Lexical Structure for Dialogue Act Recognition

Pavel Král[1,2], Christophe Cerisara[1], Jana Klečková[2]

[1]LORIA UMR 7503, BP 239 - 54506 Vandoeuvre, France
Email: {kral, cerisara}@loria.fr

[2]Dept. Informatics & Computer Science, University of West Bohemia, Plzeň, Czech Republic
Email: {pkral, kleckova}@kiv.zcu.cz

***Abstract*— This paper deals with automatic dialogue acts (DAs) recognition in Czech. Dialogue acts are sentence-level labels that represent different states of a dialogue, such as questions, hesitations, ... In our application, a multimodal reservation system, four dialogue acts are considered: statements, orders, yes/no questions and other questions. The main contribution of this work is to propose and compare several approaches that recognize dialogue acts based on three types of information: lexical information, prosody and word positions. These approaches are tested on a Czech Railways corpus that contains human-human dialogues, which are transcribed both manually and with an automatic speech recognizer for comparison. The experimental results confirm that every type of feature (lexical, prosodic and word positions) bring relevant and somewhat complementary information. The proposed methods that take into account word positions are especially interesting, as they bring global information about the structure of the sentence, at the opposite of traditional n-gram models that only capture local cues. When word sequences are estimated from a speech recognizer, the resulting decrease of accuracy of all proposed approaches is very small (about 3 %), which confirms the capability of the proposed approaches to perform well in real applications.**

***Index Terms*— dialogue act, language model, prosody, sentence structure, speech recognition**

I. Introduction

This work deals with automatic dialogue act recognition from the speech signal. A *dialogue act (DA)* represents the meaning of an utterance at the level of illocutionary force [1]. For example, "question" and "answer" are both possible dialogue acts. Automatically recognizing such dialogue acts is of crucial importance to interpret users' talk and guarantee natural human-computer interactions. For instance, this information might be used to check whether the user is requesting some information and is waiting for it, or to evaluate the feedback of the user. Another application is to animate a talking head that reproduces the speech of a speaker in real-time, by giving it facial expressions that are relevant to the current state of the discourse. In the following, a Czech train ticket reservation application has been used to assess the proposed methods.

As summarized in section II, two main types of features are generally used in the literature to automatically recognize dialogue acts: word sequences and prosody. A probabilistic dialogue grammar is also often used as additional stochastic information. Word sequences are most of the time modeled by statistical n-gram models, which encode the relationship between words and dialogue acts locally. In this work, we investigate a new kind of information for dialogue act recognition, that is the words position in the utterance. In contrast to n-grams, this information is global at the sentence level. Intuitively, this information is quite important for this task, as for instance, the word "who" is often at the beginning of sentences for questions, and at other positions for declarative sentences. A standard approach that takes into account this information consists in analyzing the sentence into a syntactic tree, but such analyzers are also known to work poorly in spontaneous speech. Hence, our approach is rather based on statistical methods.

We have already studied this problem in [2], [3], and proposed two approaches to solve it. In this work, we shortly present again both methods in section III, and further propose a third one that decouples the position from the lexical models, with the objective of optimizing the available training corpus. This paper also analyzes the gain obtained by merging lexical information with prosody, and discusses the combination of the proposed dialogue act recognition approach with a state-of-the-art speech recognizer, in order to deploy this system in realistic speech-driven applications. Section IV evaluates and compares these methods. In the last section, we discuss the research results and propose some future research directions.

II. Related Work

To the best of our knowledge, there is very little existing work on automatic modeling and recognition of dialogue acts in the Czech language. Alternatively, a number of studies have been published for other languages, and particularly for English and German.

Different sets of dialogue acts are defined in these works, depending on the target application and the avail-

This paper is based on "Automatic Dialog Acts Recognition based on Sentence Structure," by P. Král, C. Cerisara, and J. Klečková, which appeared in the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, May 14-19, 2006.

able corpora. In [4], 42 dialogue acts classes are defined for English, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [5]. Switchboard-DAMSL tag-set [6] (SWBD-DAMSL) is an adaptation of DAMSL in the domain of telephone conversation. The Meeting Recorder DA (MRDA) tag-set [7] is another very popular tag-set, which is based on the SWBD-DAMSL taxonomy. MRDA contains 11 general DA labels and 39 specific labels. Jekat [8] defines for German and Japanese 42 DAs, with 18 DAs at the illocutionary level, in the context of the VERBMOBIL corpus.

These general sets are usually further reduced into a much smaller number of broad classes, either because some classes occur rarely, or because the target application does not require such detailed classes. For instance, a typical regrouping may be the following [9]:

- statements
- questions
- backchannels
- incomplete utterance
- agreements
- appreciations
- other

Automatic recognition of dialogue acts is usually achieved using one of, or a combination of the three following models:

1) DA-models of the words sequences
2) dialogue grammars that model sequences of DAs
3) DA-models based on the utterance prosody

The first class of models infers the DA from the words sequences. These models are usually either probabilistic models, such as n-gram language models [4], [10], or knowledge-based approaches, such as semantic classification trees [10].

The methods based on probabilistic language models exploit the fact that different DAs use distinctive words. Some cue words and phrases can serve as explicit indicators of dialogue structure. For example, 88.4 % of the trigrams "<start> do you" occur in English in *yes/no questions* [11].

Semantic classification trees are decision trees that operate on words sequences with rule-based decision. These rules can be either trained automatically on a corpus, or manually coded.

A dialogue grammar is used to predict the most probable next dialogue act based on the previous ones. It can be modeled by Hidden Markov Models (HMMs) [4], Bayesian Networks [12], Discriminative Dynamic Bayesian Networks (DBNs) [13], or n-gram language models [14].

Prosodic models [9] can be used to provide additional clues to classify sentences in terms of DAs. For instance, some dialogue acts can be generally characterized by prosody as follows [15]:

- a falling intonation for most statements
- a rising F0 contour for some questions (particularly for declaratives and yes/no questions)
- a continuation-rising F0 contour characterizes a (prosodic) clause boundaries, which is different from the end of utterance

In [9], the duration, pause, fundamental frequency (F0), energy and speaking rate prosodic attributes are modeled by a CART-style decision trees classifier. In [16], prosody is used to segment utterance. The duration, pause, F0-contour and energy features are used in [17], [18]. In both [17] and [18], several features are computed based on these basic prosodic attributes, for example the max, min, mean and standard deviation of F0, the mean and standard deviation of the energy, the number of frames in utterance and the number of voiced frames. The features are computed on the whole sentence and also on the last 200 ms of each sentence. The authors conclude that the end of sentences carry the most important prosodic information for DAs recognition. Furthermore, three different classifiers, hidden Markov models, classification and regression trees and neural networks, are compared and give similar DAs recognition accuracy.

Shriberg et al. show in [9] that it is better to use prosody for DA recognition in three separate tasks, namely question detection, incomplete utterance detection and agreements detection, rather than for detecting all DAs in one task.

Lexical and prosodic classifiers are combined in [4] as follows:

$$\begin{aligned} P(W,F|C) &= P(W|C).P(F|W,C) \\ &\simeq P(W|C).P(F|C) \end{aligned} \quad (1)$$

where C represents a dialogue act and W and F, which respectively represent lexical and prosodic information, are assumed independent.

III. Lexical position for dialogue act recognition

Syntax information is often modeled by probabilistic n-gram models. However, these n-grams usually model *local* sentence structure only. Syntax parsing could be used to associate sentence structures to particular dialogue acts, but conceiving general grammars is still an open issue, especially for spontaneous speech.

In our system we propose to include information related to the position of the words within the sentence. This method presents the advantage of introducing valuable information related to the *global* sentence structure, without increasing the complexity of the overall system.

A. Sentence structure model

The general problem of automatic DAs recognition is to compute the probability that a sentence belongs to a given dialogue act class, given the lexical and syntactic information, i.e. the words sequence.

We simplify this problem by assuming that each word is independent of the other words, but is dependent on its position in the sentence, which is modeled by a random variable p.

We can model our approach by a very simple Bayesian network with three variables, as shown in Figure 1. In this figure, C encodes the dialogue act class of the test sentence, w represents a word and p its position in the sentence.

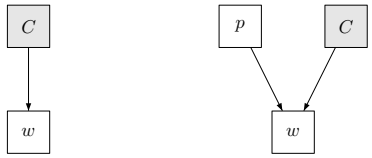

Figure 1. Graphical model of our approaches: grayed nodes are hidden

In the left model of Figure 1, $P(w|C,p)$ is assumed independent of the position: $P(w|C,p) \simeq P(w|C)$. This system only considers lexical information, and the probability over the whole sentence is given by equation 2.

$$P(w_1, \cdots, w_T|C) = \prod_{i=1}^{T} P(w_i|C) \tag{2}$$

Dialogue act recognition then consists in finding the dialogue act $\hat{C}$ that maximizes the a posteriori probability:

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|w_1, \cdots, w_T) \\ &= \arg\max_C P(C) \prod_{i=1}^{T} P(w_i|C) \end{aligned} \tag{3}$$

This system is referred to as the "unigram" or "Naive Bayes" classifier [19].

On the right part of Figure 1, information about the position of each word is included. Then, the following issues have to be solved:

- Sentences have different length.
- The new variable p greatly reduces the ratio between the size of the corpus and the number of free parameters to train.

The first issue is solved by defining a fixed number of positions N_p: N_p likelihoods $P(w_i|C,p)$ are thus computed for each sentence. Let us call T the actual number of words in the sentence. The T words are aligned linearly with the N_p positions. Two cases may occur:

- When $T \leq N_p$, the same word is repeated at several positions.
- When $T > N_p$, several words can be aligned with one position. The likelihood at this position is the average over the N_i aligned words $(w_i)_{N_i}$:

$$P(w|C,p) = \frac{1}{N_i} \sum_{i}^{N_i} P(w_i|C,p) \tag{4}$$

We propose and compare three methods to solve the second issue. The first *multiscale position* method considers the relative positions in a multiscale tree to smooth the models likelihoods. The second *non-linear merging* method models the dependency between W and p by a non-linear function that includes p. The third *best position* method decouples the positions from the lexical identities to maximize the available training corpus.

1) Multiscale position: In this approach, p can take a different number of values depending on the scale. All these scales can be represented on a tree, as shown in Figure 2. At the root of the tree (coarse scale), p can take only one value: the model is equivalent to unigrams. Then, recursively, sentences are split into two parts of equal size and the number of possible positions is doubled.

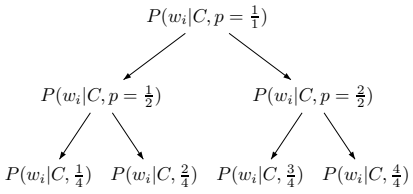

Figure 2. Multiscale position tree

For each word w_i, a threshold is applied on its number of occurrences and $P(w_i|C,p)$ for this word is computed at the finest scale that contains that minimum number of occurrences. This corresponds to the standard back-off technique [20] to solve the problem of lack of data.

Classification is then realized based on the following equation:

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|w_1, \cdots, w_T, p_1, \cdots, p_T) \\ &= \arg\max_C P(C) \prod_{i=1}^{T} P(w_i|C, p_i) \end{aligned} \tag{5}$$

where each likelihood is estimated at the finest scale possible.

2) Non-linear merging: In this approach, unigram probabilities are computed for each word and passed to a muti-layer perceptron (MLP), where the position of each word is encoded by its input index: the i^{th} word in the sentence is filled into the i^{th} input of the MLP. The output of the MLP corresponds to the *a posteriori* probabilities $P(C|w_1, \cdots, w_T, p_1, \cdots, p_T)$ and the best class is simply given by:

$$\hat{C} = \arg\max_C P(C|w_1, \cdots, w_T, p_1, \cdots, p_T) \tag{6}$$

B. Best position approach

We now give a slightly different definition for p: for any utterance W, let p be the best position amongst every possible position, i.e. the position that minimizes the DA recognition error rate.

Our objective is still to maximize:

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)} \quad (7)$$

$$= \frac{P(C)\sum_p P(W,p|C)}{P(W)} \quad (8)$$

$$= \frac{P(C)\sum_p P(W|C,p)P(p|C)}{P(W)} \quad (9)$$

Now, once the best position p has been defined for a given utterance, the decision about the winning DA class can be taken based solely on this best position:

$$P(W|C,p) = P(w_p|C)$$

where w_p is the word of the current sentence at the best position p. Hence,

$$P(C|W) = \frac{P(C)\sum_p P(w_p|C)P(p|C)}{P(W)} \quad (10)$$

Finally, maximization gives:

$$\hat{C} = \arg\max_C P(C)\sum_p P(w_p|C)P(p|C) \quad (11)$$

In this equation, the lexical likelihood $\prod_i P(w_i|C)$ used so far is replaced by the weighted sum of each word likelihood. The weights intuitively represent the importance of each position, for a given DA class.

Compared to the previously proposed solutions that take into account the global position of the words, this alternative presents the advantage of decoupling the position model from the lexical model. The lexical models $P(w_i|C)$ are thus still trained on the whole corpus, which is not divided into position-relative clusters as in the multiscale tree.

Two factors might be considered to compute these weights: they can of course be trained on a labeled corpus, but we can also use some expert knowledge to define them. For instance, it is well-known that the words at the beginning of a sentence are important to recognize questions. This expert knowledge can be easily introduced as an *a priori* probability.

A posteriori weights can also be obtained after training on a development corpus. In the following experiments, the weights are trained based on the minimum DA error rate criterion, using a gradient-descent algorithm. The initial values of the weights are obtained by first evaluating on the development corpus the DA recognition accuracy when considering only the word at position p, for every possible p. The position p that gives the best recognition accuracy represents the most important position in the sentence. The gradient descent procedure then starts from this original position.

C. Prosody

Following the conclusions of previous studies [21], only the two most important prosodic attributes are considered: F0 and energy. The F0 curve is computed from the autocorrelation function. The F0 and energy values are computed on every overlapping speech window. The F0 curve is completed by linear interpolation on the unvoiced parts of the signal. Then, each sentence is decomposed into 20 segments and the average values of F0 and energy are computed within each segment. This number is chosen experimentally [22]. We thus obtain 20 values of F0 and 20 values of energy per sentence. Let us call F the set of prosodic features for one sentence.

Two models are trained on these features and compared. The first one if a Muti-Layer Perceptron that outputs $P(C|F)$. The best class is then:

$$\hat{C} = \arg\max_C P(C|F) \quad (12)$$

The second one is a Gaussian Mixture Model (GMM) that models $P(F|C)$. The best class is then:

$$\hat{C} = \arg\max_C P(C|F) = \arg\max_C P(F|C)P(C) \quad (13)$$

D. Combination

The outputs of the lexical, position and prosodic models are normalized so that respectively approximate $P(C|W)$, $P(C|W,P)$ and $P(C|F)$.

These probabilities are then combined with a Multi-Layer Perceptron (MLP), as suggested in our previous works [23].

IV. Evaluation

A. LASER speech recognizer

The LASER (LICS Automatic Speech Extraction/Recognition) software is currently under development by the Laboratory of Intelligent Communication Systems (LICS) at the University of West Bohemia. The goal is to develop a set of tools that would allow training of acoustic models and recognition with task dependent grammars or more general language models.

The architecture is based on a so called *hybrid* framework that combines the advantages of the hidden Markov model approach with those of artificial neural networks. A typical hybrid system uses HMMs with state emission probabilities computed from the output neuron activations of a neural network (such as the multi layer perceptron).

1) Neural network acoustic model: According to many authors (see e.g. [24]) the use of a neural network for the task of acoustic modeling has several potential advantages over the conventional Gaussian mixtures seen in today's state-of-the-art recognition systems. Among the most notable ones are its economy – a neural network has been observed to require less trainable parameters to achieve the same recognition accuracy as a Gaussian

mixture model, and context sensitivity – the ability to include features from several subsequent speech frames and thus incorporate contextual information.

A three layer perceptron serves as an acoustic model in the latest version of the recognizer. It has 117 input neurons (there are 13 MFCC coefficients per speech frame and 9 subsequent frames are used as features), 400 hidden neurons and 36 output neurons corresponding to our choice of 36 context independent phonetic units (which roughly correspond to Czech phonemes). Experiments with larger hidden layer sizes have been carried out but the 400 hidden neurons were chosen as a good trade-off between modeling accuracy and computational requirements.

The incremental version of the back-propagation algorithm has been found as the fastest converging training strategy for this task. Also in order to further speed up the convergence, the cross entropy error criterion is used instead of the usual summed square error. Training this multi layer perceptron requires the precise knowledge of phoneme boundaries. These can be obtained via forced Viterbi alignment from the transcriptions of the training utterances. An already trained recognizer is necessary for this process. It is also beneficial to generate a new set of phonetic labels using the newly trained hybrid recognizer and repeat the training process once more.

Similarly to other automatic speech recognition systems, three-states HMMs phonetic units are modeled. However, all three states share the same emission probability computed from the activation value of one neuron in the output layer of the MLP. This can be viewed as a minimum phoneme duration constraint which, according to our experiments, significantly increases recognition accuracy. Because each state is tied to a neuron representing one phonetic class, the outputs of a well trained MLP can be interpreted as state posterior probabilities $P(S_j|o)$[1], which can be changed to state emission probabilities:

$$P(o|S_j) = \frac{P(S_j|o) \cdot P(o)}{P(S_j)}. \qquad (14)$$

where S_j denotes the j^{th} HMM state. The term $P(o)$ remains constant during the whole recognition process and hence can be ignored. The emission likelihoods are then computed by dividing the network outputs by the class priors (relative frequencies of each class observed in training data).

The HMM state transition probabilities are not trained since their contribution to recognition accuracy is negligible in speech recognition applications, according to our experiments. Uniform distribution is assumed instead.

2) Language model: Training words n-gram language models is not a good option in our case, because of the small size of our corpus, which is composed of manual transcriptions of a railway application (see Section IV-B). The chosen solution has been to merge words into classes and train an n-gram model based on those classes. This should compensate the lack of training data for infrequent word n-grams.

The method tries to automatically cluster words into classes according to their functional position in sentence. The algorithm (see [25]) begins with assigning each word into separate class and then starts merging two classes at a time. The process is stopped when the desired number of classes is reached. In the following experiments, the number of classes has been empirically set to 100 classes, and a trigram language model has been trained on these classes.

[1] o represents the observation, i.e. in our case the feature vector

B. Dialogue acts corpus

The Czech Railways corpus contains human-human dialogues recorded in Czech, in the context of a train ticket reservation application. The number of sentences of this corpus is shown in column 2 of Table I.

The LASER recognizer is trained on 6234 sentences (c.f. first part of Table I), while 2173 sentences pronounced by different speakers (c.f. second part of Table I) are used for testing. Sentences of the test corpus have been manually labeled with the following dialogue acts: statements (S), orders (O), yes/no questions (Q[y/n]) and other questions (Q). The word transcription given by the LASER recognizer is used to compare the performances of DAs recognition experiments with the scores obtained from manual word transcription.

All experiments for DAs recognition are realized using a cross-validation procedure, where 10 % of the corpus is reserved for the test, and another 10 % for the development set. The resulting global accuracy has a confidence interval of about $\pm$ 1 %.

DA	No.	Example	English translation
1. Training part			
Sent.	6234		
2. Testing part (labeled by DAs)			
S	566	Chtěl bych jet do Písku.	I would like to go to Písek.
O	125	Najdi další vlak do Plzně!	Look at for the next train to Plzeň!
Q[y/n]	282	Řekl byste nám další spojení?	Do you say next connection?
Q	1200	Jak se dostanu do Šumperka?	How can I go to Šumperk?
Sent.	2173		

TABLE I.
COMPOSITION OF THE CZECH RAILWAYS CORPUS

C. Sentence structure experiments

1) Multiscale position: The multiscale position approach trains a model of $P(w_i|C, p)$ at different scales, as shown in Figure 2. Recognition is then performed based on equation 5.

Figure 3 shows the recognition accuracy of this method in function of the minimum number of word occurrence at each scale: this number defines the threshold used in the multiscale tree to select the finest possible scale to

estimate the observation likelihood. The depth of the tree used in this experiment is 3, which defines 8 segments. The unigram model recognition accuracy is also reported on this figure for comparison.

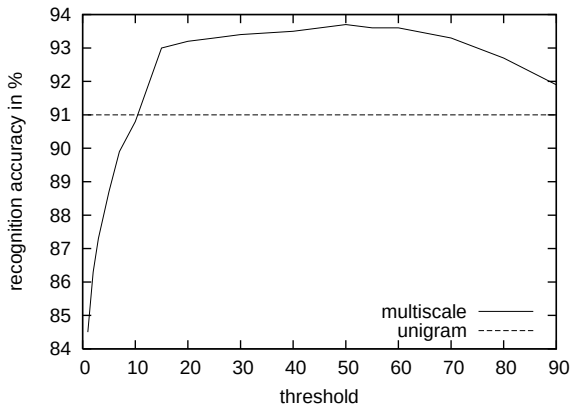

Figure 3. Dialogue acts recognition accuracy of the multiscale position system. The X-axis represents the minimum number of words in the tree, and the Y-axis plots the DA recognition accuracy

The recognition accuracy of each class is shown in Table II.

These experimental results confirm that taking into account the global position of each word improves the recognition accuracy. Furthermore, the proposed multiscale tree seems to be a reasonable solution to the issue concerning the lack of training data.

2) Non-linear merging: In the second experiment, the *Non-linear* model that merges lexical and position information is implemented by a Multi-Layer Perceptron (MLP). The chosen MLP topology is composed of three layers: 4 (for each DA class) times 8 (equal-size segments of the sentence) input neurons, 12 neurons in the hidden layer and 4 output neurons, which encode the *a posteriori* class probability. The dialogue act class is given by equation 6.

The recognition results of these methods are shown in Table II, along with the results obtained with the baseline unigram model.

3) Best position approach: The third position-based approach proposed is the *best-position* method, which recognizes dialogue acts based on equation 11. In this method, the number of positions allowed is not limited by the size of the training corpus. Hence, twenty positions (instead of eight positions for the two previous approaches) are considered.

In order to compute the initial values of the weights $P(p|C)$, recognition is first performed on the development corpus using only one position at a time:

$$P(p = i|C) = 1 \text{ and } P(p \neq i|C) = 0 \text{ for all } C$$

where i is one of the twenty possible positions. This experiment is repeated for every possible i, and the recognition accuracies obtained with each i are shown in Figure 4.

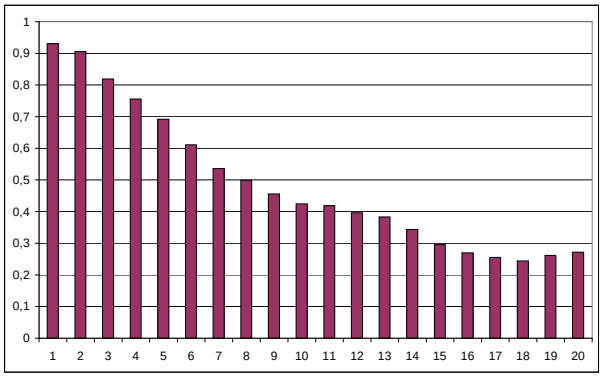

Figure 4. DA recognition accuracy on the development corpus when only a single position is considered.

Based on this experiment, the initial values chosen for the gradient descent algorithm are:

$$P(p = 1|C) = 1 \text{ and } P(p > 1|C) = 0 \text{ for all } C$$

After the gradient descent algorithm, the resulting weights are shown in Figure 5.

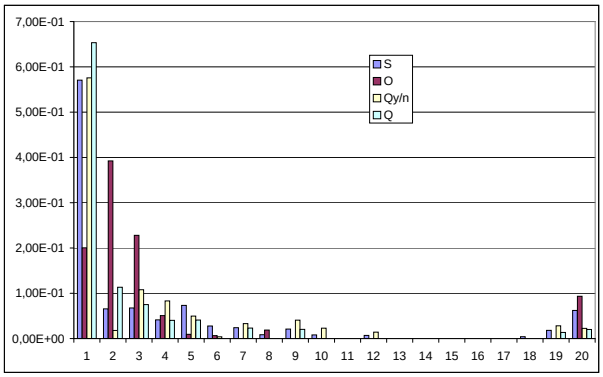

Figure 5. Weights obtained after the gradient-descent algorithm.

In this figure, it is clear that the most important positions for all DA classes are close to the beginning of the utterance. The last words of the utterance also have some importance, especially for the "order" class. The very first position is the most important for questions. These results are conforming to our intuition.

Then, using the weights shown in Figure 5, recognition is performed on the test corpus. The results are given in the fifth section of Table II.

When considering lexical information only, the best performance is obtained with the *best position* approach.

D. Prosody

The third section of Table II shows the recognition accuracy obtained when only a prosodic model is used

to classify dialogue acts. Two prosodic models are compared: the GMM (equation 13) and the MLP (equation 12).

The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The best results are obtained with the 3-mixtures GMM. It is difficult to use more Gaussians, because of the lack of training data, mainly for class O.

Although these recognition scores are much lower than the ones obtained with lexical features, it is shown next that prosody may nevertheless bring some relevant clues that are not captured by lexical models.

E. Combination

The fourth part of Table II shows the recognition accuracy when the prosodic GMM and the MLP-position models are combined with another MLP (as described in [23]).

One can conclude without loss of generality that the combination of models gives better recognition accuracy than both the lexical and prosodic models taken individually, which confirms that different sources of information bring different important clues to classify DAs.

	accuracy in [%]				
Approach/ Classifier	S	O	Q[y/n]	Q	Global
1. Lexical information					
1 Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Sentence structure					
2.1 Multiscale	94.7	70.4	96.1	95.3	**93.8**
2.2 Non-linear	90.3	83.2	91.1	98.8	**94.7**
3. Prosodic information					
3.1 GMM	47.7	43.2	40.8	44.3	**44.7**
3.2 MLP	38.7	49.6	52.6	34.0	**43.5**
4. Combination of 2.2 and 3.1					
MLP	91.5	85.6	94.0	98.7	**95.7**
5. Best position approach					
Best position	93.6	95.2	97.2	94.3	**95.8**

TABLE II.
DIALOGUE ACTS RECOGNITION ACCURACY FOR DIFFERENT APPROACHES/CLASSIFIERS AND THEIR COMBINATION WITH MANUAL WORD TRANSCRIPTION

F. Recognition with LASER recognizer

Table III shows DAs recognition scores, when word transcription is estimated by the LASER recognizer. The results are obtained with word class based trigram language model (see Section 4.2). Sentence recognition accuracy is 39.78 % and word recognition accuracy is 83.36 %.

Table III structure is the same as Table II.

The errors in transcriptions induced by the automatic speech recognizer do not have a strong impact on the results presented so far: the final accuracy only decreases from 95.7 % down to 93 %, and the ordering of the methods' accuracy is preserved. This validates the use of the proposed approaches in human-computer speech-based applications that use such a speech recognizer.

	accuracy in [%]				
Approach/ Classifier	S	O	Q[y/n]	Q	Global
1. Lexical information					
1 Unigram	93.1	68.8	94.7	86.3	**88.2**
2. Sentence structure					
2.1 Multiscale	93.8	63.2	92.9	92.9	**91.4**
2.2 Non-linear	85.5	72.0	86.8	98.0	**91.8**
3. Prosodic information					
3.1 GMM	47.7	43.2	40.8	44.3	**44.7**
3.2 MLP	38.7	49.6	52.6	34.0	**43.5**
4. Combination of 2.2 and 3.1					
MLP	88.5	77.6	90.4	97.3	**93.0**
5. Best position approach					
Best position	92.1	86.4	95.3	92.2	**93.6**

TABLE III.
DIALOGUE ACTS RECOGNITION ACCURACY FOR DIFFERENT APPROACHES/CLASSIFIERS AND THEIR COMBINATION WITH WORD TRANSCRIPTION FROM LASER RECOGNIZER

V. CONCLUSIONS

In this work, we studied the influence of word positions in a dialogue act recognition task. Two previously proposed approaches and a third new one have been described and compared, both in terms of their respective theoretical advantages and drawbacks, and also experimentally on a Czech corpus for a train ticket reservation. It has thus been demonstrated that the global position of the words in sentences is an important information that improves automatic dialogue act recognition accuracy, at least when the size of the training corpus is too limited to train lexical n-gram models with a large n, which is the most common situation in dialogue act recognition.

One of the systems that combines both lexical and position information has then been enhanced by further considering prosodic information. Yet, several prosodic models have been compared, and the combined approach still improves the results over the position and lexicon approach alone.

Finally, the manual transcription has been replaced by an automatic transcription obtained from a Czech speech recognizer, in order to validate the use of the proposed dialogue act recognition approach in realistic applications that are often based on automatic speech recognition. The resulting decrease in performances is very small, which confirms the validity of the proposed approaches.

The focus of this work has been on modeling global words position, but local statistical grammars have not been largely exploited, mainly because of the lack of training data. However, these grammars shall also bring relevant information, and it would be quite advantageous to further combine the proposed global model with such local grammars. Another important information that has not been taken into account in this work is a dialogue act grammar, which models the most probable sequences of dialogue acts. It is straightforward to use such a statistical grammar with our system, but we have not yet done so because it somehow masks the influence of the statistical

and prosodic features we focus on in this work, and also in order to keep the approach as general as possible. Indeed, such a grammar certainly improves the recognition results but is also often dependent on the target application. We also plan to test these methods on another corpus (broadcast news), another language (French) and with more DA classes.

ACKNOWLEDGMENT

This work has been partly supported by the European integrated project Amigo (IST-004182), a project partly funded by the European Commission, and by the Ministry of Education, Youth and Sports of Czech republic grant (NPV II-2C06009).

REFERENCES

[1] J. L. Austin, "How to do Things with Words," *Clarendon Press, Oxford*, 1962.

[2] P. Král, C. Cerisara, and J. Klečková, "Automatic Dialog Acts Recognition based on Sentence Structure," in *ICASSP'06*, Toulouse, France, May 2006, pp. 61–64.

[3] P. Král, J. Klečková, T. Pavelka, and C. Cerisara, "Sentence Structure for Dialog Act recognition in Czech," in *ICTTA'06*, Damascus, Syria, April 2006.

[4] A. Stolcke *et al.*, "Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech," in *Computational Linguistics*, vol. 26, 2000, pp. 339–373.

[5] J. Allen and M. Core, "Draft of Damsl: Dialog Act Markup in Several Layers," in *http://www.cs.rochester.edu/ research/cisd/resources/damsl/RevisedManual/RevisedManual. html*, 1997.

[6] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation (Coders Manual, Draft 13)," University of Colorado, Institute of Cognitive Science, Tech. Rep. 97-01, 1997.

[7] R. Dhillon, B. S., H. Carvey, and S. E., "Meeting Recorder Project: Dialog Act Labeling Guide," International Computer Science Institute, Tech. Rep. TR-04-002, February 9 2004.

[8] S. Jekat *et al.*, "Dialogue Acts in VERBMOBIL," in *Verbmobil Report 65*, 1995.

[9] E. Shriberg *et al.*, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" in *Language and Speech*, vol. 41, 1998, pp. 439–487.

[10] M. Mast *et al.*, "Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams," in *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, 1996, pp. 217–229.

[11] D. Jurafsky *et al.*, "Automatic Detection of Discourse Structure for Speech Recognition and Understanding," in *IEEE Workshop on Speech Recognition and Understanding*, Santa Barbara, 1997.

[12] S. Keizer, A. R., and A. Nijholt, "Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues," in *3rd ACL/SIGdial Workshop on Discourse and Dialogue*, Philadelphia, USA, July 2002, pp. 88–94.

[13] G. Ji and J. Bilmes, "Dialog Act Tagging Using Graphical Models," in *ICASSP'05*, vol. 1, Philadelphia, USA, March 2005, pp. 33– 36.

[14] N. Reithinger and E. Maier, "Utilizing Statistical Dialogue Act Processing in VERBMOBIL," in *33rd annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1995, pp. 116–121.

[15] R. Kompe, *Prosody in Speech Understanding Systems*. Springer-Verlag, 1997.

[16] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke., "Dialog Act Classification with the Help of Prosody," in *ICSLP'96*, Philadelphia, USA, 1996.

[17] H. Wright, "Automatic Utterance Type Detection Using Suprasegmental Features," in *ICSLP'98*, vol. 4, Sydney, Australia, 1998, p. 1403.

[18] H. Wright, M. Poesio, and S. Isard, "Using High Level Dialogue Information for Dialogue Act Recognition using Prosodic Features," in *ESCA Workshop on Prosody and Dialogue*, Eindhoven, Holland, September 1999.

[19] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue Act Classification using a Bayesian Approach," in *9th International Conference Speech and Computer (SPECOM'2004)*, Saint-Petersburg, Russia, September 2004, pp. 495–499.

[20] J. Bilmes and K. Kirchhoff, "Factored Language Models and Generalized Parallel Backoff," in *Human Language Technology Conference*, Edmonton, Canada, 2003.

[21] V. Strom, "Detection of Accents, Phrase Boundaries and Sentence Modality in German with Prosodic Features," in *Eurospeech'95*, Madrid, Spain, 1995.

[22] J. Kleckova and V. Matousek, "Using Prosodic Characteristics in Czech Dialog System," in *Interact'97*, 1997.

[23] P. Král, C. Cerisara, and J. Klečková, "Combination of Classifiers for Automatic Recognition of Dialog Acts," in *Interspeech'2005*. Lisboa, Portugal: ISCA, September 2005, pp. 825–828.

[24] H. Bourlard and N. Morgan, "Hybrid hmm/ann systems for speech recognition: Overview and new research directions," in *Summer School on Neural Networks*, 1997, pp. 389–417.

[25] J. Allen, *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, 1988.

Pavel Král is a Ph.D. candidate from University of West Bohemia at Dept. Informatics & Computer Science in Plzeň (Czech Republic) and from Henri Poincaré University in Nancy (France). He is also a lecturer at the University of West Bohemia and a member of the Speech Group at LORIA-INRIA in Nancy. His research domain is on speech recognition, more precisely on automatic dialog acts recognition.

He received his M.Sc. degree in 1999 with honours in Dept. Informatics & Computer Science at the University of West Bohemia.

Christophe Cerisara is graduated from the engineering school ENSIMAG in computer science in Grenoble in 1996, and obtained the Ph.D. at the Institut National Polytechnique de Lorraine in 1999. He worked as a researcher from 1999 to 2000 at Panasonic Speech Technology Laboratory in Santa Barbara. He is now a research scientist at CNRS, and belongs to the Speech Group in LORIA. His research interests include multiband models and robust automatic speech recognition to noise. He is the author or co-author of more than forty scientific publications.

Associated Professor **Jana Klečková** is a member of Department of Computer Science and Engineering, Faculty of Applied Sciences at the University of West Bohemia in Pilsen, Czech Republic. Her research fields are database systems, computational neuroscience (binding problem), speech recognition and understanding. She received her M.Sc. degree in 1972 at Electro technical faculty of VSSE Pilsen and her Ph.D. in 1997 at Faculty of applied sciences, UWB Pilsen.