Question types and some prosodic correlates in 600 questions in the Spontal database of Swedish dialogues

Jens Edlund, David House, Sofia Strömbergsson[1]

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

{edlund,davidh}@speech.kth.se, sostr@csc.kth.se

Abstract

Studies of questions present strong evidence that there is no one-to-one relationship between intonation and interrogative mode. We present initial steps of a larger project investigating and describing intonational variation in the Spontal database of 120 half-hour spontaneous dialogues in Swedish, and testing the hypothesis that the concept of a standard question intonation such as a final pitch rise contrasting a final low declarative intonation is not consistent with the pragmatic use of intonation in dialogue. We report on the extraction of 600 questions from the Spontal corpus, coding and annotation of question typology, and preliminary results concerning some prosodic correlates related to question type.

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

1. Introduction

Posing questions and the resulting question-response plays a central role in dialogue [1]. Analyses of such sequences have given rise to the theory of adjacency pairs forming the basic units for conversation [2]. The signaling of interrogative mode in speech through intonation as a contrast to declarative statements is a common topic amongst 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 and different types of questions (e.g. wh, yes/no or echo questions) can result in different kinds of question intonation [3].

In many languages, yes/no 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 large number of contours [4]. In Dutch, a relationship has been documented between incidence of final rise and question type in which wh-questions, yes/no questions and declarative questions obtain increasing numbers of final rises in that order [5]. There are also languages with no morphosyntactic differences between yes/no questions and statements, which 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.

Much question intonation work use elicited speech, but an increasing number of studies use conversational speech. [7] used around 200 questions from the Survey of English Usage, and rising intonation was found not to be very frequent in y/n questions. In a study of some 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 word 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. By re-synthesizing complementary pitch patterns in the two structures, Kohler establishes 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 explains the difference in distribution between the two structures with their different semantic and pragmatic functions. The wh-question is information and fact oriented, while the y/n-question is asking for a decision from the addressee and is 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 [11] showed that high and late focal accent peaks in a wh-question are perceived as more friendly and socially interested than low and early peaks.

2. Extraction from the Spontal corpus

The Spontal corpus contains in excess of 60 hours of dialogue: 120 half-hour sessions [12]. The subjects are all native speakers of Swedish, and balanced (1) as to whether the interlocutors are of same or opposing gender and (2) as to whether they know each other or not. The recordings contain high-quality audio and video. Spontal subjects were allowed to talk about anything they wanted at any point in the session, including meta-comments on the recording environment.

Orthographic transcriptions of the database are made using a transcription tool separating the two speakers into separate audio channels and dividing the temporal progression of the dialogue into talkspurts - stretches of speech delimited by silence (i.e. talkspurts in the sense of [13]). To ensure consistency and quality, each dialogue is first transcribed by one annotator, and then checked by another. For a subset of 24 dialogues, primary and secondary annotators were asked to mark questions with a simple question tag while annotating. The definition of “question” was deliberately kept quite open: “Anything that resembles, structurally or functionally, in whole or in part, a question”.

All in all, 908 talkspurts were labeled as questions by the annotators. This set of questions should not be taken to be well-defined, but rather exploratory, nor can we ensure that every talkspurt containing question-like material has been labeled as such. However, the large number of questions and the fact that they come from many dialogues that are often quite different in style lead us to believe that the talkspurts hold a good coverage of different types of questions and question-like talkspurts in conversational dialogue.

[1] Authors in alphabetical order.

3. Question selection

The initial 908 question-like talkspurts displayed a great range and variety, from long and complicated multipart questions to brief feedback-eliciting utterances such as “oh yeah?”. In some cases, the reason for the annotator labeling a talkspurt as a question was less than obvious, which necessitates careful selection of the targeted 600 instances. On the other hand, the goal of the extraction of questions was to gather 600 question-like talkspurts with an open mind to avoid starting out from theory-laden preconceptions. This openness was chosen since we are looking for all kinds of prosodic variability in questions and therefore should not be restrictive concerning question types.

We went about the selection by having three independent annotators label all 908 instances with respect to four relatively simple queries (Q1-Q4), 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. In this process, 12 talkspurts were skipped by all three annotators, 46 by at least two, and another 110 talkspurts were skipped by at least one annotator. The number of talkspurts that were labeled by all annotators, then, was 740, and the targeted 600 were selected from these so that there were 150 in each of the original Spontal balance groups KNOWN and SAMEGENDER, KNOWN and DIFFERENTGENDER, UNKNOWN and SAMEGENDER, and UNKNOWN and DIFFERENTGENDER, but otherwise at random.

4. Question markup

The queries Q1-Q4 were kept simple in the hope that naïve annotation would help categorize at least part of the questions without relying too heavily on preconceptions. The idea was that certain sequences of answers to Q1-Q4 might map to certain question types as described in the literature, and that these answers could then be used to categorize questions 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] and used to code and compare questions in ten different languages. 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 yes/no 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 also includes questions seeking acknowledgement and also questions contained in reported speech, results are not entirely predictable. Q2 concerned whether a response was required (REQUIRED), possible (OPTIONAL), or prohibited (PROHIBITED). Q3 should 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 should be answered in the positive if the question-like talk was a case of reported speech (REPORTED), and in the negative if not (DIRECT). We added a final query, Q5: does the talkspurt include nothing but the question-like speech (and all of it; QUESTIONONLY) or does it contain more or less than that (MOREORLESS).

5. Disagreement between annotators

A first inspection of the resulting annotations shows that annotators generally agree on the more traditional question types in Q1, that is on the WH and Y/N labels – pairwise comparisons show agreements of above 80% for these. There is one exception: one annotator consistently labeled more Y/N than the other two, who instead labeled OTHER in many of these cases. The rest of the alternatives for the first query show less agreement, varying between 14% and 60%. This is to be expected – ALT, MULTIPLE, and OTHER are much less well-defined and also clearly overlap with WH and Y/N in some cases. Q2 shows above 80% pairwise agreement between all annotators for the REQUIRED and PROHIBITED responses. The agreement for OPTIONAL is lower, around 50% on average. For the queries about attitude and reported speech – Q3 and Q4 – there are only a small number of occurrences. Q3 was included for explorative reasons, and Q4 because prosody is different in citation voice than in other speech, and is probably best modeled separately. Taking Q1-Q4 together (e.g. Y/N-REQUIRED-NOATTITUDE-DIRECT), all annotators agreed on each of them in 200 cases, and two out of three annotators agreed in another 191, leaving 209 cases where all judges disagreed on at least some query.

6. Four-label combinations

Table 1 shows the most frequent four-label combinations (Q1-Q4). We will discuss question types and characteristics in terms of these combined labels, and in groups based in part on their frequency in the material, in part on the impressions of the annotators, and in part on examination of the talkspurts contained in each four-label combination group. The label resulting from Q5 is not included in the combinations as it does not concern the question per se. Instead, the proportion of QUESTIONONLY vs. MOREORLESS has been taken to be a feature of the four-label combination groups, based on talkspurts that received identical four-label combinations from at least two annotators. Numbers for DURATION, NUMBEROFWORDS, and WORDS-PERSECOND are calculated for the QUESTIONONLY talkspurts only.

6.1. Traditional Y/N and Wh-questions

A general inspection of talkspurts in the Q1-Q4 groups reveals that the two most frequent by far are Y/N-REQUIRED-NOATTITUDE-DIRECT and WH-REQUIRED-NOATTITUDE-DIRECT. These represent direct yes/no and wh-questions to which answers are expected, mapping well to y/n-questions and wh-questions in the sense of Stivers and Enfield. The groups show the highest agreement between annotators – 153 of the talkspurts were given one of these combinations by all three annotators in unison, and two out of three annotators agreed on them in an additional 90 cases. The talkspurts have been labeled as belonging to these groups in 46% of the cases, and 243 of all talkspurts were deemed to belong to them by at least two annotators. 80% of these talkspurts were deemed by at least two annotators to be QUESTIONONLY on Q5. The average duration of these talkspurts was 1.9 s, the average number of words in a talkspurt 7.4, and the average number of graphemic vowels (for a rough estimation of syllables) 10.5. This gives a speech rate estimate of 3.8 words/s and 5.5 vowels/s.

6.2. Optional responses

The third and seventh most common groups, Y/N-Optional-NoAttitude-Direct and Wh-Optional-NoAttitude-Direct, are heterogeneous groups of talkspurts with less agreement between annotators. This is in part explained by the fact that all three annotators had a hard time applying Q3 and almost exclusively used NoAttitude initially. After judging a number of talkspurts, all annotators reported independently that they started interpreting the query more liberally, using the Attitude response every time the talkspurt in one way or another questioned the contents of the previous speaker's contribution. Talkspurts were labeled as belonging to these groups in 12% of the cases, and 57 of all talkspurts were deemed to belong to one of these combinations by at least two annotators. Only 54% of them were deemed by at least two annotators to be QuestionOnly on Q5. Of these, the average duration, words/talkspurt, and vowels per talkspurt are all smaller than those of traditional y/n and wh-questions: 1.6, 5.8 and 8.3, respectively.

Table 1: Ordered listing of the most common four-label combinations

Rank	Count	%	# 2+ same	Label
1	506	28	145	Y/N-Required-NoAttitude-Direct
2	335	18	98	Wh-Required-NoAttitude-Direct
3	165	9	45	Y/N-Optional-NoAttitude-Direct
4	111	6	31	Multiple-Required-NoAttitude-Direct
5	92	5	21	Wh-Required-Attitude-Direct
6	86	4	22	Y/N-Required-Attitude-Direct
7	63	3	12	Wh-Optional-NoAttitude-Direct
8	59	3	12	Y/N-Optional-Attitude-Direct
9	50	2	15	Wh-Prohibited-NoAttitude-Reported
10	49	2	19	Alt's-Required-NoAttitude-Direct
11	47	2	15	Y/N-Prohibited-NoAttitude-Reported
12	33	1	8	Wh-Optional-Attitude-Direct
13	30	1	1	Other-Optional-Attitude-Direct
14	29	1	4	Wh-Prohibited-NoAttitude-Direct
15	22	1	1	Multiple-Required-Attitude-Direct
16	19	1	4	Other-Required-NoAttitude-Direct
17	18	1	1	Other-Optional-NoAttitude-Direct

6.3. Multiple responses

The fourth most common group was Multiple-Required-NoAttitude-Direct, which together with the 1% labeled as Multiple-Required-Attitude-Direct in the 15th most common group makes up the only combinations containing Multiple. The questions in this group often give an insistent impression, and annotators perceive them as high in tempo. Talkspurts were labeled as belonging to these groups in 7% of the cases, and 32 of all talkspurts were deemed to belong to them by at least two annotators. 71% of these talkspurts were deemed by at least two annotators to be QuestionOnly on Q5. Out of these, the average duration, average word count and average vowel count were substantially greater than for traditional y/n and wh-questions: 3.8 s, 16 words and 22.4 vowels, respectively. This indicates a higher speaking rate, supporting the intuitions of the annotators: 4.2 words per second and 5.8 vowels per seconds.

6.4. Feedback elicitation and clarification requests: Attitude responses

In addition to the 15th group above, the fifth, sixth, eighth, twelfth and 13th make up the groups containing Attitude. The agreement between annotators is lower in these groups, as noted above. Wh-Required-Attitude-Direct, is made up largely of the token "What?" and other talkspurts of similar meaning. Y/N-Required-Attitude-Direct contains a large proportion of clarification requests of the type "Did you say X?", and the remaining groups containing the Attitude label are also largely made up of talkspurts that have to do with grounding and clarification of what was said. Many schemes would not label these as questions to begin with. The talkspurts were labeled as belonging to these groups in 14% of the cases, and 54 of all talkspurts were deemed to belong to one of these combinations by at least two annotators. They are short, and contain nothing else: 89% were deemed by at least two annotators to be QuestionOnly on Q5. Their average duration was 1.1 s, the average number of words in a talkspurt 3, and the average number of vowels 4.3. This gives a speech rate estimate of 3 words/s (the lowest recorded in the data) and a vowel based speech rate of 4.3 – again very low.

6.5. Reported, rhetorical and self-directed speech: Prohibited responses

The ninth, eleventh and 14th most common groups are the only ones containing the Prohibited label. Wh-Prohibited-NoAttitude-Reported and Y/N-Prohibited-NoAttitude-Reported, Prohibited is combined with Reported speech, reflecting the fact that annotators agreed that questions in reported speech were a clear instance of talkspurts to which giving an answer would be strange and unexpected. Wh-Prohibited-NoAttitude-Direct is the only one combining Prohibited without Reported. Inspection reveals that this group contains self-directed speech and rhetorical questions. The talkspurts were labeled as belonging to these groups in 5% of the cases, and 34 of all talkspurts were deemed to belong to them by at least two annotators. Unsurprisingly, the category almost always contains more material than the question in itself. Only 7% of these talkspurts were deemed by at least two annotators to be QuestionOnly on Q5. Their average duration was 2.3 s, the average number of words 8.3, and the average number of vowels 12.3.

6.6. Alt responses

The tenth group, Alt's-Required-NoAttitude-Direct, is the only one containing the Alt's label. Talkspurts were labeled as belonging to this group in 2% of the cases, and 19 of all talkspurts were deemed to belong to is by at least two annotators. They are often the only speech found in the same talkspurt. 82% were deemed by at least two annotators to be QuestionOnly on Q5. Their average duration was 4.4 s, the average number of words 16.1, and the average number of vowels 24.1. Together with the numbers for multiple questions, these are the highest numbers in the data, although in this case there is no clear increase in speech rate: the numbers are 3.6 words/s and 5.4 syllables/s, which is relatively typical for the data.

6.7. OTHER responses

The remaining two groups, ranking 16th and 17th in frequency, have been selected in a mere 2% of the cases, and 5 of all talkspurts were deemed to belong to them by at least two annotators. They are the only groups apart from the 13th group containing the OTHER label. They include interesting cases of questions that are passed back to the person originally asking the question (e.g. “What about you?”) and questions based on the exclusion of the requested word (e.g. “You said you live in…?”).

7. Conclusions and future work

We have extracted 600 talkspurts from the Spontal corpus that were (a) question-like according to Spontal transcribers, (b) possible to label by three annotators for four queries regarding their nature, and (c) balanced in the same way as the Spontal database. The scope was deliberately kept wide in order to steer clear, initially, of theory-laden decisions.

We view both agreement *and* disagreement among annotators as results in this exploratory study. Examining systematic differences between the annotators will lead us to new insights about the categories of questions. One such example concerns feedback elicitation and clarification requests. The ATTITUDE label was initially difficult for annotators, in their own view due to an unrealistic expectation that speakers would sound completely taken aback or totally distrusting – events that rarely occur in conversational dialogue. With time, annotators realized that a large number of the talkspurts had as their main purpose to question or comment on what had just been said, and so matched the definition for ATTITUDE.

Another disagreement surrounded the same group of talkspurts. One of the annotators often labeled a surprised “Oh really?” or “Yeah?” with Y/N, whereas the other two consistently selected OTHER. Although this discrepancy could be easily resolved by simply agreeing on a convention, the Y/N label is interesting from a prosodic point of view. A number of these utterances have a final rise in pitch and sound slightly inquisitive – if responded to, the answer would likely be “Yeah, it’s true!” or possibly “No, I was just fibbing”. This is consistent with a Y/N question, and maps to the second level – perception – of the process Clark [15] calls grounding and Allwood et al. [16] interactive communication management. Other talkspurts with similar lexical content, however, are longer and have their high peak earlier, and seem to mean “I don’t know how you can say that”. Answering these with a simple “yes” or “no” seems insufficient; instead a more full response explaining the previous statement is needed. This is consistent with the third level of grounding – understanding. The pitch contours and the meanings noted by the annotators are also consistent with the contours investigated in [17].

We also note that the longer and faster talkspurts labelled with MULTIPLE and ALTERNATIVE are interesting categories, but difficult to analyse prosodically because of their length and their variability, and that we need a systematic description of their internal components, and finally, two of the most common question types with the OTHER label – questions passed on, or back to the person first asking and questions formed by simply omitting the requested word for the interlocutor to fill in – are relatively frequent and seem to warrant their own types. Annotators perceived at least the latter as prosodically marked.

8. 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.”

9. 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] Clark, H. H. Using language. Cambridge, UK: Cambridge University Press, 1996.

[16] Allwood, J., Nivre, J., and Ahlsen, E. “On the semantics and pragmatics of linguistic feedback”, Journal of Semantics, 9(1), 1-26, 1992.

[17] Skantze, G., House, D., and Edlund, J. “User responses to prosodic variation in fragmentary grounding utterances in dialogue”, In Proceedings of Interspeech 2006—ICSLP, 2002-2005, Pittsburgh PA. USA, 2006.