

In analogy to the analysis of **Var**, the questions were clustered by **Q2** and **Q3** labels for the analysis of **Diff**. Figure 3 shows **Diff** within these four categories. The figure suggests a difference between a rising intonation slope for the two **Backward** categories and falling (or flat) intonation slopes for the other categories. A one-way ANOVA confirmed the difference: $F(3, 590) = 7.84$, $p < .001$, and a Bonferroni post-hoc analysis showed that the **Diff** value within **Optional_Backward** is

significantly larger than the **DIFF** values in the two **FORWARD** categories. (The difference between the categories **OPTIONAL_BACKWARD** and **REQUIRED_BACKWARD** is not significant, p = .15.)

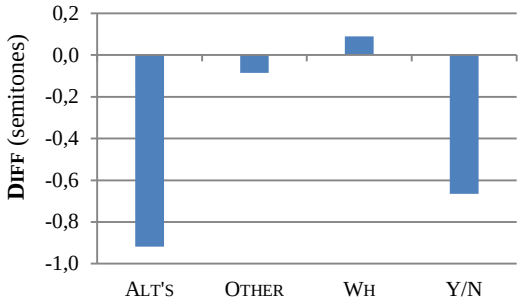

Figure 5. ***DIFF*** *(semitones) of questions grouped by* ***Q1****.*

In order to explore the distribution of intonation slopes within categories, each of the four categories were split by their **Q1** label (Figure 4). An observation is that within **REQUIRED_BACKWARD**, there are both negative and positive values. (Please note, however, that within this category, there is only one instance each of the types **ALT's** and **OTHER**.) **WH_REQUIRED_BACKWARD** (e.g. *What? What did you say?*) generally have rising intonation, whereas **Y/N_REQUIRED_BACKWARD** (e.g. *After Thursday? Are you?*) generally have falling intonation. Figure 4 also indicates that all of **FORWARD** have falling intonation.

As stated above, **DIFF** is dependent on **Q1** label. The distributions across **Q1** labels are displayed in Figure 5. A Bonferroni post-hoc analysis showed a significant difference in **DIFF** only between **WH** questions and **Y/N** questions; **WH** questions generally have a rising intonation, whereas **Y/N** questions generally have a falling intonation.

4. Discussion

In this study, we have extracted and annotated questions from the Spontal corpus, and presented analyses of the resulting question types in terms of different prosodic features. Ample evidence suggests that there is no one-to-one correspondence between traditional questions type (**Q1**) and intonation of questions, and a prosodically relevant categorization scheme should therefore include other functional aspects of the question. The annotation scheme we suggest allows categorization of questions across different dimensions: to what type of response the question elicitates (e.g. whether it is a wh-question or a y/n question); to the level of optionality of a response; and to whether it is forward or backward looking. By clustering labels across the different dimensions, we have explored the interplay between these orthogonal characteristics of a question and its prosodic realization.

An important finding that contrasts to previous findings reported for other languages (e.g. [5] and [10]) is that y/n questions in the Spontal corpus generally have falling intonation, whereas wh-questions have rising intonation. Although we use a rough estimate of intonation slope our findings are based on a relatively large corpus of questions. We have also confirmed that there is not a simple correspondence between the traditional question type and its prosodic realization – specifically, we have uncovered a more complex interplay between the form and the referential function of a question (i.e. whether it is backward or forward looking). Backward looking questions generally have a rising intonation, especially if answering the question is optional (e.g. *Okay?*) and/or if it is a wh-question (e.g. *What?*). Inspection of the questions within these categories suggests that a function of the rising intonation is to signal non-understanding or non-acceptance of the preceding context.

5. Acknowledgements

The work presented here is funded by the Swedish Research Council, Humanities and Social Sciences (VR 2009-1764) "Intonational variation in questions in Swedish."

6. References

[1] Sachs, H., Schegloff, E.A. and Jefferson, G. "A simplest systematics for the organization of turn-taking for conversation", Language 50, 696-735, 1974.

[2] Schegloff, E.A. "On some questions and ambiguities in conversation", In J.M. Atkinson and J. Heritage [Eds] Structures of social action: studies in conversation analysis, 28-52, Cambridge: Cambridge University Press, 1984.

[3] Ladd, D.R. Intonation phonology. Cambridge: Cambridge University Press.

[4] Cruttenden, A. Intonation. Cambridge: Cambridge University Press. 1986.

[5] Heuven, V.J. van, Hann, J. and Kirsner, R.S. "Phonetic correlates of sentence type in Dutch: Statement, question and command", Proceedings of ESCA International Workshop on Dialogue and Prosody, 35-40, Veldhoven, The Netherlands, 1999.

[6] D'Imperio, M. and House, D. "Perception of questions and statements in Neapolitan Italian", In Proceedings of Eurospeech 97, 251-254, Rhodes, Greece. 1997.

[7] Geluykens, R. "On the myth of rising intonation in polar questions", Journal of Pragmatics 12, 467-485, 1988.

[8] Selting, M. "Prosody in conversational questions", Journal of Pragmatics 17, 315-345, 1992.

[9] Selting, M. "Prosody as an activity-type distinctive cue in conversation: the case of so-called 'astonished' questions in repair initiation", In E. Couper-Kuhlen and M. Selting [Ed], Prosody in Conversation, 231-270, Cambridge: Cambridge University Press, 1996.

[10] Kohler, K.J. "Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions", In G. Fant, H. Fujisaki, J. Cao and Y. Xu [Eds.], From traditional phonology to modern speech processing, 205-214, Beijing: Foreign Language Teaching and Research Press, 2004.

[11] House, D., "Phrase-final rises as a prosodic feature in wh-questions in Swedish human–machine dialogue", Speech Communication, 46, 268-283, 2005.

[12] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., and House, D., "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture", In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2992-2995, Valetta, Malta, 2010.

[13] Brady, P. T. "A statistical analysis of on-off patterns in 16 conversations", The Bell System Technical Journal, 47, 73-91, 1968.

[14] Stivers, T. and Enfield, N.J., "A coding scheme for question-response sequences in conversation", Journal of Pragmatics 42, 2620-2626, 2010.

[15] Edlund, J., House, D., and Strömbergsson, S. "Question types and some prosodic correlates in 600 questions in the Spontal database of Swedish dialogues". In Proc. of Speech Prosody 2012. Shanghai, China. 2012.

[16] Edlund, J., & Heldner, M., "Underpinning /nailon/ - automatic estimation of pitch range and speaker relative pitch". In C. Müller [Ed.], Speaker Classification I: Fundamentals, Features, and Methods. Springer, 2007.