PROSODIC CORRELATES OF DISCONFIRMATIONS

Emiel Krahmer, Marc Swerts, Mariët Theune, Mieke Weegels

IPO, Center for Research on User-System Interaction, Eindhoven, The Netherlands
{E.J.Krahmer/M.G.J.Swerts/M.Theune/M.F.Weegels}@tue.nl

ABSTRACT

In human-human communication, dialogue participants are continuously sending and receiving signals on the status of the information being exchanged. These signals may either be positive ('go on') or negative ('go back'), where it is usually found that the latter are comparatively marked to make sure that the dialogue partner is made aware of a communication problem. This paper focuses on the users' signaling of information status in human-machine interactions, and in particular looks at the role prosody may play in this respect. Using a corpus of interactions with two Dutch spoken dialogue systems, prosodic correlates of users' disconfirmations were investigated. In this corpus, disconfirmations may serve as a positive signal in one context and as a negative signal in another. Our findings show that the difference in signaling function is reflected in the distribution of the various types of disconfirmations as well as in different prosodic variables (pause, duration, intonation contour and pitch range). The implications of these results for human-machine modeling are discussed.

1. INTRODUCTION

From human-human communication it is known that dialogue participants are continuously sending and receiving signals on the status of the information being exchanged. This process is often referred to as *information grounding* ([3,13]) and typically proceeds in two phases: a *presentation phase* in which the current speaker sends a message to his conversation partner, and an *acceptation phase* in which the receiver signals whether the message came across unproblematically or not. In the former case (there is no problem), the receiver transmits a positive signal ('go on'), in the latter case (there is a problem), he or she sends a negative signal ('go back'). Various studies of human-human communication (e.g., [11]) revealed that the negative signals are comparatively marked, as if the speaker wants to devote additional effort to make the other aware of the apparent communication problem. A plausible explanation for this is that missing a negative cue may cause breakdown of the communication.

One of the central shortcomings of current spoken dialogue systems is that they are insufficiently able to spot communication problems (either resulting from poor recognition or from incorrect default assumptions) and hence have difficulty in responding to them. We conjecture that the ability of spoken dialogue systems to distinguish between positive and negative cues from the user is linearly correlated with the fluency of the interaction, since these cues provide important information about the status of the information currently under negotiation. We have studied a corpus of human-machine dialogues ([14]), obtained with two Dutch train time table information systems, in order to find out which cues people actually use in human-machine communication. In a companion to the current paper ([7]) a number of positive and negative cues have been singled out and their (joined) information potential for spotting communication problems was studied. It was indeed found that human speakers who converse with a spoken dialogue system put more effort in 'go back' signals than they do in 'go on' signals.

The current paper focuses on the prosodic features of positive and negative cues. We expect that speakers use more prosodic effort (higher pitch, longer duration, more pauses, marked intonation contours, ...) in the case of a 'go back' signal than they would for a 'go on' signal. To test this hypothesis, we concentrated on *one* type of utterance which may serve as a 'go back' signal in one context while it serves as a 'go on' signal in another context, namely a "no" answer to different types of system prompts. To illustrate this, consider the following two questions from the corpus of [14].

(1) a. Do you want to go from Eindhoven to Swalmen?
b. Do you want me to repeat the connection?

Both (1.a) and (1.b) are yes/no questions and to both "no" is a perfectly natural answer. However, the two questions serve a rather different goal. Question (1.a) is an (explicit) attempt of the system to verify some pieces of information that it has recently gathered (the departure and arrival station). If the user would respond to this question with a "no" this would definitely be a 'go back' signal: the user indicates that at least one of the system's beliefs is incorrect. Question (1.b), on the other hand, is not an attempt of the system to verify its beliefs, and hence it cannot represent incorrect system beliefs. A subsequent "no" answer from the user thus serves as a 'go on' signal. In this way, the two types of "no" answers constitute ideal, natural occurring, speech materials for investigating the role of prosody in information grounding, because, being lexically similar

Table 1: Positive vs. negative cues

POSITIVE ('go on')	NEGATIVE ('go back')
short turns	long turns
unmarked word order	marked word order
confirm	disconfirm
answer	no answer
no corrections	corrections
no repetitions	repetitions
new info	no new info

but functionally different, they constitute interesting minimal pairs from a dialogue perspective. They allow us to check whether the various occurrences of this utterance vary prosodically as a function of their context. The current paper focusses on the hypothesis that the 'go back' signals are prosodically marked with respect to the 'go on' signals. Before we describe the method used to test this hypothesis (section 3) and the results that were obtained (section 4), we present a brief overview of the context of this work.

2. EFFORT IN DIALOGUE

As said, [7] is in many ways a companion paper to the current one. The basic assumption of [7] is that both user and system want the dialogue to be finished successfully as soon as possible, and that they do not want to spend more effort than necessary for current purposes, in line with e.g., the *Principle of Minimal Collaborative Effort* of [2] or the more general *Principle of Least Effort* of [15]. Since a spoken dialogue system can never be certain that it understood the user correctly, it is in constant need of verification. If a verification question of the system contains a problem, users are expected to spend more effort on their signals in order to prevent complete breakdown of the communication. This leads to the distinction between positive and negative cues in table 1. In all cases, the positive cues can be seen as unmarked settings of the features. For instance, the default word order in a sentence is unmarked (thus, no topicalization or extraposition). Additionally, it follows from the Principle of Minimal Collaborative Effort that it is a positive signal to present new information (which may speed up the dialogue), but not to repeat or correct information (which will definitely not lead to a more swift conclusion of the conversation).

The central hypothesis of [7] is that users more often employ the 'go back' signals when the preceding system utterance contains a problem, whereas the 'go on' signals are used in response to unproblematic system utterances. For nearly all of the cues of table 1 this was indeed found. Many of these cues have a high informativity. For instance, if the user's answer contains a marked word-order, then it is highly likely that the preceding system utterance contained a problem. The downside, however, is that some of the highly informative cues occur rather infrequently. Therefore we also studied boolean combinations of cues. It turned out that the complex condition "the user's utterance contains more than eight words *or* uses a marked word order *or* contains corrected information" was the overall best cue for spotting communication problems, with a precision and recall of 92%. Recent experiments using memory based learning showed that it is possible to predict in 97% of the cases whether or not the preceding system utterance was problematic on the basis of the user's utterance (for more details of these experiments the reader is referred to the appendix). On the one hand, these results are certainly encouraging. They show that taking certain cues into account provides a reliable indicator of problems. On the other hand, one has to keep in mind that there is a certain gap between the hand-annotated data used in the experiments and the raw output of a speech recognition engine (a word graph).

Table 2: List of prosodic features and their expected settings for positive and negative cues

Features	¬ PROBLEMS	PROBLEMS
Boundary tone	low	high
Pitch range	low	high
Duration	short	long
Pause	short	long
Delay	short	long

It remains an empirical question to what extent the positive and negative signals from table 1 can be recovered automatically from a word graph. It is to be expected that shifting the analysis from hand-annotated data to word graphs will worsen the precision and recall scores for spotting communication problems. This implies that there is definitely room for improvement. Therefore, one possible extension to our previous work is to include another set of characteristics of user utterances in our prediction: a number of prosodic features.

To this end, the current paper looks at possible prosodic differences between positive and negative signals, using different types of disconfirmations as analysis materials. A previous study of repetitive utterances in Japanese human-human dialogues ([11]) showed that speakers more often provide negative signals with marked or prominent prosodic features than they do with positive signals. Consequently, we expect that in human-machine interactions the difference in signaling function will also be reflected in a difference in prosodic effort ([12]). This expectation is also based on recent work on hyperarticulate speech ([8,9,10]), a speaking style which can be seen both as the result of speech recognition errors and as an important source of such errors. Typically, hyperarticulate speech has an increased pitch and longer duration. All this leads to the expectations in table 2 regarding prosodic features and the predicted settings for positive and negative signals. Apart from testing these hypotheses, we also look at distributional differences between various types of negative responses as a function of their dialogue context.

3. METHOD

For the analysis, a corpus (see [14]) was used consisting of 120 dialogues with two speaker-independent Dutch spoken dialogue systems which provide train time table information. The systems prompt the user for unknown slots, such as departure station, arrival station, date, etc., in a series of questions. The two systems differ mainly in verification strategy (one primarily uses implicit verification, the other only uses explicit verification), length of system utterances and speech output (concatenated vs. synthetic speech). Twenty subjects were asked to query both systems via telephone on a number of train journeys. They were asked to perform three simple travel queries on each system (in total six tasks). Two similar sets of three queries were constructed, to prevent literal copying of subjects' utterances from the first to the second system. The order of presenting systems and sets was counterbalanced.

A random selection of 109 negative answers to yes/no questions from both systems was analysed (7 speakers). If the preceding yes/no question was a verification of the system's assumptions (e.g, (1.a) above), the user's disconfirmation indicates that the yes/no question contained a problem (due to speech recognition or incorrect assumptions on the system's part). If the yes/no question was not a verification (such as example (1.b), but also questions like *Do you want other information?* or *Do you want information about another connection?*), then the user's disconfirmation just serves as an answer to that question and does not indicate problems.

Regarding their structure, the users' disconfirmations were divided into three categories: (1) responses consisting of an explicit disconfirmation marker "no" ("nee") only ('single no'), (2) responses consisting of an explicit disconfirmation marker followed by other words ('no+stuff', Hockey *et al.* 1997), (3) responses containing no explicit disconfirmation marker ('stuff').

The speech data were digitized with a 16 kHz sampling frequency. Fundamental frequency (F_0) was determined using a method of subharmonic summation (Hermes, 1988). Durations of speech segments and of pauses were measured directly in the digitized waveform. The users' responses to the yes/no questions were analysed in terms of the following features: (1) type of boundary tone in "no" (high or low); (2) duration (in ms) of "no"; (3) duration (in ms) of pause after "no" before stuff; (4) duration (in ms) of pause between system's prompt and user response; (5) F_0 max (in Hz) at energy peak of major pitch accent in stuff; (6) number of words in stuff. It was our original intention to also investigate pitch range in the "no" part of the different responses, but this turned out to be too difficult given that many of the cases were realized with a low-anchored pitch accent followed by a high boundary tone (L*H-H%). For these utterances, it was not possible to adequately measure pitch range, given that the F_0 maximum in the energy peak in the pitch accent basically undershoots the perceived pitch range, whereas the real F_0 maximum at the end of the high boundary tone would overshoot it. See the discussion of figure 1 below.

4. RESULTS

This section first presents the results, and then illustrates some of the main effects with two typical examples.

Table 3: Numbers of negative answers following an unproblematic system utterance (¬ PROBLEM) and following those containing one or more problems (PROBLEM)

Type	¬ PROBLEMS	PROBLEMS	TOTAL
no	18	11	29
stuff	0	24	24
no+stuff	23	33	56
TOTAL	41	68	109

Table 4: Distribution of high and low boundary tones for positive and negative cues

Boundary tone	¬ PROBLEMS	PROBLEMS	TOTAL
Low	32	7	39
High	9	37	46
TOTAL	41	44	85

Table 3 gives the distribution of different types of disconfirmations following either an unproblematic system utterance or one which contains one or more problems. A χ^2 test reveals that this distribution is highly significant (χ^2 = 22.146, df = 2, $p < 0.001$). First, this table shows that the minimal response, a single no, is in the majority of the cases used as a positive signal. Second, single stuff responses are exclusively reserved for responses following a system utterance with one or more problems. The majority of the responses to yes/no questions in our data, however, is of the no+stuff type, which may serve either as a positive or as a negative cue. The lexical material in the stuff is quite different for the two signals: for the positive cases, the subsequent words are mostly some polite phrases ("thank you", "that's right"); for the negative cases, the stuff usually is an attempt to correct the information which is misrecognized or which is wrongly assumed by the system.

Table 4 displays the number of high and low boundary tones on the word "no" (for the single no and no+stuff cases) for positive and negative signals. A χ^2 test reveals that this distribution is again well above chance level (χ^2 = 33.004, df = 1, $p < 0.001$). In responses following an unproblematic system question, "no" is generally provided with a 'declarative' L% boundary tone, while in responses following a problematic question, the "no" generally receives a 'question-like' H% boundary tone. These results are in agreement with observations in Japanese human-human conversations ([11]).

The results for the continuous prosodic features of interest are given in table 5. Taking the utterances of all subjects together, a t-test reveals a significant difference for each of these features. The numbers of unproblematic and problematic utterances are insufficient and/or unequally distributed in order to test intra-individual differences. However, when looking at the mean within-subject differences, the findings mostly point in the expected direction, thus warranting an overall t-test. For all speakers, the mean duration of "no" and of pauses, F_0 max in stuff, and the number of words in stuff are usually higher in problematic than in unproblematic cases.

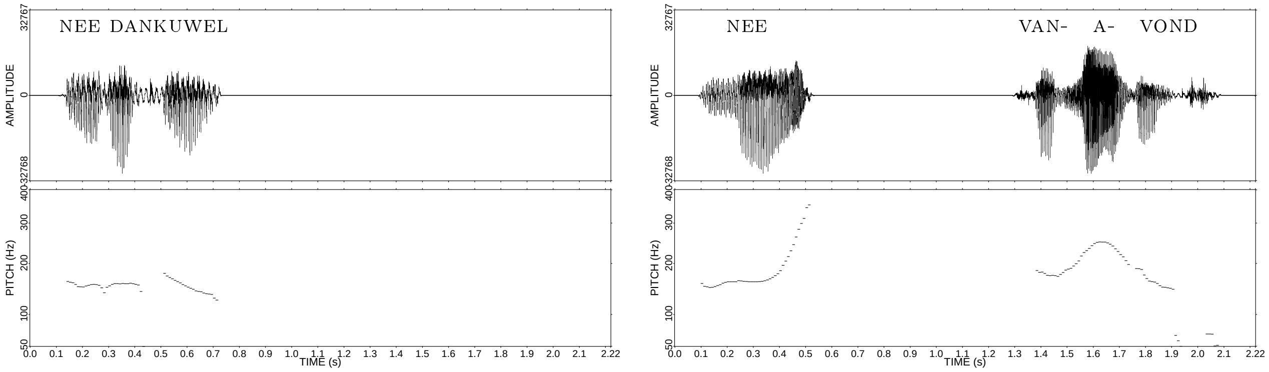

Figure 1: No+stuff responses of one speaker to two different yes/no questions from the system: left shows the 'go on' utterance "nee dankuwel"(*no thanks*) and right is the 'go back' utterance "nee vanavond" (*no tonight*).

Table 5: Average values for different features of all occurrences of "no" (single no and no+stuff). Standard deviations are given between brackets.

Feature	¬ PROBLEMS	PROBLEMS
Duration of "no" (ms)*	226 (83)	343 (81)
Preceding pause (ms)*	516 (497)	953 (678)
Following pause (ms)**	94 (93)	311 (426)
F_0 max in stuff (Hz)**	175 (37)	216 (46)
Words in stuff*	2.61 (3.65)	5.42 (8.14)

$^{*}p < 0.001$, $^{**}p < 0.05$

Let us recapitulate the findings given in table 5. First, negative signals differ from positive ones, in that the word "no" —when it occurs— in these utterances is comparatively longer. Second, compared to positive signals, there is a significantly longer delay after a problematic system prompt before users respond. Both results are in line with the data for Japanese ([11]). Third, in the no+stuff utterances, the interval between "no" and the remainder of the utterance is longer following a problematic system utterance than following an unproblematic one. Fourth, after a problematic yes/no question, the stuff part of the answer usually contains a high-pitched narrow focus accent to mark corrected information, whereas in the unproblematic case the stuff is usually prosodically unmarked. Finally, in reaction to a problem, the stuff part tends to be longer in number of words, which is in agreement with our previous, more general finding ([7]).

To illustrate some of these effects more clearly, consider figure 1 which visualizes the waveforms and corresponding F_0 contours of two typical disconfirmations produced by one of our speakers, one being a 'go on' signal (left), the other a 'go back' signal (right). Both utterances consist of a disconfirmation marker ("no") followed by stuff, but it is clear that they are realized with quite different prosody. In line with our hypothesis, the word "no" in the 'go on' case is comparatively short (185 ms), it is not provided with a prominent high boundary tone, and it is immediately followed by the stuff without a clear silence interval. In addition, the stuff part of this response does not contain a prominent pitch accent. On the other hand, the utterance on the right-hand side of the figure is a 'go back' signal and accordingly contains a relatively long "no" (441 ms), which is produced with a clear high boundary tone, and is followed by a fairly long pause of 762 ms. Note that the contour on the word "no" is of the type referred to above, L*H-H%, which does not permit a straightforward specification of pitch range. Also, the stuff contains a clear narrow focus pitch accent which serves to highlight corrected information. What cannot be derived from this figure is that in the 'go back' mode speakers generally tend to produce their responses after a longer delay than in 'go on' mode, and also that the stuff part is generally longer in words in the former case.

5. DISCUSSION

The main finding of this article can be summarized as follows: in the case of communication problems, speakers put much more prosodic effort in their reaction. If the preceding system utterance contained a problem (either a speech recognition error or an incorrect default assumption), then (1) the user's utterance of the word "no" has a longer duration, (2) there is a longer pause between the system's utterance and the user's reaction, (3) in the case of a no+stuff answer, the delay between the "no" and the stuff is longer, (4) the stuff part contains a narrow focus, high-pitched (corrective) accent and (5) the stuff contains more words. Additionally, various distributional differences between 'go on' and 'go back' signals were found: for instance, single stuff answers are solely reserved as responses to problematic system utterances and, moreover, users who respond to problematic utterances primarily use H% boundary tones.

These findings can easily be related to the respective functions of the two kinds of disconfirmation. A 'go on' disconfirmation is simply an answer to the question and does not address any underlying assumptions of the system. In principle, a single "no" is a sufficient answer. The stuff is exclusively reserved for politeness phrases, which follow more or less automatically and provide no further information. This explains the short pauses between the "no" and the stuff as well as the lack of accents in the stuff. If a yes/no question from the system contains a problem, just answering "no" might be sufficient but is not very cooperative. Assuming that the user wants the dialogue to be over as soon as possible it is more efficient to immediately *correct* the system. To do that, single stuff adequately serves the purpose, whilst an explicit "no" may be added to strengthen the problem signalling.

The findings related to prosodic effort are in line with the findings of the companion paper [7] in which it was shown that subjects use the negative ('go back') variants of the features described in table 1 more often when the preceding system utterance contains a problem, whereas the positive cues ('go on') are more often used in response to unproblematic system utterances. Taking these two results in combination, we have found evidence for the claim that people devote more effort to negative cues on various levels of communication.

An interesting question is how generalizable the prosodic results are. We contend that our findings are not specific for "no" nor for Dutch nor for the domain of train travelling. Support for this is found, for instance, in the recent collaboration of the second author with Hirschberg and Litman. One of the findings from their study of American English human-machine dialogues is that utterances following speech recognition errors can be reliably distinguished from 'normal' utterances using a set of automatically obtained acoustic/prosodic characteristics (pitch range, amplitude, timing, *inter alia*). For instance, 'corrections' appear to be more prosodically marked than other utterances (higher, longer, louder, slower, ...), which is in agreement with our current results.

The current analysis suggests that the presence of cues such as a prolonged delay before answering or a high-pitched narrow focus accent are good indicators of problems. In combination with the findings of [7], the present results provide potentially useful information for spoken dialogue systems which monitor whether or not the communication is in trouble: if a question is followed by a user's utterance which has various marked properties (such as relatively many words, disconfirmations, corrections, long delays, words with a narrow focus, high-pitched accent), the system can be fairly certain that the information it tried to verify is not in agreement with the user's intentions. If, on the other hand, the user's utterance does not contain such features, then it is highly likely that the verified information is correct. Indeed, the memory-based learning experiments mentioned in section 2 and further described in the appendix show that it is possible to predict in 97% of the cases whether or not the user signals a communication problem. Knowing whether or not there are communication problems may be very useful in a number of situations. It can be used as a basis for choosing the verification strategy employed by the system, but it may also be a cue to switch to a different recognition engine. Levow [8] found that the probability of experiencing a recognition error after a correct recognition is .16, but immediately after an incorrect recognition it is .44. This increase is probably caused by the fact that speakers use hyperarticulate speech when they notice that the system had a problem recognizing their previous utterance.

It should be stressed that before such techniques can be used in practice, the gap between the hand-annotated and interpreted data used in the analyses and the raw data which comes out of a speech recognition engine has to be bridged. We certainly do not wish to claim that all the cues discussed here and in [7] can be extracted automatically from a word graph. However, we conjecture that many cues, and certainly the most important ones, can be extracted automatically. To see whether this is indeed the case we intend to redo the experiments described here with word graphs. As said, it is likely that the percentage of cases correctly classified as problems will decrease. The prosodic cues discussed in this paper may provide a means of compensating for this loss of accuracy.

Acknowledgments The experiments with memory-based learning techniques described in the appendix were carried out with the expert help of Antal van den Bosch. The authors are mentioned in alphabetical order. Weegels and Theune were supported by the Priority Programme Language and Speech Technology (TST), sponsored by NWO (The Netherlands Organization for Scientific Research). Swerts is also affiliated with UIA and with the FWO - Flanders.

6. REFERENCES

1. Aha, D., D. Kibler, & M. Albert (1991), Instance-based learning algorithms, *Machine Learning*, 6:37-66.
2. Clark, H.H. & D. Wilkes-Gibbs (1986), Referring as a collaborative process, *Cognition* 22:1-39.
3. Clark, H.H. & E.F. Schaeffer (1989), Contributing to discourse, *Cognitive Science* 13:259-294.
4. Daelemans, W., J. Zavrel, K. van der Sloot, & A. van den Bosch (1999), *TiMBL: Tilburg Memory Based Learner, version 2.0, reference guide*, ILK Technical Report 99-01. http://ilk.kub.nl/ ilk/papers/ilk9901.ps.gz
5. Hermes D.J. (1988), Measurement of pitch by subharmonic summation, *Journal of the Acoustical Society of America* 83:257-264.
6. Hockey, B., D. Rossen-Knill, B. Spejewski, M. Stone & S. Isard (1997), Can you predict answers to y/n questions? Yes, no and stuff, *Proc. Eurospeech'97*, Rhodos, Greece.
7. Krahmer, E., M. Swerts, M. Theune & M. Weegels (1999), Problem spotting in human-machine interaction, *Proc. Eurospeech'99*, Budapest, Hungary.
8. Levow, G.-A. (1998), Characterizing and recognizing spoken corrections in human-computer dialogue, *Proc. COLING-ACL*, Montreal, Canada.
9. Oviatt, S., M. MacEachern, & G.-A. Levow (1998), Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication*, 24:87-110.
10. Soltau, H. & A. Waibel (1998), On the influence of hyperarticulated speech on recognition performance, *Proc. ICSLP'98*, Sydney, Australia.
11. Swerts M., H. Koiso, A. Shimojima & Y. Katagiri (1998), On different functions of repetitive utterances, *Proc. ICSLP-98*, Sydney, Australia.
12. Swerts, M. & M. Ostendorf (1997), Prosodic and lexical indications of discourse structure in human-machine interactions, *Speech Communication* 22:25-41.

13. Traum, D.R. (1994), *A computational theory of grounding in natural language conversation*, Ph.D thesis, Rochester.

14. Weegels, M. (1999), Users' (mis)conceptions of a voice-operated train travel information service, *IPO Annual Progress Report*, Eindhoven, The Netherlands, pp. 45-52.

15. Zipf, G.K. (1949), *Human behavior and the principle of least effort*, Addison-Wesley, Cambridge, MA.

APPENDIX: MEMORY-BASED ERROR SPOTTING

In this appendix some experiments with memory-based learning techniques for the spotting of communication problems are discussed, based on the findings of [7]. Memory-based learning techniques can be characterized by the fact that they store a representation of some set of training data in memory, and classify new instances by looking for the most similar instances in memory. In the current context an instance is the representation of an utterance pair using a vector of 14 feature value pairs. The 14 features are described in [7]: four represent properties of the systems' questions such as verification strategy and presence or absence of defaults, the ten others represent properties of the users' replies (number of words, (dis)confirmations, corrections, repetitions etc.). Various experiments were carried out on the complete set of 487 utterance pairs described in [7], each time training on 486 cases and testing on the remaining one ("leave one out"). The category to be predicted during the test phase is whether or not there are communication problems. If X is the test case, a *distance metric* $\Delta(X,Y)$ determines which group k of cases Y in memory is the most similar to X. The most frequent value for the relevant category in k is the predicted value for X. Since some features are more important than others, a weighting function w_i may be used. In sum: the distance between vectors X and Y of length n is determined by the following equation:

$$\Delta(X,Y) = \sum_{i=1}^{n} w_i \, \delta(x_i, y_i) \qquad (1)$$

where $\delta(x_i, y_i)$ gives a point-wise distance between features which is 1 if $x_i = y_i$ and 0 otherwise.

For the actual experiments we used the IB1-GR algorithm from [4]. IB1-GR is a combination of (an extension of) the instance-based learning algorithm IB1 of [1] with *gain ratio* (GR) as weighting function. The gain ratio for a feature i is derived from the *information gain* for that particular feature, computed by looking at the difference in uncertainty (*entropy*) for situations with and without feature i. A consequence of this measure is that features which have a minority of infrequent but highly informative values, and a majority of uninformative values (such as marked versus unmarked word order), tend to have low information gain, and thus mostly play a minor role in classification. Moreover, the information gain metric has a tendency to overestimate the benefits of features with a large number of values. As an extreme case, consider a feature with unique values (for the current domain, an utterance identification number between 1 and 487, say). Such a feature will have a maximal information gain, but is useless for value prediction of new cases. The gain ratio metric normalizes the information gain in this respect (for further details we refer to [4]).

Table 6: Percentages correct classifications (problems/no problems) obtained using leave-one-out on tokens with the IB1-GR algorithm.

Features	Percentage correct
All features	97%
confirm + correct	96%
correct	90%
confirm	83%

Using the IB1-GR algorithm four experiments were carried out, in which the number of features stored in memory is varied. Table 6 displays the results. The baseline strategy is always guessing that there are no problems, which would be correct for 287 of the 487 cases. Thus, the chance level lies at 59%. All experiments went well above this level, the best results being obtained using *all* features with 97% correct categorizations. In the data under consideration, the features with the highest gain ratio by far were 'confirm' (whether or not the user's utterance contains a confirmation) and 'correct' (the number of slots the user corrects). This means that these features play first fiddle when all features are considered. Looking *only* at these features leads to slightly lower percentage of correct predictions (although we should be careful to draw conclusions from that, given the relatively small amount of data). Interestingly, the two features only perform well in combination, in isolation their respective performances are much lower.

The conclusion must be that on the basis of the hand-annotated data it is very well possible to predict whether the user signals a communication problem or not. The results indicate that the presence of all features is beneficial, but the relatively small amount of data does not warrant any definite conclusions in this respect. As noted in the main text: there is a considerable gap between the hand-annotated data and the raw data coming out of a speech recognition engine. It is expected that it will be quite hard to extract certain feature values automatically from a word graph (e.g., marked word-order). However, we suspect that other (and more important) features *can* be extracted from the word graph automatically provided that the context (the preceding system utterance) is taken into account. To find out whether this is indeed the case, we intend to redo the experiments in the future with word graphs. Additionally, we want to further investigate the usefulness of prosodic cues (boundary tone, duration, F_0, pause) for error-spotting, which, we conjecture, are relatively easy to extract from the speech signal. It will be very interesting to see how the memory-based learning techniques will perform when applied directly on the level of such raw data.