CHARACTERIZATION AND PREDICTION OF DIALOGUE ACTS USING PROSODIC FEATURES

Katharina Mittelhammer[1], *Uwe D. Reichel*[2]

[1]*Institute of Phonetics and Speech Processing, University of Munich*
[2]*Research Institute for Linguistics, Hungarian Academy of Sciences*
K.Mittelhammer@phonetik.uni-muenchen.de, uwe.reichel@nytud.mta.hu

Abstract: This study investigates the classification of dialogue acts using prosodic features. Prosodic elements contain information about the type of dialogue acts. For this reason they can be used to increase the performance of speech recognition and speech synthesis systems. In the present study, two feature sets and two classifiers were compared. The first feature set comprised standard prosodic features, namely the F0 arithmetic mean, and its standard deviation. The overall dialogue act related F0 mean values support Ohala's Frequency Code hypothesis. For the second feature set, intonation features accounting for F0 shapes were derived by polynomial intonation stylization. To compare the usability of both feature sets for classification, the sample of dialogue acts was classified using k-nearest neighbor and random tree classifiers. For both feature sets and classifiers accuracies around 77% were achieved. We found no significant difference between the classifiers and feature sets. One can conclude that standard F0 features already contain a reasonable amount of information about dialogue acts, and that a more elaborated F0 analysis is not beneficial for the given data.

1 Introduction

Prosody plays an important role in the human-human-communication. Discussion partners transmit information not only via semantic but also via prosodic elements. Knowing more about prosodic realizations can help to understand the intended meaning of an utterance, as for instance, if the utterance is meant as question or request. For instance, discussion partners can notice the meaning of an utterance. The relation between fundamental frequency (F0) and the pragmatic and semantic aspects of an utterance across languages and cultures is amongst others described in terms of the Frequency Code [7]. High F0 is related to respect, politeness and submission and is used for marking questions – possibly to achieve the dialogue partners' goodwill according to [3]. Low or decreasing F0 symbolizes self-confidence, threat or authority. Such tendencies were found in human as well as in nonhuman vocalization. This shows that different prosodic realizations of one utterance can lead to conceptual differences. Discussion partners perceive prosodic features in a dialogue intuitively whereby they often notice the dialogue act type (i.e. question or request) independent of the semantic content and recognize the meaning of what is being said. This additional information source that humans perceive intuitively can be recorded and applied to improve the performance of speech recognition and speech synthesis systems. For example, the integration of prosody can improve the word recognition significantly. Furthermore, more detailed knowledge with regard to prosodic features of different types of utterance can lead to a more natural output of speech synthesis systems [14]. Several studies [15, 16] have shown that dialogue act classification can be improved using

prosodic features. Among the most common standard prosodic features are F0 mean and its standard deviation. In the current study we focus on F0-related prosodic features in order to test whether a more elaborate F0 analysis leads to higher classification accuracies compared to the standard F0 features.

2 Data

For the characterization and the prediction of dialogue acts in this study, 21 dialogues (170 minutes) of the Illinois Game Corpus [8] were used. Tangram game dialogues between students at the age of 18 to 29 were recorded. They all grew up monolingually in American English and were undergraduates by the time of recording. The applied dialogues have been manually text-transcribed, chunk-segmented and annotated in dialogue acts adapting the tag set of [4]. This tag set was developed to encode conversational moves, i.e. initiations and responses, with certain discourse purposes. For this study, the label inventory of Carletta et al. is taken as a basis and augmented by additional tags: COMMENT, COMMENT-POSITIVE, COMMENT-NEGATIVE, and OBJECT [12]. Two further added dialogue acts, namely US (unspecified) and OT (offtalk) were removed since they are not clearly defined dialogue acts and therefore are very variable. The complete label set of the sixteen dialogue acts is shown in Table 1. Please consult [4] for a more comprehensive label introduction. The data was signal-text aligned on the phoneme and word level using the WEBMAUS web service [13, 6].

dialogue act label	abbreviation
Instruct	IN
Explain	EX
Comment	CO
Comment Positive	COP
Comment Negative	CON
Align	AL
Check	CH
Query-YN	QY
Query-W	QW
Ready	RE
Acknowledgement	AC
Clarify	CL
Reply-Y	RY
Reply-N	RN
Reply-W	RW
Object	OB

Table 1 - Dialogue act labels adapted from the tagset of [4]

3 Method

Since the current study addresses dialogue act classification and not segmentation, the latter is simply taken over from the manual annotation. We comparatively evaluated two feature sets and two classifiers. The standard and the stylization-based feature sets are introduced in sections 3.2 and 3.3 The classifiers will be described in section 3.4.

3.1 F0 Preprocessing

F0 was extracted by autocorrelation (PRAAT 5.3.16 [2], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation. The contour was then smoothed by moving average filtering and transformed to semitones relative to a base value. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

3.2 Standard F0 feature set

The standard feature set contains the F0 arithmetic mean and its standard deviation for local intonation segments within a dialogue act to be defined in the subsequent section. Furthermore, the position of the local segment within the dialogue act (initial, medial or final) is added to the feature set by two boolean variable isInitial and isFinal.

3.3 F0 stylization-based feature set

The second feature set was derived from the CoPaSul intonation proposed by [10] which is illustrated in Figure 1. Within this framework intonation is stylized as a superposition of linear global contours, and third order polynomial local contours. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. The domain of local contours is determined by placing boundaries behind each content word determined by POS tagging. Thus these local contour domains roughly correspond to syntactic chunks [1] and generally contain at most one pitch accent. The global linear component is given by the F0 baseline. This line is robustly fitted to F0 values below the 10th percentile in a window shifted along the F0 contour as proposed in [11]. The baseline is then subtracted from the F0 contour, and a third order polynomial is fitted to the F0 residual within each local segment. As in [10] the global and local contour parameter vectors are clustered to derive intonation contour classes. Intonation patterns thus can be described in parametric as well as in category terms. From this stylization the following features were taken for dialogue act prediction: the 4 local contour polynomial coefficients, its contour class, and analogously to the standard feature set, the information whether the local contour is dialogue act initial, medial, or final. To allow for the comparison of the two feature sets with respect to dialogue act classification, in both sets a feature vector corresponds to a local intonation contour segment, for which is to be decided which kind of dialogue act it is part of. Both feature sets comprise 11,568 local segments labeled by 16 dialogue act targets.

3.4 Classifiers

For dialogue act classification we choose two classifiers implemented in the WEKA toolkit version 3.6 [5]: the decision tree 'random tree' and an instance-based k-nearest neighbor learner 'ibk'. All parameters were set to the default values. To correct for the very uneven dialogue act distribution the data was balanced using the WEKA filter function 'resample'. As the classification of the two feature sets were subsequently compared by a paired t-test, resampling was synchronously applied to both the feature sets.

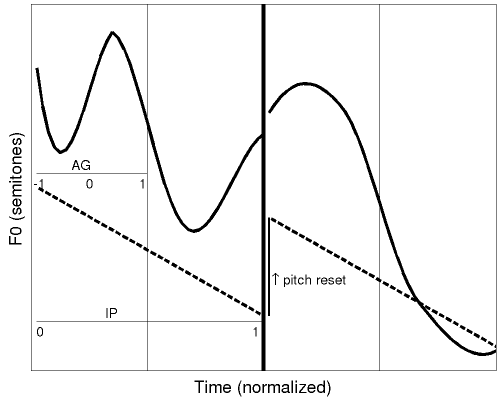

Figure 1 - CoPaSul stylization of global and local intonation components.

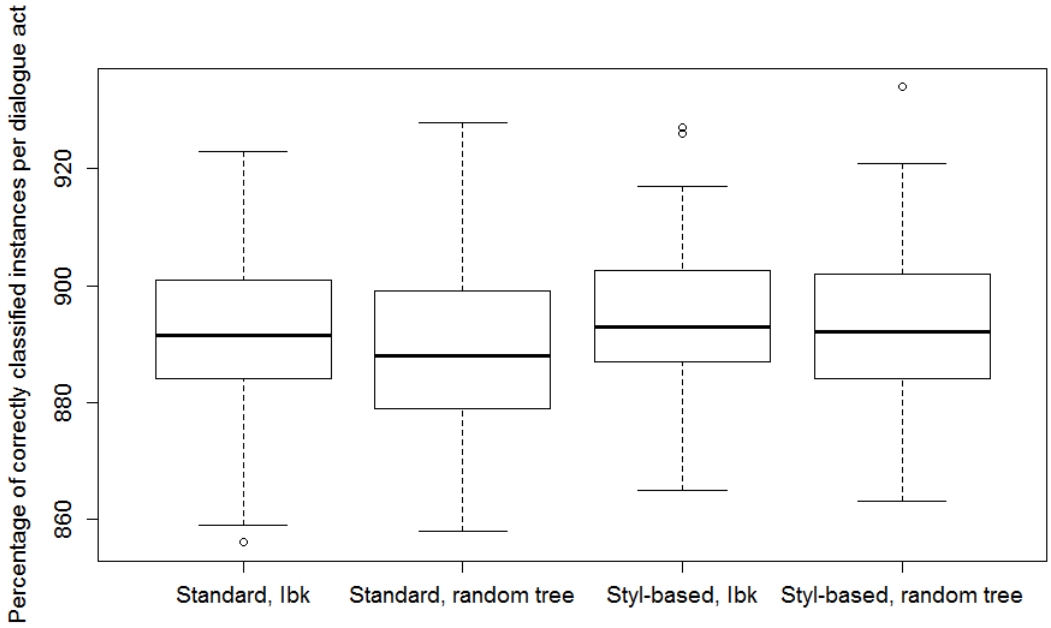

Figure 2 - Percentage of correctly classified instances per dialogue act

4 Results

The two feature sets and classifiers were comparatively evaluated by 10-fold cross-validation in the WEKA 'Experimenter' mode. Figure 2 shows the percentage of correctly classified instances per classifier and feature set on the held-out data. Furthermore, confusion matrices were created using the WEKA 'Explorer' mode, which are shown for both feature sets in Tables 2 and 3.

4.1 Standard F0 feature set

By the F0 arithmetic mean and its standard deviation, 77% of all instances were classified correctly from both the classifiers 'random tree' and 'k nearest neighbor'. The k nearest neighbor-classifier produced 892 out of 1,157 correctly classified instances, compared to 889 instances from the 'random tree'. The confusion-matrix for the k nearest neighbor classifier is depicted

in Figure 2. Rows correspond to classes, columns to classification. The confusion-matrix of thee classifier 'ibk' (Table 2) shows that the classification differs strongly depending on the dialogue act. While some dialogue acts were almost completely correctly classified, others showed numerous confusions. For example, all the instances of the dialogue acts AL and OB were matched to the correct class. Only a few confusions are caused by the classification of the dialogue acts IN, COP, and RN, however, still more than 90% of them are correctly classified. More confusions occur in the dialogue acts CH and AC. 50% of them were classified correctly. CH was incorrectly allocated to EX and QW up to 8%, AC was incorrectly allocated mostly to RE (up to 7%). The classification of the dialogue act EX seems to be the most difficult. Only 20% of this dialogue act was classified correctly. Confusions occur in all classes, especially in the dialogue acts CH and QY to 10% and more, respectively. These findings suggest that the prosodic feature F0 arithmetic mean and its standard deviation is suitable to correctly classify certain dialogue acts. Other dialogue acts such as IN, CO, COP, CON, QW, CL, RN, and RW seem to be more characteristic in terms of the average fundamental frequency than other dialogue acts like EX, AC, and RE.

IN	EX	CO	COP	CON	AL	CH	QY	QW	RE	AC	CL	RY	RN	RW	OB	
98.98	0.00	0.00	0.00	0.00	0.13	0.26	0.00	0.13	0.13	0.13	0.13	0.13	0.00	0.00	0.00	IN
2.97	24.79	6.80	1.70	3.12	0.57	11.76	10.48	9.07	3.26	5.10	7.37	4.82	1.13	6.37	0.71	EX
1.94	2.99	78.66	0.90	0.60	0.15	1.64	3.43	1.19	1.49	1.64	2.69	1.19	0.30	0.90	0.30	CO
0.41	0.27	0.14	97.56	0.41	0.00	0.00	0.14	0.14	0.14	0.14	0.00	0.41	0.14	0.14	0.00	COP
0.00	0.00	0.00	0.00	99.43	0.00	0.00	0.14	0.00	0.14	0.00	0.28	0.00	0.00	0.00	0.00	CON
0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	AL
1.42	8.24	3.84	1.85	1.42	0.00	47.87	7.81	6.25	1.99	5.26	5.82	2.84	0.99	4.26	0.14	CH
1.54	4.88	2.95	1.16	0.77	0.64	4.49	64.83	5.13	1.93	2.82	2.82	1.67	1.28	2.95	0.13	QY
1.27	4.38	2.12	0.56	0.42	0.28	6.07	4.24	67.23	1.27	1.69	4.94	1.84	0.99	2.68	0.00	QW
0.84	1.97	1.41	1.97	1.13	0.14	1.97	1.41	0.56	74.82	5.63	1.97	3.66	1.27	0.84	0.42	RE
2.44	5.70	1.90	3.12	0.95	0.00	4.61	3.26	2.44	7.33	54.27	4.48	5.02	2.04	2.31	0.14	AC
0.45	5.40	2.55	0.90	1.20	0.15	4.50	2.10	4.50	0.60	2.70	69.87	1.80	0.60	2.10	0.60	CL
0.56	2.38	1.40	1.82	0.42	0.14	0.98	2.10	1.96	3.37	3.79	1.68	77.00	0.28	1.96	0.14	RY
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	99.87	0.00	0.00	RN
0.54	2.98	0.68	0.95	0.68	0.00	2.17	1.08	1.36	1.08	0.54	2.44	0.95	0.27	84.15	0.14	RW
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	OB

Table 2 - Confusion Matrix for standard F0 feature set and classifier Ibk. Rows represent classes, columns represent classifications (indicated in percent).

4.2 F0 stylization-based feature set

The classification by the intonation stylization yielded likewise for both of the two classifier very similar results. Here again, 77% of all instances are classified correctly. Thus, classification accuracy could be not improved by using the stylization-based features (paired t-tests). The k nearest neighbor-classifier allocated 895 of 1,157 instances to the correct class, the 'random tree' 893 of 1,157 instances. The confusion-matrix of the k nearest neighbor-classifier depicted in Table 3 is illustrated in the following. As with the classification by the feature F0 arithmetic mean and its standard deviation, some dialogue acts could be classified correctly-others, however, are assigned to wrong dialogue acts. Similarly by the intonation stylization, the AL and OB were classified correctly with all of their instances. With more than 90% the dialogue acts IN, COP, CON, and RN were allocated to the the correct class. The biggest difficulty occurred in the classification of the dialogue acts AC, CH, and IN which were classified up to 50% to

the right class. Confusions were distributed over all classes. These classification results prove the ability of classifying dialogue acts by the intonation stylization. Although there are few dialogue acts which were frequently confused to others, the most dialogue acts could be classified reliable using the intonation stylization.

IN	EX	CO	COP	CON	AL	CH	QY	QW	RE	AC	CL	RY	RN	RW	OB	
99.10	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.13	0.00	0.26	0.13	0.00	0.13	0.00	0.00	IN
2.83	23.51	6.94	2.69	2.97	0.57	11.33	10.34	7.65	2.69	5.95	8.78	4.39	1.56	7.51	0.28	EX
0.60	3.58	78.06	0.75	0.60	0.30	2.39	4.03	2.09	1.34	1.79	1.64	0.60	0.30	1.94	0.00	CO
0.14	0.27	0.41	97.96	0.00	0.00	0.00	0.14	0.14	0.14	0.14	0.14	0.41	0.00	0.14	0.00	COP
0.00	0.00	0.00	0.00	99.43	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	CON
0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	AL
1.14	7.95	4.12	1.56	1.85	0.57	48.15	7.39	5.68	2.41	4.55	5.40	2.70	1.70	3.98	0.85	CH
0.77	6.03	3.59	0.77	1.16	0.51	4.62	65.08	5.39	0.77	2.05	2.70	2.05	1.03	2.82	0.64	QY
0.99	5.65	1.98	0.28	0.71	0.00	5.23	5.79	68.64	0.71	1.98	2.82	2.26	0.42	2.40	0.14	QW
1.41	2.39	1.55	2.25	0.98	0.14	2.53	1.13	0.56	71.73	5.20	1.55	5.77	1.13	1.41	0.28	RE
1.90	3.53	3.26	3.26	1.36	0.14	3.80	3.53	2.04	8.41	54.55	3.66	5.43	2.04	2.99	0.14	AC
0.60	6.30	1.65	0.45	0.30	0.15	4.95	2.40	2.55	1.05	3.00	71.36	1.65	0.15	3.15	0.30	CL
0.42	4.07	1.26	1.96	0.28	0.14	1.40	1.82	1.96	2.52	3.93	0.56	76.86	0.98	1.82	0.00	RY
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	99.87	0.00	0.00	RN
0.68	3.93	0.54	0.27	0.68	0.27	1.63	1.76	2.03	1.08	0.68	2.30	0.95	0.27	82.93	0.00	RW
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	OB

Table 3 - Confusion matrix for F0 stylization based feature set and classifier Ibk. Rows represent classes, columns represent classifications (indicated in percent).

4.3 F0 arithmetic mean and its standard deviation per dialogue act

For each dialogue act the F0 arithmetic mean and its standard deviation were calculated over all instances to allow for a dialogue act interpretation in terms of Ohala's Frequency Code which will be outlined in the Discussion below. The values are shown in Table 4.

5 Discussion

For both classifiers and feature sets classification accuracies of 77% were achieved. These results show that standard F0 features already contain a reasonable amount of information about dialogue acts, and that a more elaborated F0 analysis is not beneficial for the given data. The findings suggest that instead of a sophisticated F0 analysis features from other prosodic domains should be taken into account. Furthermore, classifiers also accounting for the dialogue act history should be employed. The results furthermore show that some dialogue acts such as AL, OB, IN COP, CON, and RN are easier to predict than others, e.g. EX, CH, and AC. This finding might be partly attributed to resampling which increases the number of dialogue acts with low occurrence but keeps the variation of their acoustic parameters small. In any case, some dialogue acts as AL and OB stand out regarding their mean F0 values which eases their correct classification. EX (EXPLAIN) turned out to be most difficult to predict which might be due to the fact that for EX much less contextual and content restrictions apply as for e.g. questions. As can be seen in Table 4 the dialogue acts show characteristic mean F0 values that can be interpreted in the terms of Ohala's Frequency code hypothesis mentioned in the Introduction. According to this hypothesis high F0 is related to respect, politeness and submission, and is used for marking questions. Low or decreasing F0 symbolizes self-confidence, threat or

arithmetic mean	standard deviation	dialogue act
11.1407338549611	1.31777990086511	AL
11.0450412019911	4.65177934450873	CON
9.97492453385058	5.34710804337955	IN
9.82111511740772	5.31529724305175	RN
9.40984560272488	5.20813257246985	CH
9.28307721322273	4.96872016293645	QY
9.27403934651614	4.96534880022615	COP
9.20629544234737	4.55100780915518	CO
9.16549417258225	5.15395787524606	QW
9.14351979405247	5.16318733687642	CL
9.12335605872035	5.1184971692763	EX
9.03513923615651	5.45581609734974	RW
8.11473442978909	4.96824440065551	RE
8.92218991742474	4.88269721812186	RY
8.84341079494335	5.36348121544495	AC
6.20592413419284	4.89337789872609	OB

Table 4 - F0 arithmetic mean and its standard deviation per dialogue act

authority. In line with this hypothesis the dialogue acts AL (ALIGN, i.e. speaker asks, whether he/she was understood) and the questions QY and QW show a higher mean F0 than the reply dialogue acts RW, RE, and RY. An explanation for the high F0 arithmetic mean of AL could be that the one dialogue partner wants to achieve the attention and compliance of the other dialogue partner. The lowest mean F0 is measured for OB (OBJECT) which might be used to express authority needed in cases of contradiction. The high mean F0 in negative comments (CON) might be explained by the state of increased arousal of the speaker [9]. Given these observed characteristic F0 mean values it can be explained that reasonable accuracies in dialog act classification are achieved based on standard F0 features only.

6 Acknowledgments

The work of the second author is funded by the Alexander von Humboldt Foundation.

References

[1] ABNEY, S.: *Parsing by chunks*. In BERWICK, R., S. ABNEY and C. TENNY (eds.): *Principle-Based Parsing*, pp. 257–278. Kluwer Academic Publishers, Dordrecht, 1991.

[2] BOERSMA, P. and D. WEENINK: *PRAAT, a system for doing phonetics by computer*. Techn. Rep., Institute of Phonetic Sciences of the University of Amsterdam, 1999. 132–182.

[3] BOLINGER, D. L.: *Intonation across languages*. In GREENBERG, J. H., C. FERGUSON and E. MORAVCSIK (eds.): *Universals of Human Language*, vol. 2: Phonology, pp. 471–524. Stanford: University Press, 1978.

[4] CARLETTA, J., A. ISARD, S. ISARD, J. KOWTKO, G. DOHERTY-SNEDDON and A. ANDERSON: *The reliability of a dialogue structure coding scheme*. Computational Linguistics, 23(1):13–31, 1997.

[5] HALL, M., E. F., G. HOLMES, B. PFAHRINGER, P. REUTEMANN and I. WITTEN: *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, 11(1), 2009.

[6] KISLER, T., F. SCHIEL and H. SLOETJES: *Signal processing via web services: the use case WebMAUS*. In *Proc. Digital Humanities*, pp. 30–34, Hamburg, Germany, 2012.

[7] OHALA, J.: *The frequency code underlies the sound symbolic use of voice pitch*. In L. HINTON, J. N. and J. J. OHALA (eds.): *Sound Symbolism*, pp. 325–347. Cambridge University Press, Cambridge, 1994.

[8] PAGE: *PAGE – Prosodic and Gestural Entrainment in Conversational Interaction across Diverse Languages*. http:page.home.amu.edu.pl, March 22nd 2015.

[9] PITTAM, J.: *Voice in social interaction*. Language and Language Behavior. SAGE Publications, Inc., 1994.

[10] REICHEL, U.: *Linking bottom-up intonation stylization to discourse structure*. Computer, Speech, and Language, 28:1340–1365, 2014.

[11] REICHEL, U. and K. MÁDY: *Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian*. In *Proc. Interspeech 2014*, pp. 111–115, Singapore, 2014.

[12] REICHEL, U., N. PÖRNER, D. NOWACK and J. COLE: *Analysis and classification of cooperative and competitive dialogs*. In *Proc. Interspeech*, p. paper 3056, Dresden, Germany, 2015.

[13] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. ICPhS*, pp. 607–610, San Francisco, 1999.

[14] SHRIBERG, E., R. BATES, A. STOLCKE, P. TAYLOR, D. JURAFSKY, K. RIES, N. COCCARO, R. MARTIN, M. METEER and C. V. ESS-DYKEMA: *Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?*. Language and Speech 41(3-4), pp. 439–487, 1998.

[15] TAYLOR, P. A., S. KING, S. D. ISARD and H. WRIGHT: *Intonation and Dialogue Context as Constraints for Speech Recognition*. Language and Speech, 41(3):493–512, 1998.

[16] WRIGHT HASTIE, H., M. POESIO and S. ISARD: *Automatically predicting dialogue structure using prosodic features*. Speech Communication, 36(1-2):63–79, 1 2002.