What's in a Question?

Marie Šafářová

1. Summary

It is generally assumed that a question in English can be expressed by other than just interrogative means (i.e., finite verb - subject inversion and/or presence of a *wh*-word for constituent questions). If so, how do we define the term 'question' and what procedures can we use to identify questions in a conversational corpus? The discussion and lack of consensus found in the literature indicate that the answer is far from trivial. One way to resolve the issue appears to be to turn to speakers' intuitions. For this purpose, an experiment was designed to test the interpretation of the term 'question' by native speakers of American English. A statistical evaluation of native speaker judgements in the categorization task suggests that, despite the common impression that "you know one when you see one," speakers actually appear to understand the term in different ways. A logistic regression analysis shows that this effect cannot be explained by the 'continuum hypothesis' according to which the difference between questions and assertions is not categorical but gradient. Instead, it seems that the semantics of interrogatives gets confounded with the pragmatic effect of response-seeking utterances (including imperatives, expressions of epistemic uncertainty and acknowledgement elicitation). In order to maintain a clear conceptualization of the terminology, it is proposed here to reserve the term 'question' for the *semantic* object denoted only by (all) interrogatives, as captured, e.g., in the partition semantics of Groenendijk & Stokhof (1997). For utterances (both interrogative and declarative) to which the addressee responds - be it in a linguistic or non-linguistic manner - the term 'response-seeking' is employed instead. Among them, the most interesting kind is that of evaluative response-seeking utterances which have the same pragmatic effect as "true" polar interrogatives in context. Some criteria for their identification in conversational transcripts are discussed.

2. Introduction

It has been noted many times in the past that there is no one-to-one relation between interrogatives and their canonical discourse function, usually labelled as 'asking a question'. For example, the utterances in (1) have different syntactic forms (only (1a) being an interrogative), yet all of them appear to have the same intended effect on the listener.

(1) a. *"Is it true?"*

b. *"I would like you to tell me if it's true."*

c. *"Tell me if it's true!"*

d. *"It's true?"* [with rising intonation]

Some researchers assume that a question is any utterance to which a speaker received and/or expected to receive a response (e.g., Geluykens, 1988). However, a superficial examination of a corpus of conversational speech shows that many utterances receive a confirmation from the addressee (as in (1d), from the Santa Barbara Corpus of American English). If non-verbal communication is taken into account, most, if not all utterances may qualify as questions according to the criterion above: this is because of the frequent facial movements (such as eye/brow/mouth movements) and head nods employed in cooperative communication. In general, the feedback serves to acknowledge that the listener follows the speaker, has understood what she said and, possibly, agrees with her or not (viz. Clark, 1996). Only a small part of the responses is, in fact, clearly answer-like in that it addresses the issue under disscussion with respect to which the other speaker appears to be ignorant.

(2) a. A: *"In other words, I should change filters."*

b. B: *"Ugh, God."*

c. A: *"I know it won't last long but it sure does make a mess in the house."*

d. B: "*Yeah, it makes a mess."*

On the other hand, if only these utterances are regarded as questions, many interrogatives no longer qualify as rightful members of the category. This is because they of their use in contexts in which the speaker is *not* ignorant about the discussed issue, merely to solicit a confirmation from the addressee or to bring up a hitherto disregarded possibility. The most typical examples of this use of interrogatives are *reversed-polarity tag questions* (viz. (3) from the Santa Barbara Corpus), and (some) *negative polar questions with preposed negation* (as in (4b), also from the SBC).[1]

(3) a. KATHY: *"I meant once you bring it over there."*

b. NATHAN: "*I know what you meant. I don't ever remember us doing anything like that though. There's like a way you always can get rid of those absolute value bars in problems, isn't there? Can I use some of this?"*

c. KATHY: *"Oh. Yeah. Mm. See, yeah. Here it's absolute values. Right here."*

(4) a. ROY: "*Do we want to be the guy in charge of the world."*

b. MARILYN: "*Unhunh."*

c. PETE: "*On the other hand, isn't he saying it's too late?"*

d. ROY: "*Well he is kind of saying it's too late."*

e. PETE: *"So-"*

1 There is another type of negative polar questions with preposed negation, so-called inner negation, which are used as true information-seeking questions (viz. Ladd, 1981, Han & Romero, 2003, van Rooij & Šafářová, 2003, and Šafářová, 2005, for a discussion).

f. ROY: "*This is why it's such a depressing book.*"

Moreover, it is not always possible to distinguish utterances addressing an issue about which the speaker is ignorant from those of which the speaker is knowledgeable because speaker's beliefs may not be manifested in the examined conversation. For this reason, Šafářová & Swerts (2004) suggest that any procedure of question identification is unreliable unless the researcher has access to speakers' intentions. However, it turns out that even this stringent requirement may not be sufficient since speakers may experience doubts if inquired about their own utterances (Grundstrom, 1973).[2] According to Haan (2001), the problem with the question definition is that there actually exists no clear boundary between questions and assertions. The two categories are not binary but gradient, in other words, they form opposing ends of a single scale. For example, questions whose function has the support of formal properties (wh-word, inversion) and questions seeking new information are more prototypical than questions lacking these formal markers and/or confirmation-seeking questions (Haan, 2001): 19). In what follows, Haan's position will be referred to as the 'continuum hypothesis'.

In the next section, I will describe the results of an experiment the purpose of which was to examine the understanding of the notion 'question' by native speakers of American English. First of all, the experiment had as its goal to confirm that speakers interpret the term in a uniform way.[3] If a statistically significant agreement were found, the utterances classified as questions were to be analyzed for the presence (and, possibly, importance ranking) of prototypical question features. As shown below, however, the level of disagreement among judges was too high. The data, however, appear to disprove the 'continuum hypothesis'. Instead, it appears that the variation in participants' responses is due to a confusion between the semantics and pragmatic effects.

3. Experimental Study

One of the goals of the experiment was to use native speakers' intuitions (if sufficiently uniform) to establish a workable procedure for question identification in conversations. The intended purpose of the procedure was to collect questions from a corpus in order to examine their intonational patterns. Since the questions have to be collected without reference to their intonational properties (to avoid bias), only conversational transcripts can be used in the identification process. Therefore, judges in the experiment were only presented with a transcript of a free conversation, with no access to its audio recording.

The experimental hypothesis was formulated as follows:

[Hypothesis] The term 'question' is used in a uniform way by native speakers.

2 Grundstrom concludesthat the distinction between a question and an assertion is not obligatory in some languages. I will not advocate this particular view here.

3 This assumption is made, e.g., in the experimental work on intonational meaning in which participants are asked to classify tunes according to whether they perceive them to be questioning or not (viz. Šafářová (2005) for a discussion).

3.1 Method

The participants in the study were 26 native speakers of American English, 13 male and 13 female, between 24 and 67 years old. Some of them were contacted directly and some of them were recruited by their departmental colleagues or via the [phonlist], a phonetics e-mail list. They were paid for their participation in the task with an Amazon gift certificate for $15. The task was estimated to take between 15-20 minutes, though some of the judges reported that they needed more time.

The task was set up as a web form with an accompanying *php*-script for input submission. A transcript of a natural free conversation, publicly available from the Santa Barbara Corpus, part I, was segmented into individual meaningful units assumed to correspond to utterances, with 875 utterances in total. There were three speakers in the conversation, all female, referred to as Lynne, Lenore and Doris in the transcript. Their individual contributions were highlighted with different colors (red, blue and green) in order to help the judges process the text. Each utterance started on a new line and finished with a "check box". The participants received the following instructions:

[Instructions:] Below is a transcript of a real conversation. Check the box next to the utterance you think was a question.

Prior to the actual task, the judges were asked to submit their name, e-mail, age and gender. Submitting their name and e-mail was necessary for the participants to receive their Amazon gift certificate. The information about their gender and age was considered of possible relevance for the evaluation of the responses.

3.2 Results

Speakers' responses were first evaluated in terms of general agreement. There were in total 218 cases classified as questions by at least one judge. Of these, one fifth (19.5%) was classified as questions by more than a half of the judges; on the other hand, one fifth (20%) of the utterances identified as questions were identified by just one judge. There were 6 utterances that were identified as questions by *all* the judges: these were all isolated utterances of just a *wh*-word (such as *what, why*) or a *wh*-word with a part of the predicate (e.g., *where is the uh*) or with subject-finite verb inversion (e.g., *did they train you that*). Overall, more than 60% of the utterances categorized as questions were so interpreted by five or less of the 26 judges.

A brief analysis of the responses revealed that most apparent points of disagreement, summarized below.

1. *you know*-phrases: These were prevalent in the speech of one of the recorded speakers (Lynne) who also contributed longer monologues into the conversation. Some judges always marked the phrases as questions, others did so only if they were turn-final and some judges never considered them to be questions.

2. uncertainty expressions: Some judges classified expressions of uncertainty as questions, e.g., (5) or (6), similarly also (7) and, possibly, also (8): [4]

(5) "*I wonder if that has something to do with it*"

(6) "*maybe it is, maybe it is*"

(7) a. LYNNE: *"so she named everybody in the class except for"*
b. LENORE: *"about four people -- well I think they asked for names"*
c. LYNNE: "*oh did they*"

(8) a. LYNNE: "*he's a pretty neat guy I thought*"
b. LENORE: "*yeah*"

3. imperatives: Some judges classified also imperatives, as in (9), as questions:

(9) *look right here*

4. hidden questions: Utterances that were used to elicit a response, as in (10) where the speaker is asking a direct witness of an event for a confirmation of information she received, were sometimes classified as questions.

(10) "*and she said like the lights looked a greenish color*"

5. incomplete utterances: Utterances with ellipsis or requests for confirmation such as (11), (12) and (13), were a frequent source of disagreement, in that some judges identified them as questions, while others consistently excluded them.

(11) "*really*"

(12) "*Debby*"

(13) "*in Roundup*"

4 It is not clear whether (6) was interpreted as expressing uncertainty or was a misunderstanding.

6. rhetorical questions: Some judges classified rhetorical *wh*-questions as questions.

The responses were evaluated in a pair-wise comparison using the Cohen's *kappa* coefficient (Cohen, 1960), where zero corresponds to an agreement between two judges which was the same as chance. The statistic is calculated on the basis of a two-by-two table, where *A* is the number of utterances both subjects considered to be questions, *B*, like *C*, is the number of utterances which had this property according to one of the judges but not according to the other one, and *D* is the number of utterances on which the two judges agreed that they were not questions. Obviously, $A+B+C+D=N$, *N* the total number of utterances (i.e., 875). The formula for calculating kappa is given in (A) below.

Table 1.: The two-by-two table for the *kappa-statistics*.

	Q for judge 1	Non-Q for judge 1	Total
Q for judge 2	*A*	B	A + B
Non-Q for judge 2	C	D	C + D
Total	A + C	B + D	N

(***A***) $\kappa = [2(AD-BC)] / [(A+C)(C+D)+(B+D)(A+B)]$

The strength of agreement, measured with the *kappa*-statistics, is often described verbally with the Landis-Koch scale (Landis & Koch (1977)), as: <0 - “poor”, 0-.20 - “slight”, .21-.4 - “fair”, .41-.6 - “moderate”, .61-.80 - “substantial” and .81-1.0 - “almost perfect” (though some statisticians oppose this interpretation).

The *kappa*-statistics, though widely used, is sometimes disputed as the appropriate measure of inter-rater agreement (Cicchetti & Feinstein (1990)). Therefore, additional measurements of agreement were made, using raw agreement indices: the observed proportion of overall agreement (which, unlike *kappa*, does not correct for chance), symbolized here as p_o, and the proportion of specific agreement for positive and negative ratings, p_s^+ and p_s^- respectively. Given that the conversational transcript contained proportionally a much larger number of utterances overall classified as non-questions, especially the positive rating p_s^+ was of relevance. The formulas for calculating the raw agreement measurements are given below:

(***B***) $p_o = (A + D)/(A+B+C+D) = (A+D)/N$

(***C***) $p_s^+ = 2A / (2A+B+C)$

(***D***) $p_s^- = 2D / (2D+B+C)$

The *kappa* coefficient and the raw agreement measurements were calculated for all pairs of judges (for 26 judges, 325 pairs) on their compared scores, using an *awk* program, and evaluated in SPSS. Cohen's *kappa* coefficient varied from 0.15 to 0.82, with the mean *kappa* being 0.52 (S.D.=.139), which is taken to correspond

to “fair” on the Landis-Koch scale. The proportion of overall agreement varied between 0.36 and 0.99, with the mean 0.94 (S.D.=.568); the agreement on positive ratings varied between 0.19 and 0.99 with the mean 0.55 (S.D.=.137) and the agreement on negative ratings varied between 0.92 and 0.99 with the mean 0.96 (S.D.=0.018).

Although the *kappa* coefficient and the raw agreement scores can give an overall impression of subjects' agreement in the task, a more detailed statistical analysis is needed to determine the difference in underlying models speakers used to identify questions in the task. For this purpose, all the utterances were tagged for the presence/absence of seven categorical features estimated to play a role in the subjects' decision process. The description of the features is summarized in table 2.

Table 2. : *Features estimated to play a role in judgements of question identification.*

inversion	Y/ N	subject - finite verb inversion in the (main) clause
uncertainty	Y/ N	contains an expression of uncertainty
question test	Y/ N	if turned into an interrogative, would be felicitous in context
wh-word	Y/ N	*wh*-word in the utterance
turn final	Y/ N	turn-final
you know	Y/ N	*you know* in the utterance
yes/no answer	Y/ N	followed by a *yes/no*-reply

The feature uncertainty was present for utterances that contained one of the following phrases: '*I guess', 'I wonder', 'I forget', 'I don't know', 'maybe'* and/or '*perhaps'*. These were considered to express the uncertainty or ignorance of the speaker; the list is based solely on expressions of uncertainty found in the transcript. The feature question test was present for declarative utterance which, if turned into an interrogative by a simple subject-finite verb inversion in the main clause, still appeared to be felicitous in their context (see also below for examples and further discussion of this feature). The annotation of the utterances was done by a single annotator, which may be problematic for the feature question test given that it involves a subjective judgement. The features wh-word, you know, turn final and uncertainty, on the other hand, were identified mechanically with a simple matching procedure. The same could be done for the feature yes/no-answer, selecting utterances immediately followed by *yes/no/mhmm/yeah* (here, also responses that contextually entailed *yes/no* were taken into consideration).

The features were inspected for possible dependencies with a *chi*-square analysis; for cases where more than 20% of the expected frequencies were below 5 or one of them below 1, Yates' correction was employed. The calculations were done with an online calculator at www.unc.edu/~preacher/chisq/chisq.htm. The analysis revealed a statistically significant association in eight cases out of the twenty-one possible ones and in these cases, the effect of a factor (as described below) thus cannot be ascribed uniquely to its presence but has to be considered together with its dependencies. The presence of an uncertainty expression was frequently accompanied with the presence of a *wh*-word, with $X^2_{(1)}2=33.506$, $p < .01$, an effect which can be explained by

the frequent embedding of *wh*-clauses under the predicates *I don't know, I forget* and *I wonder*. Turn-finality was significantly associated with all the other features, except for uncertainty, with, for the pairs wh-word - turn-final, yes/no-answer - turn-final, inversion - turn-final, question test - turn-final, you know - turn-final, respectively, $X^2_{(1)}2=10.839$, $p < .01$; $X^2_{(1)}2=177.509$, $p < .01$; $X^2_{(1)}2=16.701$, $p < .01$; $X^2_{(1)}2=25.690$, $p < .01$; $X^2_{(1)}2=23.468$, $p < .01$. The presence of a *wh*-word was not independent of inversion, with $X^2_{(1)}2=59.020$, $p < .01$ and, finally, inversion was not independent of the yes/no-answer feature, (as could be expected, since interrogatives are frequently followed by a *yes/no* answer), with $X^2_{(1)}=21.880$, $p < .01$.

In order to determine the role of the seven features as factors in speakers' individual, as well as combined responses the data were analyzed with logistic regression in SPSS 11.0.3, with 'question' as the dependent (binary) variable and the seven features as categorical predictors. For all subjects, their models were significant (the null hypothesis being that no utterance is a question), but there were individual differences with respect to the significance of the different features (viz. table 5 in the appendix). For example, while many judges considered the feature inversion to be the most important for question identification, for other judges it was rather the fact that an utterance contained the *'you know'* phrase or that it contained a *wh*-word or could in the context be turned into an interrogative (the question test feature). The uncertainty feature was not significant in any of the individual models.

Following the individual analyses, the responses by all twenty-six subjects to the 875 utterances were merged together, thus giving 22 750 cases and evaluated with the logistic regression analysis, using the seven binary features as before. The purpose of the analysis was to identify features relevant for question identification across subjects.

With the null hypothesis that all utterances are not questions, which could account for 93 percent of the responses, the model with the seven features could account for 94,2 percent of the responses, correctly identifying 38,1 percent of utterances subjects classified as questions (compared to 0 percent based on the null hypothesis). The model was significant with $X^2_{(7)}=4537.340$ and $p <.01$. The importance of the individual features, which were all significant, is summarized in table 3.

Table 3.: *Importance of individual features in a model for merged responses.*

	B	S.E.	Wald	df	Sig.	Exp(B)
uncertainty	2.935	.394	55.472	1	<.01	18.820
you know	-3.101	.095	1070.196	1	<.01	.045
wh-word	-2.200	.095	538.719	1	<.01	.111
yes/no ans	-1.462	-1.462	216.404	1	<.01	.232
inversion	-3.534	.128	767.614	1	<.01	.029
quest test	-1.525	.122	156.176	1	<.01	.218
turn final	-2.025	.092	481.178	1	<.01	.132

The results reveal that in the general model, the two most important factors for question classification were inversion and the presence of '*you know*', while the presence of an uncertainty expression, though significant, was not as important as any of the other factors.

In the logistic regression analysis described above, as questions in the model were considered all utterances that were so classified by at least one judge. Since there was a large amount of disagreement between subjects, the results were also analyzed from the perspective of the majority opinion. Again, the logistic regression analysis on all results was performed, but this time, only utterances that were categorized as questions by more than half of the judges (i.e., more than 13) were considered as questions in the model, using as predictors the seven binary features as before. With the null hypothesis that all utterances are not questions, which could account for 95.3 percent of the responses, the model could account for 99.3 percent of the responses, correctly identifying 61 percent of questions (compared to 0 percent based on the null hypothesis). The classification in this model was thus clearly better than in the previous one (cmp. 61% of questions correctly categorized to 31.8% in the previous model). The model was significant with $X^2_{(7)}=218.090$ and $p<.01$. Compared to the model summarized in table 2 the importance of the individual features differed, in that the feature uncertainty and you know were no longer statistically significant. Of the remaining features, again inversion was the most relevant one, followed by turn-finality and the presence of a *wh*-word. The results for the individual features are summarized in table 4 below.

Table 4.:*Importance of individual features in a model for merged responses with majority question categorization.*

	B	S.E.	Wald	df	Sig.	Exp(B)
uncertainty	7.642	28.703	.071	1	.790	2083.452
you know	-.553	1.194	.215	1	.643	.575
wh-word	-3.482	.632	30.329	1	<.01	.031
yes/no ans	-2.019	-2.019	10.286	1	.01	.133
inversion	-5.233	-5.233	27.369	1	<.01	.005
quest test	-2.238	-2.238	9.420	1	.02	.107
turn final	-3.615	-3.615	17.307	1	<.01	.027

3.3 Discussion

To sum up the results, the level of agreement regarding which utterances were questions was quite poor among the native speakers who participated, and exhibited great variation. This result suggests that speakers, if consistent in their judgements, employed different definitions of the term 'question'. This possibility received further support through a logistic regression analysis of the data, using seven binary features associated with questions as predictors. Individual models varied in that some judges appeared to give most weight to subject-finite verb inversion, while for others it was the presence of '*you know*' in the utterance (a factor which, on the

other hand, was not statistically significant in the model of some judges). In a general model with merged responses, where as questions were considered all utterances that were classified as such by at least one judge, the seven predictor features all appeared to play a role in question identification, the two most important being subject-finite verb inversion and the presence of *'you know'*. Since the responses exhibited a great variation, the features were also evaluated in a model with merged responses where as questions were considered only utterances so classified by the majority of judges. In this model, 99.3 of the utterances were correctly classified (61 % of those considered questions); only five of the seven predictor features were statistically significant, with inversion being the most important one.

The 'continuum hypothesis', formulated on the basis of the discussion in Haan (2002), does not appear to be supported by the experimental results because the ranking of the predictor features differed for the individual subjects. In fact, a feature (e.g., the presence of the *you know* phrase) could be the most significant question predictor in some models and not have any significance in others. Presumably, if the problem with categorizing questions was that the difference between questions and statements is a continuum, the experimental participants would be expected to differ only in how strict they are in their judgements (some of them employing a very narrow definition, others a very broad one), but to preserve the same ranking.

One possible explanation that appears to be in line with the results reported above is that the semantic and the pragmatic interpretations of the term 'question' were confounded. Under the semantic interpretation, as questions are considered only interrogatives (i.e., utterances with subject - finite verb inversion and possibly the presence of a *wh*-word). In fact, this appeared to be the core interpretation in that the predictor features inversion and wh-word, together with the statistically dependent turn-finality, were the most significant ones in a model with merged responses based on majority classification. Another interpretation concerns utterances which appear to be *response seeking* in a broad sense, i.e., ranging from acknowledgment that the addressee has heard and understood an utterance to the addressee performing a non-verbal action in the sense of Clark (1996). Note that for "true" interrogatives, both the semantic and the pragmatic component are present (they have the semantics of questions and the discourse function to signal response-seeking). Rhetorical questions, however, lack the pragmatic component and this may have been the reason why some judges classified them as questions and others did not.

The purely response-seeking-based interpretation would be the reason why some judges classified all occurrences of the phrase *you know* as questions, since the phrase usually serves to elicit a low-level acknowledgment (at least signalling that the speaker has the listener's attention, e.g., in the form of a small nod or eye gaze in the speaker's direction). Also, it accounts for the otherwise unexpected categorization of imperatives as questions, as in (9), where the imperative is response-seeking in that it asks for a non-verbal action from the listener. Other points of disagreement noted above, such as expressed uncertainty or eliciting a direct witness account of reported events also fall into the category of response-seeking utterances.

3. Conclusion

In order to employ a more transparent terminology, it appears desirable to reserve the term ‘response-seeking’ for the pragmatic function of interrogatives, as well as other utterances which ask for the listener's response, be it a low-level acknowledgment (e.g., “yes, I heard you”), a higher level acknowledgment (e.g., “yes, I understood you”), an evaluative response (e.g., “yes, I believe that it is true”) or a non-verbal action response (e.g., looking where the speaker is pointing (9)). Clearly, there is no one-to-one relation between the semantic denotation and the pragmatic function, in that not all interrogatives are response-seeking (e.g., rhetorical questions). In the same way, not all response-seeking utterances denote questions. In fact, as suggested in Šafářová (2005), only interrogatives denote questions and that the response-seeking effect of some declaratives is of purely pragmatic nature.[5]

5 Is there at all a direct relation between the question denotation and response-seeking? Or is the primary pragmatic function of the use of interrogatives the raising of a discourse topic, and the response-seeking a derived pragmatic effect?

4. Appendix

Table 5.: The importance of different utterance features for question categorization (by experimental subjects).

ID	***Ranking***
1	Inversion > you know > turn-final > wh-word > question test > yes/no-answer
2	Inversion > question test > turn-final > wh-word > yes/no-answer
3	Inversion | you know > yes/no-answer > turn-final > wh-word
4	Inversion > question test > wh-word > turn-final > you know
5	Inversion > turn-final > wh-word
6	Inversion > you know > wh-word > turn-final > question test > yes/no-answer
7	Inversion > wh-word > turn-final> question test > yes/no-answer
8	Question test > turn-final > wh-word > you know > inversion
9	Wh-word > inversion > turn-final
10	You know > inversion > yes/no-answer > turn-final > wh-word
11	Inversion > you know > turn-final > question test > wh-word > yes/no-answer
12	You know > inversion > turn-final > yes/no-answer > wh-word
13	Question test > wh-word > inversion > turn-final > yes/no-answer
14	Inversion > question test > wh-word > turn-final
15	Inversion > you know > wh-word > turn-final
16	You know > inversion > wh-word > yes/no-answer > turn-final
17	Inversion > turn-final > wh-word > you know > question test > yes/no-answer
18	Inversion > turn-final > wh-word > question test > yes/no-answer
19	Inversion > wh-word > turn-final
20	Inversion > turn-final > wh-word
21	Inversion > you know > turn-final > wh-word > yes/no-answer
22	You know > inversion > wh-word > turn-final
23	Inversion > you know > turn-final > wh-word > yes/no-answer
24	Inversion > wh-word > turn-final > question test
25	Inversion > turn-final > wh-word > question test
26	Inversion > turn-final > wh-word > yes/no-answer

Bibliography:

Cicchetti, D.V. & A.R. Feinstein (1990): High agreement but low kappa: Resolving the paradoxes. Journal of Clinical Epidemiology, 43, s. 551-558.

Clark, H.H. (1996) Using Language. Cambridge University Press.

Cohen, J. (1960): A coefficient of agreement for nominal scales. Educ. Psychol. Meas., 20, s.27-46.

Geluykens, R. (1988): On the myth of rising intonation in polar questions. Journal of Pragmatics, 12, s. 467-485.

Groenendijk, J. & M. Stokhof (1997): Questions. In J. van Benthem & A. ter Meulen (eds.): Handbook of Logic and Language, Elsevier/MIT Press, s. 1055-1124.

Grundstrom, A. (1973): L'intonation des questions en Français Standard. In A. Grundstrom & P. L. Léon (eds.): Interrogation et intonation, 8, s. 19-51, Didier, Paris.

Haan, J. (2001): Speaking of Questions: An Exploration of Dutch Question Intonation. PhD thesis, University of Nijmegen.

Ladd, D.R. (1981): A First Look at the Semantics and Pragmatics of Negative Questions and Tag Questions. In Papers from the Seventeenth Regional Meeting of the Chicago Linguistic Society. Chicago Linguistics Society, s. 164-171.

Landis, R.J. & Koch, G.G. (1977): The measurement of observer agreement for categorical data. Biometrics, 33, s. 159-174.

van Rooij, R. & M. Šafářová (2004) On Polar Questions. In R. Young & Y. Zhou (eds.): Proceedings of SALT XIII, CLC Publications, Cornell University, Ithaca, N.Y.

Romero, M. & C.-H. Han (2003): On Negative Yes/No-Questions. Linguistics & Philosophy 27:5, s. 609-658.

Šafářová, M. (2005): Rises and Falls: Studies in the semantics and pragmatics of intonation. PhD thesis, ILLC, University of Amsterdam.

Šafářová, M. & Swerts, M. (2004) On recognition of declarative questions in English. In Proceedings of Speech cand Prosody, Nara, Japan.