Henri Poincaré University – Nancy 1

University of West Bohemia in Pilsen

Doctoral Dissertation
under Joint Supervision

2007   **Pavel KRÁL**

Université Henri Poincaré – Nancy 1
Département de formation doctoral en informatique

University of West Bohemia in Pilsen
Faculty of Applied Sciences

AUTOMATIC RECOGNITION OF DIALOGUE ACTS

by
Pavel KRÁL

A dissertation under joint supervision submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in "Computer Science" and "Computer Science and Engineering"

Presented and defended publicly on November 12, 2007 before the board of examiners.

Régine ANDRÉ-OBRECHT	**reviewer**	**Université Paul Sabatier**
Luděk MÜLLER	**reviewer**	**University of West Bohemia**
Jean-Paul HATON	**examiner**	**Université Henri Poincaré**
Václav MATOUŠEK	**examiner**	**University of West Bohemia**
Jan NOUZA	**examiner**	**Technical University of Liberec**
Christophe CERISARA	**examiner**	**CNRS Loria**
Yves LAPRIE	**supervisor**	**CNRS Loria**
Jana KLEČKOVÁ	**supervisor**	**University of West Bohemia**

Nancy / Pilsen 2007

Université Henri Poincaré – Nancy 1
Département de formation doctoral en informatique

Université de Bohême de l'Ouest à Plzeň
Faculté des Sciences Appliquées

RECONNAISSANCE AUTOMATIQUE DES ACTES DE DIALOGUE

par
Pavel KRÁL

Thèse en cotutelle présentée pour l'obtention du grade de Docteur de l'Université Henri Poincaré – Nancy 1 (spécialité Informatique) et de l'Université de Bohême de l'Ouest (spécialité Informatique et ingénierie)

Soutenue publiquement le 12 novembre 2007 devant la commision d'examen.

Régine ANDRÉ-OBRECHT	reviewer	Université Paul Sabatier
Luděk MÜLLER	reviewer	University of West Bohemia
Jean-Paul HATON	examiner	Université Henri Poincaré
Václav MATOUŠEK	examiner	University of West Bohemia
Jan NOUZA	examiner	Technical University of Liberec
Christophe CERISARA	examiner	CNRS Loria
Yves LAPRIE	supervisor	CNRS Loria
Jana KLEČKOVÁ	supervisor	University of West Bohemia

Nancy / Plzeň 2007

Université Henri Poincaré – Nancy 1
Département de formation doctoral en informatique

Západočeská univerzita v Plzni
Fakulta aplikovaných věd

AUTOMATICKÉ ROZPOZNÁVÁNÍ DIALOGOVÝCH AKTŮ

Ing. Pavel KRÁL

Disertační práce pod dvojím vedením k získání akademického titulu doktor v oboru „Informatika" a „Informatika a výpočetní technika"

Předneseno a obhájeno veřejně před zkušební komisí dne 12. listopadu 2007.

Régine ANDRÉ-OBRECHT   Université Paul Sabatier
Luděk MÜLLER   KKY ZČU v Plzni
Jean-Paul HATON   Université Henri Poincaré
Václav MATOUŠEK   KIV ZČU v Plzni
Jan NOUZA   Technická univerzita v Liberci
Christophe CERISARA   CNRS Loria
Yves LAPRIE   CNRS Loria
Jana KLEČKOVÁ   KIV ZČU v Plzni

Nancy / Plzeň 2007

Declaration

I submit this doctoral thesis for review and defense in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Henri Poincaré University in Nancy, France and at the University of West Bohemia in Pilsen, Czech Republic.

I declare that this doctoral thesis is completely my own work and that I used only the cited sources.

Pilsen, September 4, 2007   Pavel Král

Acknowledgements

I wish to express my thanks to Mrs. Jana Klečková and to Mr. Yves Laprie, my thesis supervisors, for their leadership during my PhD studies.

Special thanks belong to Mr. Christophe Cerisara for his support and advices during my studies and for his valuable remarks during writing this document.

I would like also thank to my family and to my partner Dana Stejskalová for their support and understanding during this studies.

My thank belong also to Mr. Michel Orlhac for his corrections of English language.

Finally, I wish to thank my colleagues from the Parole team, particularly to Emmanuel Didiot and to Joseph Razik, for their help to work in a friendly atmosphere.

iii

Abstract

This thesis deals with automatic Dialogue Act (DA) recognition in Czech and in French. Dialogue acts are sentence-level labels that represent different states of a dialogue, such as questions, statements, hesitations, etc.

The first main contribution of this work is to propose and compare several approaches that recognize dialogue acts based on three types of information: lexical, prosodic and word positions. These approaches are tested on the 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 a sentence, at the opposite of traditional n-gram models that only capture local cues. We propose three approaches to model this information: the first one, the *multiscale position* approach, exploits a description of the sentence at several levels and smoothes the probabilities across these levels. The second one, the *non-linear merging* approach, models the dependency between the words in the sentence and their position with a Multilayer Perceptron. The third one, the *best position* approach, exploits the Bayesian framework and assumes conditional independence between the words and their position to infer the probability of the dialogue act. We also propose a solution to the lack of training data problem, which is a common issue in DA recognition systems. We develop the *clustered unigram model*, which clusters the words in the sentences into several groups by maximizing mutual information between two neighbor word classes. We show that this method is especially efficient when the DA corpus is small. 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 to perform well in real applications.

One of the main issue in the domain of automatic dialogue act recognition concerns the design of a fast and cheap method to label new corpora. The next main contribution is to apply a general semi-supervised training approach based on the Expectation Maximization algorithm to the task of labeling a new corpus with pre-defined DAs. We further propose to filter out incorrect examples with two confidence measures, the *maximum a posteriori probability* and the *a posteriori probability difference* methods. Experimental results show that the proposed method is an interesting approach to create new dialogue act corpora at low costs.

Resumé

Ce mémoire concerne la reconnaissance automatique des Actes de Dialogues (AD) en tchèque et en français. Les actes de dialogues sont des unités au niveau de la phrase qui représentent les différents états d'un dialogue, comme par exemple les questions, les affirmations, les hésitations, etc.

La première contribution de ce travail est de proposer et comparer plusieurs approches de reconnaissance des actes de dialogues qui sont basées sur trois types d'informations : lexical, prosodique et relative à la position des mots dans une phrase. Ces approches ont été testées sur un corpus tchèque de dialogues entre utilisateurs et personnels dans le domaine de la réservation de billets de chemins de fer. Ce corpus a été transcris en mots manuellement, et avec un moteur de reconnaissance automatique afin de valider les approches dans des conditions réelles. Les résultats expérimentaux confirment que chaque type d'attributs (lexical, prosodique et syntaxique de position) apporte des informations pertinentes et complémentaires. Les méthodes proposées exploitant la position des mots dans la phrase sont particulièrement intéressantes, parce qu'elles utilisent une information globale sur la structure de la phrase, alors que les modèles statistiques traditionnels de type n-gram modélisent seulement les dépendances locales. Nous avons proposé trois modèles de ce type : la première approche, *position multi-échelle*, décrit une phrase sur plusieurs niveaux et lisse les probabilités au travers de ces niveaux. La deuxième approche, *fusion non-linéaire*, modélise la dépendance entre les mots dans une phrase et leur position avec un réseau de neurones de type perceptron multi-couches. La troisième approche, *meilleure position*, utilise un formalisme bayésien : elle suppose l'indépendance conditionnelle entre les mots et leur position dans une phrase pour inférer la probabilité a posteriori d'un acte de dialogue étant donnés les mots et leurs positions. Nous proposons aussi une solution au problème du manque de données pour l'apprentissage, qui est un problème très courant dans les systèmes de reconnaissance automatique des actes de dialogues. Un *modèle unigramme de classes* a été notamment développé dans ce but. Ce modèle rassemble les mots des phrases dans plusieurs groupes en maximisant l'information mutuelle entre les classes de mots voisins. Nous avons démontré que cette méthode est particulièrement efficace sur un petit corpus d'AD. Si les séquences de mots sont estimées par un moteur de reconnaissance automatique, la précision de toutes les approches proposées ne diminue que très peu relativement au cas idéal de la transcription manuelle. Ceci confirme la validité et l'applicabilité des approches proposées dans des applications réelles.

Une autre contribution conséquente, également relative au manque de corpus étiquetés

dans le domaine de la reconnaissance automatique des actes de dialoques, concerne le développement et l'étude de méthodes d'étiquetage semi-automatique de nouveaux corpus. Cette méthode est basée sur l'algorithme d'Espérance-Maximisation avec des AD prédéfinis spécifiques à la tâche visée. Nous proposons deux mesures de confiance pour sélectionner les exemples qui ont le plus de chance d'être classifiés correctement : la première utilise le critère de maximisation de la probabilité *a posteriori*, et la seconde un critère basé sur une différence de probabilités *a posteriori*. Les résultats expérimentaux démontrent que la méthode proposée est une approche intéressante pour la création de nouveaux corpus d'actes de dialogues à moindre coût.

Abstrakt

Tato dizertační práce se zabývá automatickým rozpoznáváním Dialogových Aktů (DA) v českém a francouzském jazyce. Dialogové akty jsou větné jednotky, které reprezentují různé stavy dialogu, jako např. otázku, sdělení, rozpaky, atd.

Prvním hlavním přínosem této práce je návrh a srovnání několika různých přístupů rozpoznávání dialogových aktů využívajíce tři druhy informace: lexikální, prozodickou a pozici slov ve větě. Navržené přístupy byly testovány na korpusu pro České dráhy (ČD), který obsahuje rozhovory lidí v přirozeném jazyce. Korpus je pro srovnání účinnosti metod transkribován ručně a pomocí automatického rozpoznávače řeči. Výsledky pokusů potvrdily, že příznaky každého typu (lexikální, prozodické i pozice slov) přináší důležitou a vzájemně se doplňující informaci. Navržené metody, které využívají pozici slov ve větě jsou velmi zajímavé, protože přinášejí informaci o globální struktuře věty, zatímco tradiční statistické modely typu n-gram modelují pouze lokální závislosti slov ve větě. Globální pozici slov ve větě modelují tři navržené metody. První metoda, *multiscale position*, využívá popis věty na několika úrovních a vyhlazuje pravděpodobnostní odhady mezi těmito úrovněmi. Druhá metoda, *non-linear merging*, modeluje závislost mezi slovy a jejich pozicí ve větě pomocí neuronové sítě typu vícevrstvý perceptron. Třetí metoda, *best bosition*, využívá Bayesovský rámec a k odvození pravděpodobnosti dialogového aktu předpokládá nezávislost mezi slovem a jeho pozicí ve větě. Navrhli jsme též řešení problému s nedostatkem trénovacích dat, což je jedním z úskalí systémů pro rozpoznávání dialogových aktů. Vyvinutá metoda, *clustered unigram model*, shlukuje slova ve větě do skupin na základě maximalizace vzájemné informace mezi dvěma sousedními slovními třídami. Ukázali jsme, že tato metoda je zvláště účinná, pokud máme k dispozici pouze malý DA korpus. Pokud jsme použili slovní sekvence získané pomocí automatického rozpoznávače řeči, přesnost všech našich přístupů zůstala téměř shodná jako v případě použití manuální transkripce (pokles pouze o 3%). Tento výsledek potvrdil schopnost navržených metod fungovat spolehlivě i v reálných aplikacích.

Jeden z hlavních nedostatků v oblasti automatického rozpoznávání dialogových aktů se týká nedostatku trénovacích dat a návrhu rychlé a levné metody pro značkování nových korpusů dialogovými akty. Dalším hlavním přínosem této práce je použití obecné metody trénování s učitelem i bez, která je založena na algoritmu Expectation Maximization, v úloze značkování nového korpusu předdefinovanými dialogovými akty. Zde jsme navrhli dvě metody míry důvěry na odstranění prvků, které by mohly být klasifikovány nekorektně. Metody se nazývají: *maximum a posteriori probability* a *a posteriori probability difference*. Výsledky experimentu ukázaly, že navržená metoda je účinným přístupem pro rychlou a levnou tvorbu korpusu dialogových aktů.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modeling and automatically identifying the structure of spontaneous dialogues is very important to better interpret and understand them. The precise modeling of spontaneous dialogues is still an open issue, but several specific characteristics of dialogues have already been clearly identified. Dialogue Acts (DAs) are one of these characteristics.

Austin defines in [6] the dialogue act as the meaning of an utterance at the level of illocutionary force. In other words, the dialogue act is the function of a sentence (or its part) in the dialogue. For example, the function of a question is to request some information, while an answer shall provide this information.

Table 1.1 shows an example of the beginning of a dialogue between two friends, with Peter (A) calling Michal (B) on the phone. The corresponding DA labels are also shown. Each utterance is labeled with a unique DA.

Speaker	Dialogue Act	Czech	French	English
A	Conventional-opening	Haló!?	Hallo!?	Hallo!?
B	Conventional-opening	Ahoj Petře!	Bonjour Pierre!	Hi Peter!
B	Statement	To jsem já, Michal.	C'est moi, Michel.	It's me, Michael.
B	Question	Jak se máš?	Ca va?	How are you?
A	Conventional-opening	Čau Michale!	Salut Michel!	Hello Michael!
A	Statement	Moc dobře.	Très bien.	Very well.
A	Question	A ty?	Et toi?	And you?
B	Statement	Taky dobře.	Aussi bien.	I'm well too.

Table 1.1: Example of the beginning of a dialogue between persons A and B in Czech, French and English with the corresponding DA labels.

1.1 Applications

There are many applications of automatic dialogue acts detection. We mention here only the most important ones: dialogue systems, machine translation, Automatic Speech Recognition (ASR), topic identification [41] and animation of talking head.

In dialogue systems, DAs can be used to recognize the intention of the user, for instance when the user is requesting some information and is waiting for it, or when the system is trying to interpret the feedback from the user. An example of dialogue management system that uses DA classification is the *VERBMOBIL* [2] system.

In machine translation, dialogue acts can be useful to choose the best solution when several translations are available. In particular, the grammatical form of an utterance may depend on its intention.

Automatic detection of dialogue acts can be used in ASR to increase the word recognition accuracy, as shown for example in [127]. In this work, a different language model is applied during recognition depending on the actual DA.

A talking head is a model of the human head that reproduces the speech of a speaker in real-time. It may also render facial expressions that are relevant to the current state of the discourse. Exploiting DA recognition in this context might make the animation more natural, for example by raising the eyebrows when a question is asked. Another easier option is to show this complementary information with symbols and colors near the head.

1.2 Motivations

Recognizing dialogue acts can be seen as the first level of dialogue understanding and is an important clue for applications, as it has been shown in the previous section. However, this information is often missing in the current systems. Our first motivation is thus to implement a module to automatically detect DAs, which can be easily integrated into different systems, and particularly into dialogue systems and animated talking heads.

One of the main issues with DA recognition comes from the fact that the optimal DA tag-set is usually not the same for different applications. Hence, manual labeling of a corpus is usually required every time DAs are considered for use in a new system. But manual labeling is a very time-consuming and expensive task. Therefore, our next motivation to propose and implement a method for semi-automatic corpus labeling. This method should be as general as possible to be able to create, at a low cost, new DA corpora in several languages, with several DA tag-sets.

1.3 Objectives

This memory deals with automatic dialogue act recognition in Czech and in French. The main goal is to study the existing dialogue act recognition approaches and to propose new

approaches that address some of their limitations. Different kinds of information can be used to recognize DAs. Another goal is thus to study the existing classifier combination methods and to propose and implement some original solutions to improve the accuracy of recognition.

The final objective is to make our proposals applicable in different languages and tasks, so that our work can be applied in other settings than the particular experimental setup developed in this thesis. Our third and last general objective is thus to propose solutions to facilitate the development of new DA recognition systems at a low cost.

1.4 Contributions

The main contributions proposed in this thesis for automatic dialogue act recognition are summarized below:

- Proposition of three new dialogue act recognition approaches based on lexical information and word position in the utterance:
    - *multiscale position*,
    - *non-linear merging* and
    - *best position approach*.
- Proposition of a new dialogue act recognition model, the *clustered unigram model*, based on word clustering.
- Analysis and comparison of several methods of classifier combination for DA recognition.

The main contributions in semi-automatic labeling are:

- Proposition of a new DA tag-set for radio broadcast news, based on the Dialogue Act Markup in Several Layers (DAMSL) [3] and Meeting Recorder [36] projects.
- Proposition and implementation of two confidence measures methods: maximum *a posteriori* probability and *a posteriori* probability difference.
- Use of these confidence measure methods to improve the performances of the Expectation Maximization (EM) algorithm for semi-supervised dialogue act tagging.
- Semi-automatic creation of a new French DA corpus based on the ESTER [34] corpus.

1.5 Framework

This work is developed in the context of two platforms: the first one at the Henri Poincaré University in Nancy in France and the second one at the University of West Bohemia in Pilsen in the Czech Republic.

The first potential outcome of this work concerns the design and implementation of a software for the deaf and hearing-impaired children to help them to better understand the teacher at school and to facilitate their integration in classrooms with normal-hearing children.

The software is based on the following principle: a microphone captures the speech signal of the teacher, which is then passed to a phonetic speech recognizer. The sequence of phones recognized by the system is then translated into "Langage Parlé Completé"(LPC [26], Cued Speech in English), which is a visual representation of the phonetic content of the sentence. This representation, well-known by part of the deaf community, is based on lips movements enriched by hands and fingers positions. In the laptop used by a child, a 3D talking head [69] reproduces these lips and hand movements. The information about the DA type will be used to enrich the LPC transcription that appears on the laptop screen, for example by displaying a DA type mark near the talking head. Another possibility is to animate the face in function of the DA (for example by displaying different types of eyebrows in function of the current DA).

The second outcome of this work is the creation of a dialogue system that could be used in the Czech railway stations. This system shall be able to communicate with the passengers with a limited vocabulary in natural language. The passengers can ask for train departure or arrival times. This system will also be able to reserve and buy train tickets.

1.6 Thesis Structure

The first chapter presents an introduction about the importance of dialogue act recognition with its main applications, our motivations, objectives and main contributions.

Chapter 2 presents the state of the art in the dialogue act domain. It defines the concept of a dialogue act. Then, several DA tag-sets are cited with a particular focus on the Meeting Recorder DA (MRDA) tag-set. Knowledge sources that are used for DA recognition are described next. In particular, Sections 2.7 and 2.8 summarize the existing DA recognition approaches, while Section 2.9 discusses several methods of classifier combination for DA recognition.

Chapter 3 focuses on our main contributions about dialogue act recognition. It deals with three proposed lexical and syntactic approaches that model utterance structure from the words and their position. The fourth approach, the clustered n-gram model, is based on word clustering and is described next. Our prosodic features and models are also described. Several methods that combine the individual outputs of both types of approaches (lexical and prosodic) are described in Section 3.5.

Chapter 4 deals with the experimental validation of the proposed approaches. The methods that are evaluated respectively exploit lexical information (with and without sentence structure), prosody information and a combination of both. They are evaluated in two cases: with manual words transcription and with the transcription obtained from the LASER speech recognizer.

Chapter 5 deals with our proposal for semi-automatic labeling of the DA corpus. Our DA tag-set is defined and an initial DA corpus is created. Then, the semi-supervised training algorithm is proposed, developed and evaluated.

Chapter 6 discusses the research results and proposes some future research directions.

Chapter 2

State of the Art

2.1 Introduction

In this chapter, we summarize the main previous studies in the automatic DA recognition domain. First, *dialogue acts* are described and the main DA tag-sets are presented with the description of DAs that are used in our work. Next, the *sentence modality* term is described with its relation to DAs. Section 2.5 describes the three main information sources used to recognize dialogue acts: lexis (and syntax), prosody and the dialogue context. The existing approaches of automatic DAs recognition are summarized and described in Sections 2.7 and 2.8. The last section deals with classifiers combination (in the general case and for DA recognition).

2.2 Dialogue Acts

Generally speaking, a dialogue can be viewed as a sequence of complex elements of communicative behavior, intended to change the dialogue context. These elements are called the Dialogue Acts (DAs). Several different definitions for DAs have further been proposed:

Dialogue acts are the functional units used by the speaker to change the context. These functional units do not correspond to natural language utterances or other instances of communication in a simple way, because utterances in general are multifunctional [18].

A dialogue act represents the meaning of an utterance at the level of illocutionary force [6] or a DA is approximately the equivalent of the speech act [109].

A speech act is the action performed by means of a language, such as describing something ("It is snowing."), asking a question ("Is it snowing?"), making a request or an order ("Could you pass the salt?", "Drop your weapon or I'll shoot you!"), or making a promise ("I promise I'll give it back.") [105].

A dialogue act is characterised by three properties:

- Communicative function
- Semantic content
- Utterance form

For example, the dialogue act "Does it snow?" takes the communicative function yes/no question, the semantic content "it is snowing" and the utterance form "Does it snow?". The communicative function informs about the way of the context changes, provided that the semantic content is given.

Each context takes global and local views. The global view remains constant from the beginning of the dialogue while the local view keeps changing. Bunt distinguishes five types of contexts [18]:

- **Linguistic context** is constituted from the previously pronounced text (local). It contains also the language used by the participant of dialogue (global).
- **Semantic context** is formed by the objects concerning the task. The global view generally includes the task. The local views are composed of specific elements, such as the state of the task at a given time.
- **Cognitive context** consists of the aptitudes, of the goals and of the confidences of the dialogue participants.
- **Physical and perceptual context** is characterised by the place and time. Its characteristics are for example, whether the inter-actors see themselves or not, or the type of communication channel that can be used, etc.
- **Social context** consists of the type of interactive situation and the roles of the participants in that situation. The institutional context is the global view of this context as well as the social status of the inter-actors. The local view means the action performed by the obligation and by the rights to answer in function to the local linguistic context.

Generally, a dialogue can not change all contexts on request. Only the linguistic, cognitive and local social contexts can be modified during a dialogue. Furthermore, local views are usually easier to change than global views.

The authors in [18] make a distinction between Task-Oriented (TO) and Dialogue Control (DC) dialogue acts. Task-oriented DAs allow to change the semantic context, while dialogue control DAs allow to change the social or physical context. Several DAs can be classified as informative DAs that correspond to information queries (such as questions), or to propositions (such as inform or answer). Informative DAs and Dialogue Control DAs form two distinct categories. For example, "It is the energy." is an informative *task oriented* utterance and "I can't hear you." is an informative *dialogue control* utterance.

2.3   Dialogue Act Tag-set

Before performing DA recognition, it is necessary to define a DAs tag-set. This is a very difficult task, because of two important requirements: the DAs tag-set should be generic enough to be applicable to many different problems; and the DA tags definition must be clear enough in order to be easily separable, which maximizes the agreement between the human labelers.

Therefore, there is no general DAs tag-set in the literature. Most researchers define their own DAs tag-sets, which is usually derived from one or more existing DAs tag-sets.

The most common DAs tag-sets are the Dialogue Act Markup in Several Layers (DAMSL) [3], Switchboard SWBD-DAMSL [57], Meeting Recorder [36] and VERBMOBIL [52] DAs tag-sets.

DAMSL, which was initially designed to be universal, is the most popular DA taxonomy. Its annotation scheme is composed of four levels (or dimensions): communicative status, information level, forward looking functions and backward looking functions. Generally, the dimensions are orthogonal and it is possible to find examples of any possible combination of labels. Communicative status states whether the utterance is uninterpretable, abandoned or is a self-talk. This feature is not used for most of the utterances. Information level provides an abstract characterization of the content of the utterance. It is composed of four categories: task, task-management, communication-management and other-level. The forward looking functions are organized into a taxonomy, in a similar way as actions in traditional speech act theory. The backward looking functions show the relationship between the current utterance and the previous dialogue, such as accepting a proposal or answering the question. DAMSL is composed of 42 DA classes.

SWBD-DAMSL is the application of DAMSL in the domain of telephone conversation. Usually, there is a correspondence between the SWBD-DAMSL and DAMSL labels. First, the dialogue utterances have been labeled with 220 tags. 130 of those labels that occurred less than 10 times have been clustered, leading to 42 classes.

The Meeting Recorder DA (MRDA) tag-set is based on the SWBD-DAMSL taxonomy. The MRDA corpus contains about 72 hours of naturally occurring multi-party meetings manually-labeled with DAs and adjacency pairs. Meetings involve regions of high speaker overlap, affective variation, complicated interaction structures, abandoned or interrupted utterances, and other interesting turn-taking and discourse-level phenomena. The tags are not organized anymore on a dimensional level (such as DAMSL), but the correspondences are rather listed at the tag level. Each DA is described by one *general* tag, which may be for several DAs completed by one (or more) *specific* tag. A specific tag is used when the utterance cannot be sufficiently characterised by a general tag only. For example, the utterance "Just write it down!" is characterised by one general tag *statement* and by an additional specific tag *command*. MRDA contains 11 general tags and 39 specific tags.

The DA hierarchy in VERBMOBIL is organized as a decision tree. This structure is chosen to facilitate the annotation process and to clarify relationships between different DAs. During the labeling process, the tree is parsed from the root to the leaves, and

a decision about the next branch to parse is taken at each node (c.f. Figure 2.1).
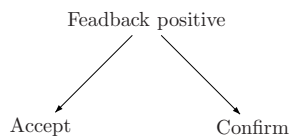
Figure 2.1: Part of the DAs decision tree hierarchy.

42 DAs for German, English and for Japanese are defined in VERBMOBIL with 18 DAs at the illocutionary level.

The DA tag-set used and defined in this work is based on a some important DAs that are described next in details.

2.3.1 Statements

The primary goal of statements is to make claims about the world as in utterances such as "Il neige, aujourd'hui." (*It is snowing, today.*) or "Mám rád sýr." (*I like cheese.*), and as in answers to questions. Usually, the content can be evaluated as being true or false. As in SWBD-DAMS, statements can be distinguished between **statement-non-opinion** and **statement-opinion**. Statement-opinions are explicitly expressing the opinion (or idea) of the speaker while statement-non-opinions are not.

Examples of statement-opinions are: "Nemohu si představit králíky jako domácí mazlíčky." (*I cannot imagine rabbits as domestic pets.*) or "Je pense que c'est bien." (*I think it is right.*). The statement-non-opinion tag is used for instance when the speaker is telling a story and the topic is personal: "J'ai quatorze ans et j'habite à Plzeň." (*I am fourteen and I'm living in Plzeň*) or "Máme kočku. Je jí asi pět let." (*We have a cat. She is probably five years old.*). However, as noted in [57], it is not clear yet that such a distinction between statements is really useful.

2.3.2 Questions

A question is an utterance normally used by a speaker to request some information from the listener. This information usually takes the form of an answer. Alternatively, one may state that the question is the request itself, which is expressed by the interrogative sentence. In the following, we will use the former definition. Questions resemble other requesting expressions as well as commands, which all normally elicit a response. Several types of questions are defined in the literature, but we describe next only two of them, which are the most common in dialogues.

Yes/no questions are questions, which possible answers are "Yes" or "No". They are usually characterised by an inversion of subject-verb or by a characteristic intonation. Examples of yes/no questions are: "Jsi to ty?" (*Is it you?*) or "Pleut-it?" (*Is it raining?*).

The DA class **Wh-question** contains questions that require a specific answer. The question usually contains a "wh" word such as: what, which, where, when, who, why, or how. Note however that some questions with a "wh" word are not wh-questions, for instance open-ended questions[1]. Wh-questions are for example: "Quelle heure est-il?" (*What time is it?*) or "Jak se máte?" (*How are you?*).

For information about other types of question, please refer to [36].

2.3.3   Action Motivators

This group contains utterances that are at the origin of future actions. Such a future action may happen immediately or after a long time. Action motivators can be classified into *commands*, *suggestions* and *commitments*. Following the MRDA tag-set classification, this DA class is usually considered as a special case of statements or questions. Therefore, it is represented by one general DA tag and the specific DA tag action motivator.

The DA class **Commands** imposes somebody to make something. It may be syntactically in the form of a question, e.g. "Chtěl bys zavřít dveře?" (*Do you want to close the door?*) or as a statement, e.g. "Fermez la porte!" (*Close the door!*). Commands are often confused with suggestions. The distinction between both DAs may be made based on the kind of responses or the social role of the speaker. Thus, rejecting a suggestion is not considered as impolite as rejecting a command. This may be considered when manually labeling the corpus. The role of the speaker has also some impact on classification, as suggestions made by the speaker who is running a meeting are often considered as commands.

The **Suggestion** tag marks proposals, offers, advices, and most obviously, suggestions. This class is often using the form "maybe we should ..." or "perhaps I can ...". Examples of suggestions are: "Voulez-vous courir aujourd'hui?" (*Do you want to run today?*) or "Snad to můžeme vzít teď." (*Maybe we can take it at once.*).

Commitments are utterances, where a speaker commits himself to some course action in the future. The main difference with suggestion is about the degree of certitude that the action will be realized. Examples of commitments are: "Budu na tom pracovat." (*I will work on that.*) or "Je vous attendrai." (*I will wait for you.*).

2.3.4   Backchannels and Acknowledgments

This group contains utterances that are most often responses and that usually confirm to the speaker who has the floor, i.e. the speaker who is currently talking, that the listener is listening and understanding. Generally, they do not elicit feedback, and the purpose of these DAs is not to interrupt the speaker who has the floor. This broad DA class contains: backchannels, acknowledgments, assessments/appreciations and rhetorical question backchannels.

---

[1]An open-ended question is a question, designed to encourage a full, meaningful answer using the subject's own knowledge and/or feelings [36].

Backchannels are utterances said by the listener to inform that the listener is following the discourse of the speaker. It is usually not possible to identify backchannels from the vocabulary only, because they share some words with other DAs, such as: accepts, floor holders, etc. Additional features are thus required to identify backchannels, e.g. the context of the DA or the corresponding audio record. Backchannels are often confused with acknowledgments and accepts. A useful contextual information to distinguish them comes from the speaker who has the floor: accepts agree the previous utterance of the speaker, and they generally occur at the end of these utterances. Acknowledgments occur usually after the end of the utterance of the speaker, and they confirm the semantic meaning of the previously pronounced content. Common backchannels are for example: "Uh-huh", "Huh" or "Hm".

Acknowledgments correspond to expressions that confirm the previous utterance (or its significant part) pronounced by another speaker. They are neither positive nor negative, in the sense that acknowledgments serve to confirm, not to agree or disagree. Common acknowledgments are for example: "Okay", "Oui" (*Yes*), "Je vois" (*I see*) or "Souhlasím" (*All right*).

2.3.5 Responses

This group contains utterances that represent the reaction to the previously pronounced utterance (or utterances) in the dialogue[1]. It is divided into three main groups: positive, negative and uncertain responses. The most important responses are described as follows:

Accepts (sometimes also called **Agreements**) are positive utterances that express agreement to or acceptance of a previous question, proposal or statement. Accepts are usually short utterances. They are often confused with backchannels and acknowledgments (c.f Subsection 2.3.4), because they share a very similar vocabulary. To distinguish accepts from other DAs, it is useful to look at the context of the utterance and to listen to the corresponding audio record. Usually, accepts have much more energy and are more assertive than backchannels and acknowledgments. Examples of accept are: "Oui volontiers" (*Yes with pleasure*) or "Presně tak" (*Exactly*).

Rejects are utterances that contain negative reactions to questions, proposals or statements. Common rejects are: "Ne" (*No*) or "Non pas du tout" (*No not at all*).

The **maybe** tag marks utterances that express the probability or possibility about the content of the previously pronounced utterance (or utterances). Probability or possibility can be represented by the word "maybe" or by other words denoting incertitude. Maybes are often confused with suggestions, such as "maybe we should ...". Examples in context (maybe after a question) are: "Savez-vous? Peut-être." (*Do you know? Maybe.*) or "Jakým přízvukem mluvíte? Pravděpodobně západním." (*What accent are you speaking? Probably western.*).

---

[1]Note that following the MRDA tag-set classification, this DA class is usually considered as a special case of statements.

2.3.6 Floor Mechanisms

Floor mechanisms contains the DAs pertaining to the mechanisms of grabbing or maintaining the floor. The main DAs in this broad class are the following:

Floor grabbers usually occur at moments without speech, when a speaker wants to gain the floor so that he may start speaking. They are often repeated by the speaker to gain attention or are used to interrupt the actual talking speaker. Usually, floor grabbers occur at the beginning of a speaker's turn. They are often louder than the neighbouring speech. They share a common vocabulary with floor holders, backchannels and accepts. Other features, such as the context or the corresponding audio record are then required to label them. Common floor grabbers are for example: "Uh", "So", etc.

Floor Holders occur mid-speech by a speaker who has the floor. It is usually a short clause like "uh" or "so" and it is used to fill the pause in the utterance, when the speaker think about the next words of the utterance. Sometimes, a floor holder is used at the end of the speaker turn in order to leave the floor. Generally, their energy is similar to the neighbouring speech but their duration is usually longer than the other words. Floor holders are not common at the beginning of a speaker turn, but more likely occur in the middle or at the end of the turn, often in the middle of an utterance. There are often confused with floor grabbers, backchannels, ..., because of the similar vocabulary. One example of floor holders in the middle of an utterance is: "Je pense **euh** c'est vrai." (*I think* **uh** *it is right.*).

2.3.7 Conventional-opening and Conventional-closing

Conventional-opening DA class contains utterances, which function is to inform about the beginning of a dialogue, such as: "Dobré odpoledne!" (*Good afternoon!*) or "Madame la conseillère, bonjour!" (*Good morning, madame the adviser!*).

Conventional-closing DA class contains utterances, which function is to inform about the end of a dialogue. Examples are: "Bonne journée!" (*Have a nice day!*) or "Na shledanou!" (*Good bye!*).

2.3.8 Politeness Mechanisms

This group contains dialogue acts that mark utterances where speaker express courteousness. It is composed of several DAs, but only one occurs in our corpus, that is the **Thanks** dialogue act, which is composed of the utterances where a speaker thanks another speaker. Examples of thanks are: "Merci beaucoup!" (*Thank you very much!*) or "Děkuji mnohokrát!" (*Many thanks!*).

2.3.9   Disruption Forms

This broad class contains utterances, which are indecipherable, abandoned or interrupted. It is possible to use one disruption form per utterance only.

Indecipherable marks indecipherable speech such as mumbled or muffled words or utterances that are too difficult to hear, for instance because of breathing noise in the microphone. There is no vocabulary for this DA, which can be determined only from the audio stream and the context.

Interruptions correspond to utterances in which the speaker is interrupted by another speaker and stops talking. Examples of interruptions are: "si on ... oui" (*if we ... yes*) or "půjdeme do ... ale ne" (*We will go to ... but no*)[1].

Abandoned labels utterances that are abandoned by the speaker. They usually occur when a speaker decides to reformulate the utterance. The current utterance is abandoned and the new one is beginning. For example, abandoned utterances can be: "Když se podívate na ..." (*If you look at ...*) or "Savez-vous que ..." (*Do you know that ...*).

The most important DA classes described previously are summarized in Table 2.1. For other DAs, please refer to [3, 57, 36].

2.3.10   Reduction of the DA Tag-set

Complete DA tag-sets with tens of DAs are usually to large for DA recognition. Several DA classes only have very few occurrences, which makes it difficult to model them. Furthermore, several other DAs are not useful for the application. Therefore, the complete DA tag-set is usually reduced for recognition into a few broad classes. Reduction is realized by removing the DA classes that are not needed by the application, and by grouping together DA classes that do not occur enough times.

Table 2.2 shows the 10 most frequent DAs from the SWBD-DAMSL corpus with examples and their relative frequencies. This table can be used to group dialogue acts for DA recognition.

An example of the DA tag-set that is often used for DA recognition is shown in Table 2.3. This tag-set is based on SWBD-DAMSL and contains seven grouped DA classes.

2.4   Sentence Modality

Several works deals with automatic sentence modality (or sentence mode) recognition [114, 50, 38], which can be considered as a subset of automatic DA recognition.

A sentence gives a specific relationship between the speaker and the other participants of a discussion, which is called modality. These relationships can be clustered into different classes. The simplest classification distinguishes declaration from interrogation. A more

---

[1] "..." marks the instant of interruption of speaker A by speaker B, who starts talking.

DA class	Example
1. Statements	
Statement opinion	In my opinion, this is a good decision.
Statement non-opinion	It is a great story.
2. Questions	
Yes/No question	Do you think that it is ok?
Wh-question	What do you mean?
3. Action motivators	
Command	Continue!
Suggestion	Maybe you have to standardize this thing also.
Commitment	I will continue to sleep.
4. Backchannels and Acknowledgments	
Backchannel	Uh-huh
Acknowledgment	Oh okay
5. Responses	
Accept	Yeah
Reject	Not at all
Maybe	Maybe
6. Floor mechanisms	
Floor grabber	Um
Floor holder	Well
7. Conventional-opening and Conventional-closing	
Conventional-opening	Hello!
Conventional-closing	Bye!
8. Politeness Mechanisms	
Thanks	Thank you!
9. Disruption forms	
Indecipherable	
Interruption	I would ... Your opinion is not good.
Abandoned	I'm sor ...

Table 2.1: Summary of the most important DA classes for our work, along with examples.

DA Type	DA Tag	Example	Amount in [%]
Statement non-opinion	sd	I live in New York.	36
Acknowledge (Backchannel)	b	Uh-huh	19
Statement opinion	sv	I hope, she is pretty.	13
Agree/Accept	aa	It is really sure.	5
Abandoned or Turn-Exit	%-	So ...	5
Appreciation	ba	I can understand.	2
Yes/No question	qy	It's true?	2
Non-verbal	x	<Laughter>, <Throat_clearing>	2
Yes answers	ny	Yes	1
Conventional-closing	fc	Well, it's been nice talking to you.	1

Table 2.2: The 10 most frequent dialogue acts from the SWBD-DAMSL tag-set.

DA Type	DA Tag
Statements	
Statement non-opinion	sd
Statement opinion	sv
Questions	
Yes/No question	qy
Wh-question	qw
Other question	all other
Backchannels	b
Incomplete utterances	%
Accepts	aa
Appreciation	ba
Other	all other

Table 2.3: Grouped SWBD-DAMSL DA tag-set with seven broad DA classes.

complex classification (for example, declaration, interrogation and order) includes subtle degrees of social relationship, of discussion and context, etc. In almost all languages, various information (syntactic, morphologic, as well as the intonation) indicate sentence modality.

The expected information given by prosody is for example [63]:

- A falling intonation for a statement.
- A rising F0 contour for a question.
- Continuation-rise characterizes a (prosodic) clause boundaries, which differ from the end of sentences.

Sentence modal classes correspond to a small DA subset. A possible correspondence between modal and DA classes is shown in Table 2.4.

Modal class	Corresponding DA class from MRDA tag-set
declarative sentence	statement
investigation question	yes/no question
imperative sentence	command

Table 2.4: Correspondence between modal classes and DA classes.

2.5 Dialogue Act Recognition Information

The most important information that is used to recognize dialogue acts is described in this section.

The first one is **lexical information**. Every utterance is composed of a sequence of words. Generally, the DA of an utterance can be partly deduced from the lists of words that form

this utterance. For example, Wh-questions often contains one interrogative word, which does not occur in other DA classes.

The second one is **syntactic information**. It is related to the *order* of the words in the utterance. For instance, in French and Czech, the relative order of the *subject* and *verb* occurrences might be used to discriminate between declarations and questions.

Another information is **semantic information**. The sense of the utterance is also correlated to the DA. However, this information is very difficult to obtain automatically, which is why, to the best of our knowledge, it is not used so far in the DA recognition domain.

Yet another information that is useful to recognize DAs is **prosody**, and in particularly the melody of the utterance. Usually, questions have an increasing melody at the end of utterance, while statements are usually characterised by a slightly decreasing melody.

The last information mentioned here is the **context** of each DA. Hence, any DA depends on the previous (and next) DAs, the most important context being the previous one. For example, a "Yes" or "No" answer is most likely to just follow a *Yes/no question*. The sequence of pronounced DAs is also called the *dialogue history*.

Most of the works on DA recognition makes use of a combination of the three following information sources [113, 110]:

1. Lexical (and syntactic) information
2. Prosodic information
3. Dialogue history

2.5.1 Lexical Information

Lexical and syntactic features can be derived from the word sequence in the dialogue. The first broad group of DA recognition approaches that uses this type of features is based on the assumption that different dialogue acts are generally composed of sequences of different words.

The correspondence between DAs and words sequences is usually represented either by Bayesian models, such as n-grams, Naive Bayes, Hidden Markov Models, Bayesian Networks, etc., or Non-Bayesian approaches, such as Neural Networks, Semantic Classification and Regression Trees, etc.

2.5.2 Prosodic Information

Most researchers agree on the fact that the lexical/syntactic information can not totally explain DAs alone. Prosodic cues [74], which are somehow independent of the words used in the sentence, are also correlated to the actual DA.

For example, questions are usually characterized by an increasing melody at the end of the utterance [88], and accepts have usually much more energy than backchannels and

acknowledgments (c.f. Section 2.3.5). Prosodic features are usually modeled with the same statistical methods as used for lexical information.

Definitions

Many different definitions of prosody exist. We quote here the three following ones, which serve as a basis for this work:

1 - The term *prosody* comprises speech attributes which are not bound to phone segments [63].

2 - The *prosody* is a set of suprasegmental phenomena of the speech. Therefore, the prosodic units are not related to acoustic units - phonemes, words, syllables, sentences, etc [74].

3 - Another definition of the *prosody* is a study of the three following phonetic correlates [74]:

- Fundamental frequency (F0)

- Energy

- Duration

Importance

While lexical information is a strong cue to recognize DAs, prosody also clearly plays an important role. The same sentence may even have a different meaning depending on the prosody. For example, the simple sentence: "Il pleut." in French, "Prší." in Czech, "It is raining." in English can be a statement or a question in Czech or in French.

DA labeling of corpora is usually realized based on transcripts only, because of practical reasons (mainly speed efficiency). However, the authors of [58] show that prosody should be taken into account during the DA labeling process. In particular, they have realized the following experiment: 44 randomly selected conversations are first labeled from text transcripts only. Then, the same sentences are relabeled in the same conditions as before, with the transcript and context, but also with listening to the sentences. Both labels are compared, and a difference of about 2% is observed in the DA labels. This difference may seem small enough to justify the use of transcripts only during the labeling process. However, it was also noted that the same differences often occur for the same DA classes, which increases the labeling error rate for these particular classes up to a significant level, as shown in Table 2.5.

General Properties

Prosody is generally not considered useful to recognize all types of DAs, but seems to play an important role for specific DA types, such as questions, accepts and incomplete

DA from transcript only	DA from listening and transcript	Count	%
backchannels	accepts	43/114	38
opinion statements	non-opinion statements	22/114	19
non-opinion statements	opinion statements	17/114	15
other		32	(3 each)

Table 2.5: Modifications of some DA labels when labeling with transcripts only and with both audio and transcript.

utterances. The following describes some general rules where prosody is involved:

- The declarative and yes/no questions are usually characterized by a rising F0 at the end of utterance.

- Several incomplete utterances have a final F0 contour similar to that of the middle of a normal utterance, which is often neither rising nor falling.

- The energy at the end of incomplete utterances is also usually higher than for complete utterances.

- Backchannels differ from accepts by the amount of effort used when speaking.

- Usually, accepts have a higher energy, a greater F0 movement, and a higher likelihood of accents and boundary tones than backchannels.

For more information about general prosodic properties, please refer to [110].

French Prosodic Rules

The basic prosodic rules concerning the French language can be summarized as in [42]:

- Statement: small decrease of melody.

- Order (or command): important decrease of melody.

- Questions (particularly yes/no): increase of melody.

- Grammar question (wh-question, ...): neutral intonation.

Let us consider for example the French sentence *Il dort* (in English "He sleeps.") pronounced in some neutral context. This sentence can belong to at least two dialogue acts:

- Statement: *Il dort.*

- Question: *Il dort ?*

Apart from prosody, all other information types (syntactic, morphological, contextual) have the same characteristics in both cases. This means that sometimes only prosody can be used to distinguish a statement from a yes/no question. Indeed, in the above example, one can observe a significant difference between the melodic contours of each DA, as illustrated in Figure 2.2. In this figure, the F0 contour is increasing for yes/no questions and slightly decreasing for statements.
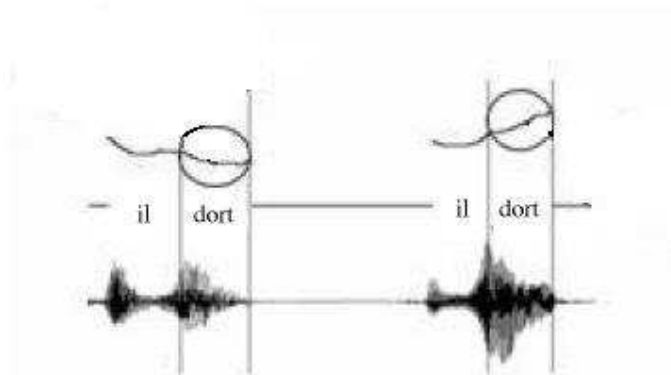
Figure 2.2: Fundamental frequency contour for a statement (left) and a yes/no question (right).

In other cases, the importance of the intonation contour decreases in favor of other phonological information. For yes/no questions, the inversion of the couple subject-verb or the interrogative form *est-ce que* are examples of such information. The sentence thus becomes: *Dort-il ?* in the first case and *Est-ce qu'il dort ?* in the second one. The F0 curve is shown in Figure 2.3. One can note that the characteristic interrogative contour is neutralized in this particular interrogative form.

We have illustrated in this section the theoretical influence of prosody on DA classification in French through a few examples only. For more information about French prosodic rules, please refer to [85, 84, 88, 125].

Czech Prosodic Rules

There are three basic melodic types of Czech utterances. The most common type is the *declarative melody* (for statements). Another melodic type is the *melody of yes/no questions* and the last one is the *melody before the clause-boundary pause*.

Declarative Melody

The basic melody of statements is decreasing. It is characterised by a significant melodic decrease after the accented syllable of the utterance core. This part of the sentence is
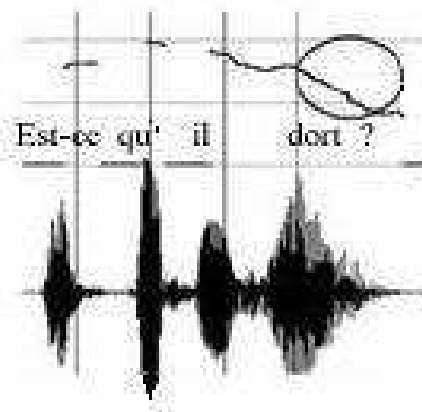
Figure 2.3: Fundamental frequency contour for yes/no question with the *est-ce que* form.

called "melodem". If the last word is mono-syllabic, the fall is only on this syllable. In the case of a poly-syllabic word (or a group of words), the fall is distributed over each syllable (or word). Two examples of declarative F0 contour are shown in Figure 2.4: *Měl s sebou psa.* (in English "He was with a dog.", left) and *Měl s sebou kočku.* (in English "He was with a cat.", right).
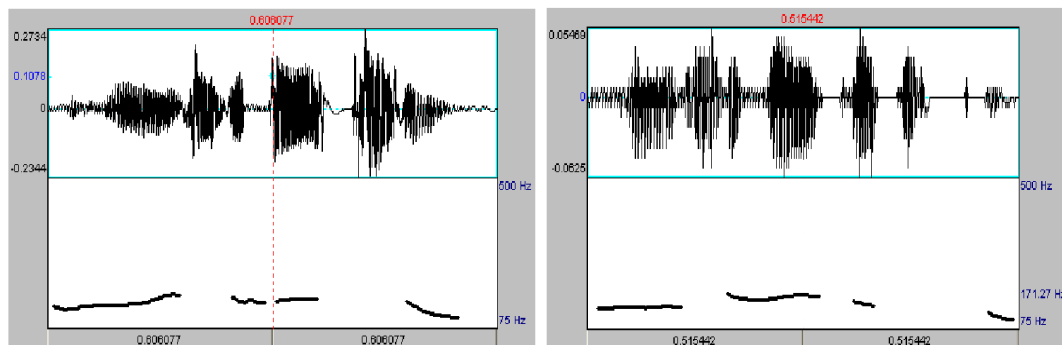
Figure 2.4: F0 contour of two statements: *Měl s sebou psa.* (in English "He was with a dog.") with one syllable in melodem (left) and *Měl s sebou kočku.* (in English "He was with a cat.") with bi-syllable in melodem (right).

A similar type of melody can be observed in orders or in questions, which are completed by an interrogative word. Figure 2.5 shows an example of an order: *Vezmi s sebou kočičku!* (in English "Take a kitten!", left) and an example of a wh-question (with an interrogative word): *Co se ti přihodilo?* (in English "What happened to you?", right).

Investigation Question Melody

The investigation question (or yes/no question) is a question without an interrogative word and that has an answer: "*yes*" or "*no*". The melody at the last accented syllable starts
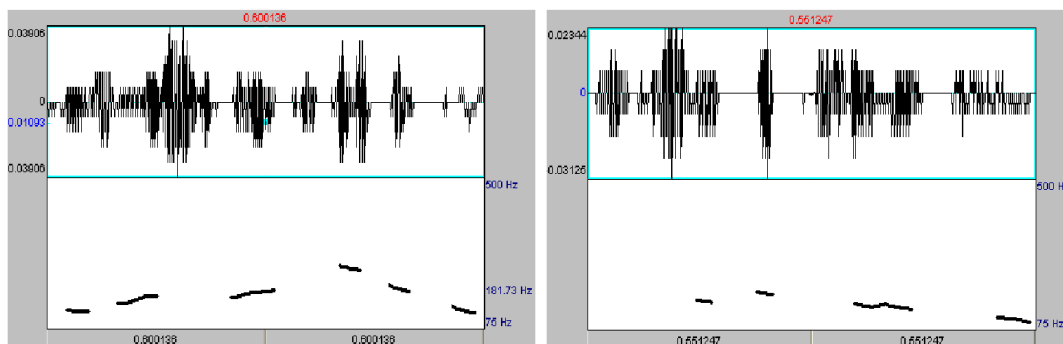
Figure 2.5: F0 contour of an order: *Vezmi s sebou kočičku!* (in English "Take a kitten!") with two syllables in melodem (left) and a wh-question: *Co se ti přihodilo?* (in English "What happened to you?") with four syllables in melodem (right).

by a low tone. If the melodem is mono-syllabic only, then the melody is increasing with a round shape. If it is bi-syllabic, the accented syllable has a lower melody than the last syllable. Figure 2.6 shows an example of F0 contours for two yes/no questions: *Prijdeš včas?* (in English "Will you come in time?", left) and *Už jsi skončil?* (in English "Have you finished?", right).
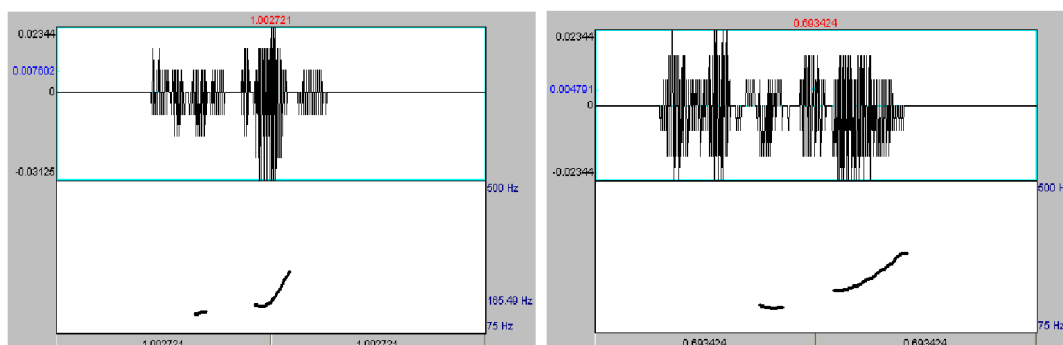
Figure 2.6: Fundamental frequency (melody) contours for two yes/no questions: *Prijdeš včas?* (in English "Will you come in time?", left) and *Už jsi skončil?* (in English "Have you finished?", right).

When the last accented word is composed of more that two syllables, two cases may occur:

- The accented syllable and the following one (not accented) have a low frequency (the second syllable can have a little higher F0). The melody is increasing only at the end of the utterance.
- The accented syllable is heavy but the next one is much higher and the other syllables at the end have a decreasing melody.

These two cases are shown in Figure 2.7: *Znáte sousedy?* (in English "Do you know your neighbours?").
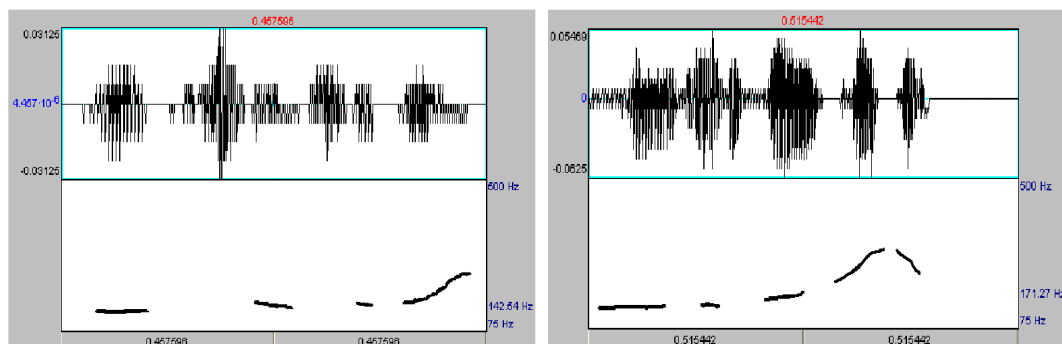
Figure 2.7: Fundamental frequency (melody) contours for two cases of yes/no questions: *Znáte sousedy?* (in English "Do you know your neighbours?"): case 1 on the left and case 2 on the right.

Melody before the Clause-Boundary Pause

The clause-boundary pause is a pause, which usually separates two simple sentences in a complex sentence, two clause elements in a simple sentence or two enumerated elements. There are several melodic forms. The most frequent and basic one is the increasing melody. It is characterised usually by a low tone before an accented syllable and it is gradually increasing towards the pause. This is not in the scope of this study (for more information, see [95]).

For more information on Czech sentence melody, please refer to [103, 95, 33].

Prosodic features

The most important prosodic features are:

- Fundamental frequency (F0)
- Energy
- Duration
- Speaking rate
- Voice quality
- Pause

Compound prosodic features are formed by the variations of these attributes over time:

- Intonation
- Accentuation
- Rhythm
- etc.

The basic prosodic features, along with their importance in DA recognition, are described in the next sections. The compound prosodic features are described in [63].

Fundamental Frequency

The Fundamental Frequency (F0) or pitch, often simply referred to as the *fundamental*, is the lowest frequency in the harmonic series of vibration of vocal folds.

The F0 of a periodic signal is the inverse of the pitch period length. The pitch period is the smallest repeating unit of a speech signal. One pitch period thus describes the periodic signal completely. The significance of defining the pitch period as the smallest repeating unit can be appreciated by noting that two or more concatenated pitch periods form a repeating pattern in the speech signal (multiples of F0). However, the concatenated signal unit obviously contains redundant information.

The fundamental frequency is a characteristic of voiced sounds only. Its value depends on the speaker, mainly in function of his age and his sex (100 - 160 Hz for a man and 150 - 300 Hz for a woman). The F0 does not allow to discriminate two vowels ("a", "e", etc.), but this can be done with formants [47]. The F0 curve is called the melody.

F0 automatic extraction is a complex problem [49]. Two families of methods are usually employed: in the first one, the computation is performed in the *time domain*, while the second one operates in the *spectral domain*.

Time Domain Methods [97] are based on the assumption that the signal of the speech can not vary quickly in a limited time frame. This is a consequence of the physical properties of the vocal tract. The principle of these methods (named short term analysis methods) consists in processing the speech signal within short time intervals. The length of these frames is generally between 10 and 20 ms. The result of the analysis of each frame is represented by one vector. The analysis of a complete sentence is represented by a vector sequence. Examples of this method are: autocorrelation functions, AMDF, etc.

Spectral Methods [86, 87] are based on the fact that, most often, harmonics of F0 exist. Therefore, the information contained in the whole speech spectrum can be exploited to extract F0. The main advantage of these methods is that they might work well even when the F0 is noisy or filtered out. Several harmonic frequencies and amplitudes measures can be used. This principle is used for example in the method based on cepstrum [93] or on the period histogram [108]. Usually, the time domain digital signal is transformed into the power spectrum domain via the Discrete Fourier Transform (DFT) [47].

These methods are more robust than the first ones, but they also increase the computational cost. This explain why time-domain methods have often been used in the past. For more information about F0 extraction algorithms, please refer to [99, 47].

After extraction of the F0 curve, several F0 features are computed for DA recognition. Examples of these features are the max, min, mean, standard deviation, etc., as shown in Table 2.6.

F0 feature	Description
F0_mean	mean of F0 values in the utterance
F0_mean_end	mean of F0 values in end segment of the utterance
F0_mean_ratio	ratio of F0 mean in the utterance divided by F0 mean in the conversation side
F0_std_dev	standard deviation of F0 in the utterance
F0_max	maximal value of F0 in the utterance
F0_min	minimal value of F0 in the utterance
F0_grad	gradient of F0
F0_perc_good_utt	ratio of number of good F0 values divided by the number of F0 values in the utterance

Table 2.6: Example of F0 features.

Energy

The energy represents the loudness of speech; the relation between them depends on the sensitivity of the human auditory system to different frequencies. The energy is often called the "force" of the speech. Its value is in relation to the quantity of air in the vocal tract. An increase of energy is often related to an increase of F0.

The most common way to calculate the energy is the *Root Mean Square Energy (RMSE)*, which is the square root of the average of the sum of the squares of the amplitude of the signal samples. Using a window of width W to segment the speech into frames, let $s_n(i)$ denotes the ith windowed speech sample in frame number n, and let E_n be the energy of frame n, the $RMSE$ of E_n is given by the equation:

$$E_n = [\frac{1}{W} \sum_{i=1}^{W} s_n^2(i)]^{\frac{1}{2}} \tag{2.1}$$

Several energy features are computed for DA recognition as for example: max, min, mean, etc.

For more information about energy computation algorithms, please refer to [97, 46]. Other energy features are shown in [110].

Duration (and Speaking Rate)

Duration is the timing interval of pronunciation of each acoustic unit. It can be measured for single phones, for syllable segments or for other acoustic units and its unit can be a millisecond. Each phone has its own mean and standard deviation for duration. The speaking rate (enrate) is the inverse value of duration. It is still not clear how duration and speaking rate are perceived [94], but an objective measurement of speaking rate can be obtained by normalizing the durations with phone intrinsic values [124].

Table 2.7 shows several duration and speaking rate features used for dialogue act recognition.

Feature	Description
ling_dur	duration of the utterance
ling_dur_minus	ling_dur minus sum of the duration of all pauses longer at 100 ms
cont_speech_frames	# of frames in continuous speech region
mean_enr	mean of speaking rate values in the utterance
min_enr	minimum of speaking rate values in the utterance
max_enr	maximum of speaking rate values in the utterance

Table 2.7: Duration and speaking rate (enrate) features.

Voice Quality

Voice quality is a way of describing and evaluating speech fidelity, intelligibility, and the characteristics of the analog voice signal itself. It involves attributes that concern the overall phone independent spectral structure, for example, jitter (small and apparently random perturbations of the F0 period), shimmer (small and apparently random perturbations of the F0 amplitude), or the relative energy of the higher harmonics with respect to F0.

Pause

Pause is a timing interval between two phones or between two other acoustic units. We can distinguish two pause types: *unfilled* and *filled*. An unfilled pause is simply a silence or it may contain breathing or background noises. Conversely, a filled pause is a relatively long speech segment of rather uniform spectral characteristics consisting of a short "eh" or sometimes followed by an "ehm" [63]. This type of pauses can be called *hesitation pause* or *hesitation*. The F0 is flat or slightly falling and it is at a comparably low level.

2.5.3 Dialogue History

The third general type of information used in classical DA recognition systems is the dialogue history. It is defined by the sequence of previous DAs that have been recognized. It may be used to predict the next DA. Different formalisms are employed to model this information: statistical models such as n-grams, Hidden Markov Models (HMMs), Bayesian Networks, etc.

2.6 Segmentation

Before DA recognition, the dialogue must be segmented into sentence-level units, or utterances [90], where each utterance represents a single DA. In this work (except in chapter 5),

we assume that the corpus has already been segmented. The issue of utterance segmentation is thus only dealt with in Chapter 5.

2.7 Bayesian Approaches

The main types of automatic DA recognition approaches proposed in the literature can be broadly classified into Bayesian and Non-Bayesian approaches. Bayesian approaches are presented in this section and Non-Bayesian approaches are described in Section 2.8.

2.7.1 Notations

The main mathematical symbols that are used throughout this thesis are reported and defined in Table 2.8.

Symbol	Definition	Description
$\mathcal{C}$	set of all DA classes c (DA tag-set)	$c \in \mathcal{C}$
C	sequence of DAs	$C = (C_1, .., C_\tau, .., C_{\mathcal{T}})$
O	observation (generic)	
W	lexical (and syntactic) information (sequence of words w_i)	
A	acoustic information	
F	prosodic information	

Table 2.8: Names and definition of symbols used in the manuscript: C, O, W, A and F are random variables.

When there is no ambiguity possible, the subscript τ for a single dialogue act C_τ might be dropped to simplify notations.

2.7.2 Principle

The most common formalism used in the DA recognition domain is the Bayesian framework. For instance in [10], the best sequence of dialogue acts $\hat{C}$ that maximizes the *a posteriori* probability $P(C|O)$ over all possible sequences of dialogue acts C on the observation O is obtained from the observation likelihood $P(O|C)$ as follows:

$$\hat{C} = \arg\max_C P(C|O) = \arg\max_C \frac{P(C).P(O|C)}{P(O)} = \arg\max_C P(C).P(O|C) \qquad (2.2)$$

This is the widely known derivation that is classically used in many pattern recognition tasks and that solves the maximum *a posteriori* criterion by training generative models of the observations.

2.7.3   Lexical (and Syntactic) N-Gram DA Models

The most common methods model $P(O|C) = P(W|C)$, where W is the word sequence in the pronounced utterance with statistic models such as n-grams. These methods are based on the observation that different DA classes are composed of distinctive word strings. For example, 92.4% of the "uh-huh" occur in Backchannels and 88.4% of the trigrams "<start> do you" occur in yes-no questions [113]. The words order and positions in the utterance may also be considered. A theory of word frequencies, which is the basis for DA modeling from word features, is described in [41].

DA Recognition from Exact Word Transcription

The following approach is based on the hypothesis that the words in the utterances are known. Then, Equation 2.2 becomes:

$$\arg\max_C P(C|W) = \arg\max_C P(C).P(W|C) \tag{2.3}$$

The "Naive Bayes assumption", which assumes independence between successive words, can be applied and leads to:

$$\arg\max_C \frac{P(C).P(W|C)}{P(W)} = \arg\max_C P(C).\prod_{i=1}^{T} P(w_i|C) \tag{2.4}$$

This equation represents the unigram model, also sometimes called the Naive Bayes classifier. In this case, only lexical information is used. More complex models, such as 2-grams, 3-grams, etc., further consider syntactic information about the dependencies between adjacent words. These n-grams usually model local structures only. The complexity and performances of each of these models depends on the size of the corpus. Usually, 4-grams or more complex models are not used.

Reithinger et al. use in [100] unigram and bigram Language Models (LMs) for DA recognition on the VERBMOBIL corpus. Their DA recognition rate is about 66% for German and 74% for English with 18 dialogue acts. In [77], a naive Bayes n-gram classifier is applied to the English and German language. The authors obtain a DA recognition rate of 51% for English and 46% for German on the NESPOLE corpus. Grau et al. use in [44] the naive Bayes and the uniform naive Bayes classifiers with 3-grams. Different smoothing methods (Laplace and Witten Bell) are evaluated. The obtained recognition rate is 66% on the SWBD-DAMSL corpus and with 42 DAs. Ivanovic also uses in [51] the naive Bayes n-grams classifier and obtains about 80% of recognition rate in the instant messaging chat sessions domain with 12 DAs classes derived from the 42 DAs of DAMSL.

One can further assume that all DA classes are equi-probable, and thus leave the $P(C)$ term out:

$$\hat{C} = \arg\max_C P(W|C) \tag{2.5}$$

This approach is referred to as the *uniform* naive Bayes classifier in [44].

DA Recognition from Automatic Word Transcription

In many real applications, the exact words transcription is not known. It can be approximately computed from the outputs of an automatic speech recognizer. Let A be a random variable that represents the acoustic information of the speech stream (e.g. spectral features).

The word sequence W is now an hidden variable, and the observation likelihood $P(A|C)$ can be computed as:

$$P(A|C) \quad = \quad \sum_W P(A|W,C).P(W|C) \qquad (2.6)$$

$$= \quad \sum_W P(A|W).P(W|C) \qquad (2.7)$$

where C is the DA class and $P(A|W)$ is the observation likelihood computed by the speech recognizer for a given hypothesized word sequence W. The summation over all W hypotheses is approximated over k best only. A dialogue act recognition approach from recognized words is shown for example in [113].

2.7.4   Dialogue Sequence N-Gram Models

The dialogue history also contains very important information to predict the current DA based on the previous ones. The dialogue history is usually modeled by a statistical discourse grammar, which represents the prior probability $P(C)$ of a DA sequence C.

Let C_τ be a random variable that represents the current dialogue act class at time τ. The dialogue history H is defined as the previous sequence of DAs: $H = (C_1, .., C_{\tau-1})$. It is usually reduced to the most recent n DAs: $H = (C_{\tau-n+1}, .., C_{\tau-1})$. The most common values for n are 2 and 3, leading to 2-gram and 3-gram models. In order to train such statistical models, the conditional probabilities $P(C_\tau|C_{\tau-n+1}, .., C_{\tau-1})$ are computed on a labeled training corpus. *Smoothing* techniques, such as standard back-off methods [11], may also be used to train high-order n-grams. N-grams are successfully used to model dialogue history in [113, 101].

Polygrams are mixtures of n-grams of varying order: n can be chosen arbitrarily large and the probabilities of higher order n-grams are interpolated by lower order ones. They usually give better recognition accuracy than standard n-grams and are shown in [89].

2.7.5   Hidden Markov Models

Hidden Markov Models [98] can be used to model sequences of dialogue acts. Let O be a random variable that represents the observations and C the sequence of DAs classes. n th-order HMM can be considered, which means that each dialogue act depends on the n previous DAs (in a similar way as for n-grams). Then, each HMM state models one DA and the observations correspond to utterance level features. The transition probabilities

are trained on a DA-labeled training corpus.

DA recognition is carried out using some dynamic programming algorithm such as the Viterbi algorithm, which estimates the most probable DA sequence $\hat{C} = \arg\max_C P(C|O)$.

HMMs with word-based and prosodic features are successfully used to model dialogue history in [112]. Wright uses in [127] intonation events and tilt features such as: F0 (fall/rise, etc.), energy, duration, etc. She achieves 64% of accuracy on the DCIEM map task corpus [8] with 12 DA classes. Ries combines in [102] HMMs with neural networks (c.f. Section 2.8.1). He obtains about 76% of accuracy on the CallHome Spanish corpus. In [37] language models and modified HMMs are applied on the Switchboard corpus [43] with the SWBD-DAMSL tag-set.

2.7.6 Bayesian Networks

A Bayesian network is represented by a directed acyclic graph. The nodes represent random variables and the arcs represent relations (dependencies) between nodes. The topology of the graph models conditional independencies between the random variables. In the following, we do not differentiate dynamic Bayesian networks (with stochastic variables) from static Bayesian networks, as most of our variables are stochastic, and when static Bayesian networks are drawn, they represent an excerpt of a dynamic Bayesian network at a given time slice. The stochastic variables are conditionally dependent of theirs descendants and independent of theirs ascendants.
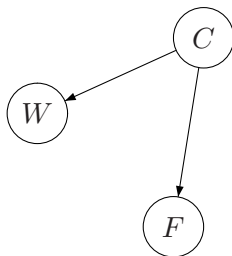
Figure 2.8: Example of Bayesian network for dialogue act recognition.

An example of Bayesian network for dialogue act recognition is shown in Figure 2.8. Node C represents the current dialogue act. Utterance features are represented by nodes W (sequence of words in the utterance) and F (prosodic features). The dialogue context is not considered there. The conditional independence assertions of this network allows the following factorization:

$$P(C, W, F) = P(W|C).P(F|C).P(C) \tag{2.8}$$

In order to build such a network, the network structure (conditional dependencies) and

the conditional probability distributions must be defined. The conditional probabilities are trained statistically on a training corpus. The topology of network can be created manually or automatically.

Bayesian networks are successfully used in [59] for dialogue act recognition. In the first experiment reported, three features are used: sentence type (declarative, yes/no question, etc.), subject type (1st/2nd/3rd person) and punctuation (question mark, exclamation mark, comma, etc). The Bayesian network is defined manually. They achieve 44% of accuracy on the SCHISMA corpus. In the second experiment, a small corpus is derived from the dialogue system used to interact with the navigation agent. Utterances are described by surface level features, mainly keywords-based features. These features are computed automatically for each utterance. Bayesian networks are further generated automatically iteratively, starting from a small hand-labeled DA corpus. This network is used to parse another large corpus, and a new network is generated from this corpus. This approach gives 77% of accuracy for classification of forward-looking functions (7 classes) and 88% of accuracy for backward-looking functions (3 classes).

Another application of Bayesian network in dialogue act recognition is shown in [53]. Two types of features are used: utterance features (words in the utterance; w_i) and context features (previous dialogue act; $C_{\tau-1}$). The authors compare two different Bayesian networks to recognize DAs (c.f. Figure 2.9).

(a) (b)
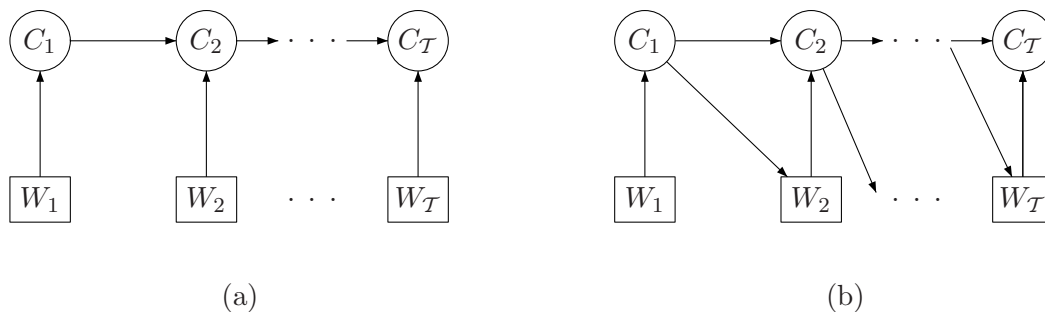
Figure 2.9: Two Bayesian networks for dialogue act recognition: C_i represents a single DA, while W_i is a sequence of words.

These networks are built manually. In the left model of Figure 2.9, each dialogue act is recognized from the words of the current utterance and from the previous DA. In the right model of Figure 2.9, the authors further consider an additional dependency between each word of the utterance and its previous dialogue act (diagonal arcs). They achieve about 64% precision on a subset of the MRDA corpus and with the reduced DA set size.

2.8 Non-Bayesian Approaches

Non-Bayesian approaches are also successfully used in the DA recognition domain, but they are not so popular as Bayesian approaches. Examples of such approaches are Neural Networks (NNs), such as Multi-Layer Perceptron (MLP) or Kohonen Networks, Decision Trees, Memory-Based Learning and Transformation-Based Learning.

2.8.1 Neural Networks

A neural network (NN) [48] is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. It can be used to model complex relationships between inputs and outputs or to find patterns in data.

Multi-Layer Perceptron

One of the most frequently used neural network technique in the DA recognition domain is the Multi-layer perceptron (MLP, c.f. Figure 2.10), which consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and one output layer. The input signal propagates through the network layer-by-layer. An MLP can represent a non linear function.
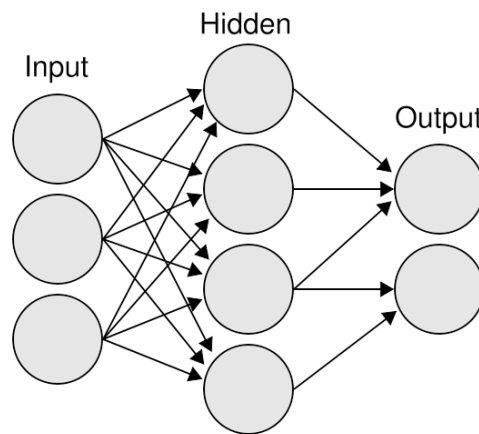
Figure 2.10: Example of multi-layer perceptron.

Wright [127] describes an approach with a one-hidden-layer MLP. 54 suprasegmental and duration prosodic features are used as inputs. She achieves 62% of accuracy on the DCIEM map task corpus [8] with 12 DA classes. Ries successfully uses in [102] an MLP both standalone, and in combination with HMMs. He obtains a similar accuracy (about 76%) on the CallHome Spanish corpus with both setups. Sanchis et al. also use in [107] an MLP to recognize DAs. The inputs of MLP are the words of the lexicon of the restricted-semantic task (138 inputs=size of the lexicon). The experiments are performed on the Spanish dialogue corpus in the train transport domain (16 DA classes). They achieve about 93%

of accuracy on the text data and about 72% of accuracy on the recognized speech. Note that this approach will be difficult to apply on a real (large) lexicon. Levin et al. use in [77] a set of binary features to train an MLP. These features are computed automatically by combining a grammar-based phrasal parsing and machine learning techniques. They obtain a DA recognition accuracy of about 71% for English and about 69% for German on the NESPOLE corpus.

Kohonen Networks

Another type of neural network used in the dialogue act classification domain is Kohonen Networks. A Kohonen network [62], also known as Self-Organizing Map (SOM), defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. A model is associated with each grid node (c.f. Figure 2.11).

The topology of the SOMs is a single layer feedforward network where the discrete outputs are arranged into a low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. A weight vector with the same dimensionality as the input vectors is attached to every neuron. The number of input dimensions is usually much larger than the output grid dimension. SOMs are mainly used for dimensionality reduction rather than expansion.

The models of the Kohonen network are estimated by the SOM algorithm [29]. A data item is mapped onto the node which model is the most similar to the data item, i.e. has the smallest distance to the data item, based on some metric.

(a)   (b)
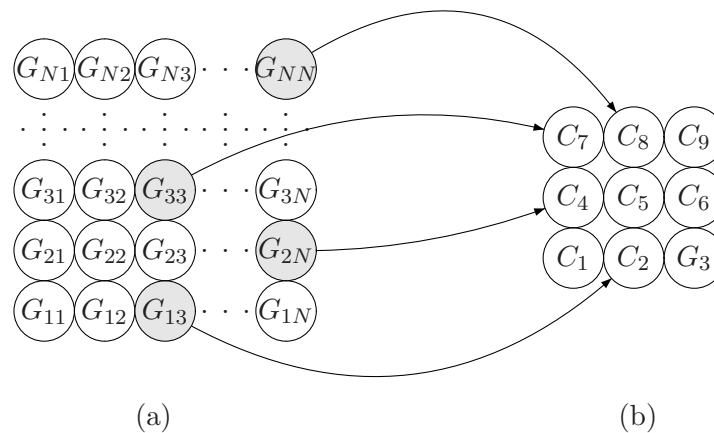
Figure 2.11: Two Kohonen networks (from [5]) with a rectangular structure to model dialogue acts: The inputs to the large network (on the left) are a set of binary utterance features. Neurons representative of DA classes are grayed. The small network on the right represents the outputs of system (DA classes). The connexions between the neighboring nodes are not shown.

Kohonen networks for dialogue act recognition are used in [5]. The authors use seven *superficial* utterance features: speaker, sentence mode, presence or absence of a wh-word,

presence or absence a question mark, etc. Each utterance is represented by a pattern of these features, which is encoded into a binary format for the SOM representation. Initially, the exact number of DA classes is not known *a priori*, and only the large network on the left is created and trained. The clustering process is interrupted after a given number of clusters have been found.

To interpret the clusters, another small Kohonen network is built (right model in Figure 2.11). This network contains as many neurons as DA classes. These neurons are initialized by the values of the weight-vectors of the representative neurons from the large network.

The quality of classification is evaluated by the Specificity Index (SI) [4] and by the Mean number of Conditions (MoC). They achieve about 0.1 for SI and about 2.6 for MoC on the SCHISMA corpus, with 15 DA classes and a network with 10*10 neurons. Another experiment has been performed with 16 DA classes and a larger network with 12*12 neurons with comparable results. Generally, unsupervised methods such as Kohonen networks are rarely used for DA recognition.

2.8.2 Decision Trees

Decision trees (or Classification and Regression Trees, CARTs) [16] are generation tools that are successfully used in operations research and decision analysis. They are usually represented by an oriented acyclic graph (c.f. Figure 2.12). The root of the tree represents the starting point of the decision, each node contains a set of conditions to evaluate, and arcs show the possible outcomes of these decisions.
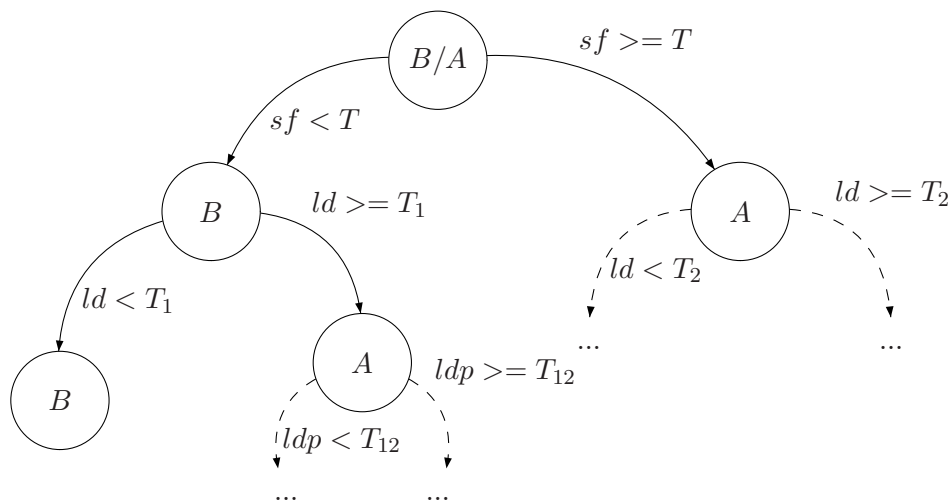
Figure 2.12: Example of a part of the decision tree in the DA recognition domain: recognition of Backchannels (B) and Accepts (A) by prosody, from [110].

In the case of DA recognition, the decisions usually concern utterance features. Each decision compares the value of some feature with a threshold. For example, in Figure 2.12,

three different prosodic features (sf, ld and ldp) are shown with their corresponding thresholds (T, T_1, T_2 and T_{12}). sf is the pause type feature and ld and ldp are the duration type features. Training of the decision tree is performed automatically on the training corpus. The output of the CART is the probability of the DA given the utterance features (lexical and prosodic), i.e. the *posterior* probability $P(C|W, F)$. Usually, only the prosodic features are used. The main advantage of CARTs is that they can combine different discrete and continuous features.

Wright uses in [127] 54 suprasegmental and duration prosodic features to train the trees on the CART algorithm [16]. She achieves 63% of accuracy on the DCIEM map task corpus with 12 DA classes. Shriberg et al. use CARTs in [110] for DA recognition with prosodic features. They use CARTs to recognize a few DAs only, which are very difficult to recognize with lexical (and syntactic) features. These DAs are recognized from prosody only. CARTs are used for example to distinguish statements from questions because questions usually differ from statements by an increasing final F0 curve. Therefore, this CART classifier is trained on statements and questions data only. Levin et al. compare in [77] CARTs with other classifiers, mainly Naive Bayes and MLP classifiers. They use binary grammatical features for this comparison. They show that CARTs outperform the Naive Bayes classifier and that they give comparable results with an MLP. The resulting DA recognition accuracy is about 68% for English and about 66% for German on the NESPOLE corpus.

2.8.3   Memory-Based Learning

Memory-Based Learning (MBL) [32] is an application of memory-based reasoning theory in the field of machine learning. This theory is based on the assumption that it is possible to handle a new sample by matching them with stored representation of previous samples. Hence, in MBL, all known samples are stored in memory for future reference, and any unknown sample is classified by comparing it with all the stored samples. The main advantage of MBL compared to other machine learning techniques is that it successfully manages exceptions and sub-regularities in data. The main drawback of the method is its high memory and computational requirements.

Several methods can be used to compare the stored and recognized samples. The most popular one is the k-Nearest Neighbor (k-NN) [30]. It consists of defining a distance measure between samples, and of finding k stored samples that have the smallest distance to the recognized sample. These k samples are assumed to be similar to the recognized one, and the recognized sample is classified into the dominant class amongst these "neighbors".

Rotaru uses in [104] MBLs in an automatic dialogue acts tagging task on the Switchboard corpus [43] of spontaneous human-human telephone speech. The utterance features are based on word bigrams computed on the whole training corpus. These bigrams are hashed to a given number of features, which optimal value is found experimentally. The hash function uses the letters present in the bigrams and the number of features. The author experiments a various number of neighbors. The best performance is about 72% of accuracy with three neighbors. Levin et al. exploit in [77] MBLs on the NESPOLE corpus.

They use the same features as described in the MLP case (c.f. Section 2.8.1) on the IB1 algorithm [1] with one neighbor. They achieve about 70% of accuracy for English and about 67% for German. MBLs are also used in [76] with the IB1 algorithm. The authors obtain an accuracy of about 74% with prosodic, lexical and context features on a corpus of Dutch telephone dialogues between users and the Dutch train timetable information system.

2.8.4 Transformation-Based Learning

The main idea of Transformation-Based Learning (TBL) [17] is to start from some simple solution to the problem, and to apply transformations to obtain the final result. Transformations are composed in a supervised way. Given a labeled training corpus and a set of possible transformation templates on this corpus, all possible transformations are generated from the templates, after what the transformations are selected iteratively. The templates can be for example: if tag X is after tag Y and/or N previous utterances contain word w, then change actual tag to Z. At each step the "best" transformation (bringing the largest improvement to precision) is selected and applied to the current solution. The algorithm stops when the selected transformation does not modify the data enough, or when there are no more transformations left.

The total number of all possible transformations can be very high. It is thus often computationally expensive to test all transformations, especially since most of them do not improve precision. A Monte-Carlo (MC) approach [126] can be used to tackle this issue: only a fixed number of transformations are selected randomly and used in the next steps. Although this may exclude the best transformation from the retained set, there are usually enough transformations left so that one of them still brings a large improvement to precision.

TBL can be applied to most classification tasks, and has been proposed for automatic DA recognition and some related works. Samuel et al. use in [106] TBL with a Monte Carlo strategy on the VERBMOBIL corpus. They use the following utterance features for DA recognition: cue phrases, word n-grams, speaker identity, punctuation marks, the preceding dialogue act, etc. The resulting DA accuracy is about 71% with 18 dialogue acts. Van der Bosch et al. use in [120] TBLs on the corpus of Dutch telephone dialogues between users and the Dutch train timetable information system, with a very limited DA tag-set. Question-answer pairs are represented by the following feature vectors: six features represent the history of questions asked by system, and the next features represent the recognized user utterance, which is encoded as a sequence of bits, with 1 indicating that the *i-th* word of the lexicon occurs at least one time in the word graph. The last feature is used for each user utterance to indicate whether this sentence gave rise to a communication problem or not, as requested by their application, which final objective is to detect communication problems (incorrect system understanding) between the user and the dialogue system. They achieve to detect about 91% of the communication problems with the rule-induction algorithm RIPPER [24]. Authors show that TBLs outperforms the MBLs technique on this task. Lendvai et al. also use in [76] TBLs with the RIPPER

algorithm. They obtain an accuracy of about 60% with prosodic, lexical and context features on the same Dutch corpus as in the previous experiments.

2.9 Combination of Classifiers

The different classifiers presented so far exploit different kinds of information (lexicon, prosody, sentence structure), which should intuitively bring complementary cues, and should be combined in order to improve the overall DA recognition performances.

A vast literature dealing with classifier combination exists. All these numerous combination methods can be classified into two broad classes, depending on whether they require parameter training or not. Thus, voting strategies [9] usually do not use any parameter training, while Bayesian based weighted product rule [60] or averaging [96] and meta-learners, such as stacking [15] or arbitration [20] do require additional training.

A general architecture for classifier combination is shown in Figure 2.13. The base (or individual) classifiers D_i can be used on the same/different training data, feature spaces, or models. The node D is the combination node.
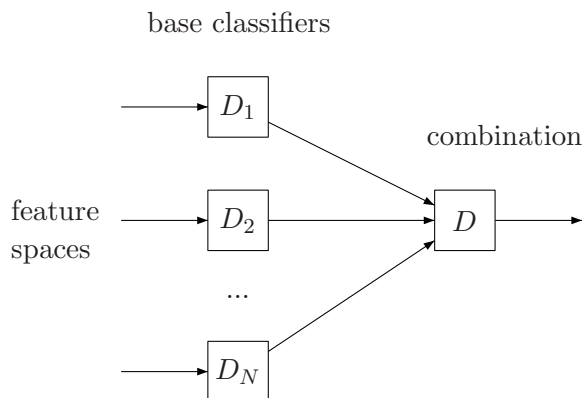
Figure 2.13: Combination of classifiers scheme.

We present next some methods of both classes that we have chosen to combine the different sources of information presented previously.

2.9.1 Naive Bayesian Classifier Combination

A Naive Bayes classifier assumes that all input features are conditionally independent when a value of the classification variable is given. Let $O_1, O_2, ..., O_N$ be a set of N input features and C the class associated to these observations. Under the Naive Bayes assumption, the joint probability of the observations and the class can be simplified as:

$$P(O_1, O_2, ..., O_N, C) = P(C) \prod_{i=1}^{N} P(O_i|C) \tag{2.9}$$

where $P(C)$ is the prior probability of the classification variable C and $P(O_i|C)$ is the conditional likelihood of the feature O_i. Both $P(C)$ and $P(O_i|C)$ are estimated from the labeled training corpus by the training process.

The recognized class $\hat{C}$ is chosen simply as:

$$\hat{C} = \arg \max_C P(C) \prod_{i=1}^{N} P(O_i|C) \tag{2.10}$$

2.9.2 Majority and Weighted Voting

The majority voting combination outputs the class that is chosen by the majority (maximum number of) base classifiers.

Let N be the number of classifiers, $P(C|O, \lambda_i)$ the *a posteriori* probability of class C given by the ith classifier, and $\triangle_{ki}$ the function defined as follows:

$$\triangle_{ki} = \begin{cases} 1 & \text{if } P(C = k|O, \lambda_i) = \max_j P(C = j|O, \lambda_i) \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

the recognized class $\hat{C}$ is then given by:

$$\hat{C} = \arg \max_k \sum_{i=1}^{N} \triangle_{ki} \tag{2.12}$$

Weighted linear voting is a variation of majority voting [9] defined as:

$$\triangle_{ki} = \begin{cases} P(C = k|O, \lambda_i) & \text{if } P(C = k|O, \lambda_i) = \max_j P(C = j|O, \lambda_i) \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

Littlestone et al. [79] propose several *majority voting* algorithms to combine different classifiers. These algorithms are similar to the weighted voting method described above, apart from the fact that the combination weights are trained on a training or development corpus. Different kinds of classifiers are supported: the classifiers are viewed as different prediction algorithms, which are not necessarily trained. The training data is only used to compute the weights. The basic algorithm, called WM, associates each classifier with an initial weight. Every example in the training set is then processed by the classifiers. The final prediction for each example is generated as in weighted voting. If the final prediction is wrong, the weights of the classifiers whose predictions are incorrect are multiplied by a fixed discount δ, where $0 < \delta < 1$, that decreases their contribution to final predictions.

2.9.3   Order Statistics

Combination based on order statistics exploits the ordering of the classes as returned by each classifier to decide on the winning class.

Let X be a random variable with probability density function $f_X(.)$. Let $(X_1, X_2, ..., X_N)$ be a random sample chosen from this distribution. This random sample is arranged in non-decreasing order as:

$$X_{1:N} \leq X_{2:N} \leq ... \leq X_{N:N}$$

The ith value in this progression is the ith order statistic $X_{i:N}$.

Let us now identify these variables with the output of N classifiers. We assume next that these outputs represent the *posterior* probability of each class, given the observation. For a given input x, the outputs of the N classifiers for each class i can be ordered in the following manner:

$$f_i^{1:N}(x) \leq f_i^{2:N}(x) \leq ... \leq f_i^{N:N}(x)$$

The combination of the N classifiers based on the kth order statistics outputs the kth output value for each class $(f_i^{k:N}(x))$ [117]. For example, the *maximum, minimum* and *median* combinations are defined as:

$$
\begin{align}
f_i^{max}(x) &= f_i^{N:N}(x) \tag{2.14} \\
f_i^{min}(x) &= f_i^{1:N}(x) \tag{2.15} \\
f_i^{med}(x) &= \begin{cases} \dfrac{f_i^{\frac{N}{2}:N}(x) + f_i^{\frac{N}{2}:N}(x)}{2} & \text{if } N \text{ is even} \\ f_i^{\frac{N+1}{2}:N}(x) & \text{if } N \text{ is odd} \end{cases} \tag{2.16}
\end{align}
$$

The previous three combination of classifiers represent important qualitative interpretations of the input space. The *maximum* combination method corresponds to the selection of the class with the highest *posterior* probability. Intuitively, this can be interpreted as choosing the output of the classifier that is the most confident about its decision. However, the performance of this method can be degraded when a single classifier always return high probabilities, which may happen when it does not assess correctly its confidence scores.

The *minimum* combination methods is based on the same principle as the *maximum* method, but it focuses on classes that are not so likely to be correct. This method eliminates the least likely classes, because its decision is based on the lowest value for a given class. This method depends less on a single error, because it performs a min-max operation, rather than a max-max operation such as in the previous method.

The *median* method is based on a "typical" representation of each class. This method

outperforms both other ones when it is applied on very noisy data, because the final decision is not compromised as much by a single large error.

More detailed explanations about the combination of classifiers based on order statistics can be found in [119].

2.9.4 Weighted Linear Combination

Let N be the number of classifiers to be combined. Given an observation O_i, each classifier i outputs a score $(P(C|O,i))_{i \in \{1,\cdots,N\}}$ for class C.

The combination of classifiers with weighted linear combination consists in estimating a new score for class C as follows:

$$P(C|O) = \sum_{i=1}^{N} g_i P(C|O,i) \tag{2.17}$$

where $g_1, g_2, ..., g_N$ are non-negatives weights ($\sum_{i=1}^{N} g_i = 1$) assigned to the different classifiers. They are usually determined by estimating how accurate classifiers perform on a validation set.

2.9.5 Combination with a Meta-Learner

The principle of this approach is to create a new classifier, called a *meta-learner*, which is trained on a corpus composed of (or eventually derived from) the classification results of the *base* classifiers. Any base classifier provides a prediction of the unknown class given an input set of test features. In [19], two types of meta-learners are defined: an arbiter and a combiner. Additional arbiters and combiners can also be trained on the set of predictions of lower-level arbiters/combiners.

Arbiter

An *arbiter* [22] is a classifier that is trained using different types of training algorithms to arbitrate between the classifications generated by the base classifiers. Its role is to provide a new winning class when the base classifiers output different classifications. The final decision is then based both on the base classifier and arbiter results, and follows the *arbitration rule*, as shown in the left model of Figure 2.14.

Nodes $D_1, .., D_N$ represent N base classifiers, node A represents the arbiter and node R the arbitration rule. An arbitration rule can be for example: the winning class is the class chosen by the majority of base classifiers D_i and the arbiter A. The output of the arbiter is then used only when the base classifiers can not find an agreement on the winning class. The arbiter is usually trained on a development set that is composed of the most confusing examples, for which the output classes of most of the base classifiers differ.

base classifiers

D_1

combination

D_2

R

feature
spaces ...

final
decision

D_N

A

base classifiers

D_1

combination

D_2

R

feature
spaces ...

final
decision

D_N
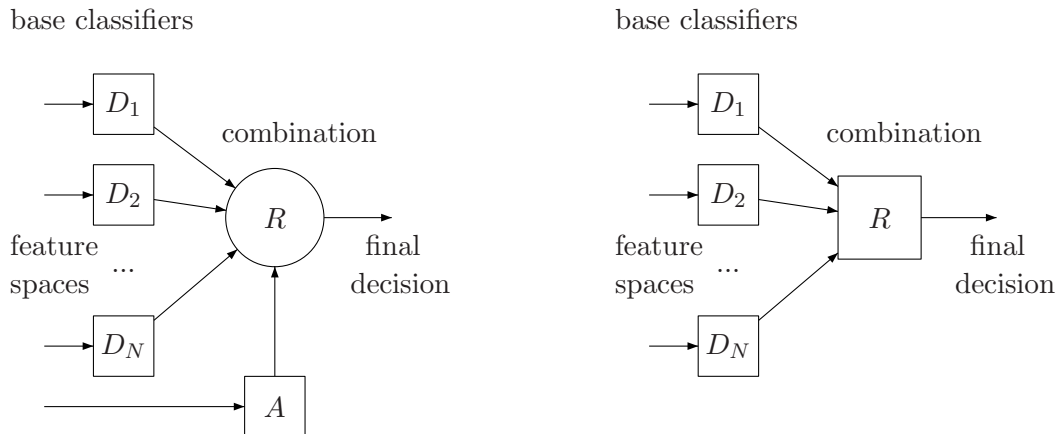
Figure 2.14: Two meta-learner techniques: an *arbiter* on the left and a *combiner* on the right.

Combiner

The principle of the *combiner* [21] is to compose the classification results of the base classifiers by learning the relation between these classes and the correct one.

For instance, when a base classifier D_k usually recognizes correctly the class c_i, the meta-learner might learn that when the base classifier D_k recognizes class c_i, this answer is likely to be correct regardless of the results of the other base classifiers.

A combiner is shown in the right model of Figure 2.14. Node R represents the combiner, the meaning of the other nodes is the same as in the previous case.

2.9.6 Combination of Classifiers for DA Recognition

The combination of knowledge sources with the objective of improving the performance of DA recognition is a research area that has not been deeply explored yet. To the best of our knowledge, only a few works have been published in this area.

Several of these works, as in [113], combine lexical W and prosodic F information only. A simple approach assumes that the lexical and prosodic features are independent:

$$
\begin{aligned}
P(F, W | C) &= P(W|C).P(F|W,C) & (2.18) \\
&\approx P(W|C).P(F|C) & (2.19)
\end{aligned}
$$

The length of the utterance may be represented by a prosodic feature (utterance duration) and also used in DA-specific language models [40].

Shriberg et al. show in [110] that it is better to use prosody for DA recognition in three

separate tasks, namely question detection, incomplete utterance detection and accepts detection, rather than for detecting all DAs in one task.

Questions Detection

Questions usually differ from statements because of their prosodic features, particularly with regard to the final F0 rise. The authors of [110] build decision trees using questions and statements data only. These trees use the F0, duration and speaking rate features to estimate the probability $P(C|F)$. Their first experiment recognizes two DA classes only (questions and statements) from these decision trees. The recognition accuracy is about 74%. In their second experiment, they classify one class of statements and three classes of questions: yes/no questions, wh-question and declarative questions[1]. Their objective is to show which type of questions is well recognized solely from prosodic features. The tree achieves in this experiment an accuracy of 47%. The importance of each feature is shown in Table 2.9.

Feature	Importance in [%]
F0	43.2
Duration	31.8
Pause	21.3
Enrate	3.7

Table 2.9: Importance of prosodic features in classification of statements, yes/no questions, wh-questions and declarative questions.

In conclusion, the most important feature is the F0. Furthermore, the final F0 rises are often associated with yes/no and declarative questions, but not with wh-questions, which confirms the initial hypothesis.

Incomplete Utterances Detection

There are three main types of incomplete utterances: turn exits, self-interruptions and other-interruptions. Although these three cases differ, they are similar in the fact that the utterance could have been completed but was not. The authors of [110] build a classification tree to recognize two classes only: complete and incomplete utterances, which includes all non complete utterance types. An accuracy of about 72% is reached when using mainly duration feature (55%), and secondly energy, speaking rate, F0 and pause features.

---

[1]A declarative question is a question with a similar utterance structure as a statement. It is however characterized by a final F0 rise that does not occur in statements.

Accepts Detection

Accepts are often confused with backchannels and acknowledgments (c.f. Section 2.3.5). The authors of [110] build a prosodic classification tree that uses duration, pause and energy features to discriminate between these classes. With this approach, accepts are discriminated from backchannels with an accuracy of about 69%. The authors further show that accepts are consistently longer in duration and have a higher energy than backchannels. The pause feature is not important in this case.

2.10   Conclusions

The main studies that have been realized in the dialogue act recognition domain have been summarized in this chapter. The concept of a *dialogue act* has been defined. The most popular DA tag-sets have been described with a particular focus on the dialogue acts that are used in our studies. The main knowledge sources that are used for DA recognition have been mentioned. The particular case of prosodic information has been detailed. Several DA recognition approaches have been presented with their advantages, drawbacks and recognition accuracy. Several methods of classifier combination for dialogue act recognition have also been described. Our contributions are based on this state of the art and are described in the next chapters.

Chapter 3

Dialogue Act Recognition with Prosody, Sentence Structure and their Combination

3.1 Introduction

This chapter details our main contributions about dialogue act recognition. The first section deals with three proposed lexical and syntactic approaches that model utterance structure from words and their positions. The fourth approach, called the *clustered unigram model*, based on word clustering is described next. This method addresses the weakness of n-grams model on small DA corpora. Section 3.4 describes our prosodic approaches. Several methods to combine the separate results of both types of approaches (lexical and prosodic) are described in Section 3.5. Our main contributions are briefly summarized in the last section of this chapter.

3.2 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 some 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.

The general problem of automatic DA 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 graphically represent our approach by a very simple Bayesian network with three variables, as shown in Figure 3.1. In this figure, C encodes the dialogue act class of the test sentence, w represents a word and p its position in the sentence.

(a)   (b)
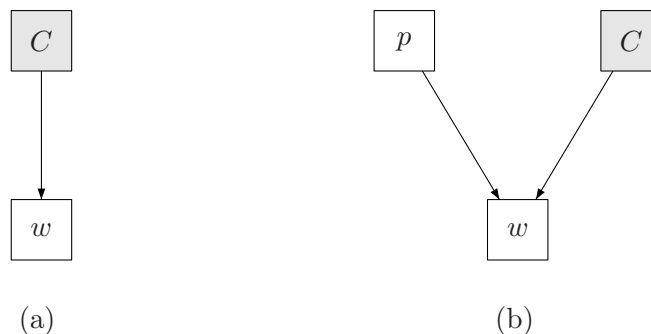
Figure 3.1: Graphical model of our approaches: grayed nodes are hidden.

In the left model of Figure 3.1, $P(w|C,p)$ 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 3.1.

$$P(w_1, \cdots, w_T|C) = \prod_{i=1}^{T} P(w_i|C) \tag{3.1}$$

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

$$\begin{aligned}\hat{C} &= \arg\max_{C} P(C|w_1, \cdots, w_T) \\ &= \arg\max_{C} P(C) \prod_{i=1}^{T} P(w_i|C) \end{aligned} \tag{3.2}$$

This system is referred to as the "unigram" or "Naive Bayes" classifier [44].

On the right part of Figure 3.1, information about the position of each word is included. Considering this additional variable induces the following issues that 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 $(w_i)_{N_i}$:

$$P(w|C, p) = \frac{1}{N_i} \sum_i^{N_i} P(w_i|C, p) \tag{3.3}$$

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.

3.2.1 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 3.2. 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.

$$P(w_i|C, p = \tfrac{1}{1})$$

$$P(w_i|C, p = \tfrac{1}{2}) \qquad P(w_i|C, p = \tfrac{2}{2})$$

$$P(w_i|C, \tfrac{1}{4}) \quad P(w_i|C, \tfrac{2}{4}) \quad P(w_i|C, \tfrac{3}{4}) \quad P(w_i|C, \tfrac{4}{4})$$
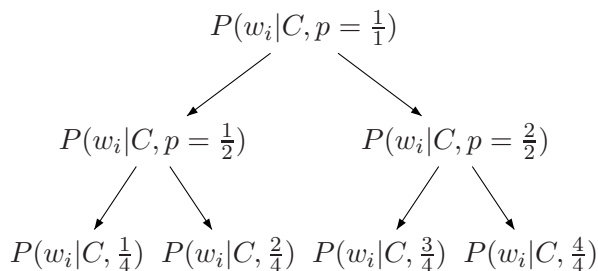
Figure 3.2: Multiscale position tree.

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 [11] to solve the problem of lack of data.

Classification is then realized based on the following equation:

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

$$= \arg\max_C P(C) \prod_{i=1}^{T} P(w_i|C,p_i) \tag{3.4}$$

where each likelihood is estimated at the finest scale possible.

3.2.2 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 ith word in the sentence is filled into the ith 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:

$$\hat{C} = \arg\max_C P(C|w_1,\cdots,w_T,p_1,\cdots,p_T) \tag{3.5}$$

3.2.3 Best Position

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:

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)} \tag{3.6}$$

$$= \frac{P(C)\sum_p P(W,p|C)}{P(W)} \tag{3.7}$$

$$= \frac{P(C)\sum_p P(W|C,p)P(p|C)}{P(W)} \tag{3.8}$$

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,

$$P(C|W) = \frac{P(C)\sum_p P(w_p|C)P(p|C)}{P(W)} \tag{3.9}$$

Finally, maximization gives:

$$\hat{C} \quad = \quad \arg\max_C P(C) \sum_p P(w_p|C)P(p|C) \tag{3.10}$$

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.

3.3   Word Clustering

Performance of classical approaches, such as n-gram models, depends on the size of the DA corpus. They are working especially well with large DA corpus. However, when the corpus size is small, the number of words per DA class is insufficient for a correct estimation of word probabilities. Our approach, a *clustered unigram model*, addresses this issue.

3.3.1   Unigram Model

Assuming that the words of the sentence are independent, the probability of the sentence is given by Equation 3.1. This case corresponds to the left model of Figure 3.3.

3.3.2   Clustered Unigram Model

The words of the application vocabulary are clustered into several groups, in order to reduce the number of parameters to estimate in the unigram models. During recognition, this approach can be modeled by a very simple Bayesian network with three variables, as
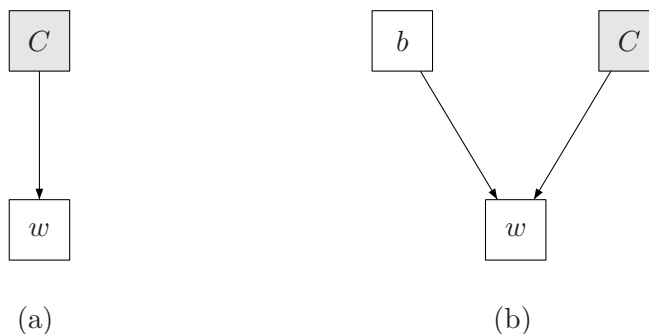
Figure 3.3: Graphical model of dialogue act recognition approaches: grayed nodes are hidden and white ones are observed.

shown in the right part of Figure 3.3. In this figure, C encodes the dialogue act class of the test sentence, w represents a word and b its cluster. Words with a similar functional position in the sentence are clustered into the same group. Mutual information between two neighbor word classes is maximized as described in [83]. The loss of mutual information between two groups b_i and b_j is computed by the following equation:

$$\text{MI-loss}(b_i, b_j) = \sum_{b_k \in \mathcal{B} \setminus \{b_i, b_j\}} I(b_k; b_i) + I(b_k; b_j) - I(b_k; b_i \cup b_j) \qquad (3.11)$$

where $\mathcal{B} = b_1, .., b_j, .., b_k$ is the set of clusters.

Two clusters that cause the minimal loss of mutual information are merged. These clusters are chosen at each step of the clustering bottom-up algorithm as follows:

$$(b_{n_1}, b_{n_2}) = \arg \min_{(b_i, b_j) \in \mathcal{B} \times \mathcal{B}} \text{MI-loss}(b_i, b_j) \qquad (3.12)$$

The clustering of all the words of the vocabulary is realized hierarchically, as shown in Figure 3.4.

The root of this tree (node b in Figure 3.4) contains all the words, and each leave of the tree (nodes $w_1, ..., w_n$) contains a single word. Nodes $b_{11}, ..., b_{1m}$ illustrate word clusters after the first step of clustering. Many levels exist between these nodes and the root of the tree.

During training of groups unigram models, group probabilities $P(b_{ij}|C)$ are estimated for each group on the training corpus.

During recognition, sentences are classified into DA classes using group models. The optimal group model in the tree is not known a priori and must be found empirically.
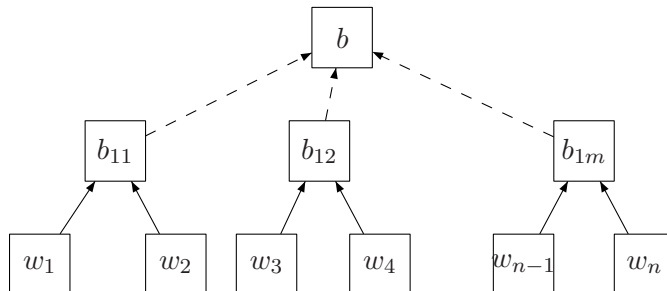
Figure 3.4: Word clusters hierarchy.

3.4 Prosodic Approaches

Following the conclusions of previous studies [114], 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 [61]. 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 is a Muti-Layer Perceptron that outputs $P(C|F)$. The best class is then:

$$\hat{C} = \arg\max_C P(C|F) \tag{3.13}$$

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

$$\hat{C} = \arg\max_C P(C|F) = \arg\max_C P(F|C)P(C) \tag{3.14}$$

When we assume that the *prior* probabilities $P(C)$ are similar for all DAs, we can simplify the previous equation as:

$$\hat{C} = \arg\max_C P(C|F) = \arg\max_C P(F|C) \tag{3.15}$$

This assumption is further used in our experiments.

3.5 Combination of Prosodic and Lexical Approaches

The conclusions of previous studies [110, 113] suggest that prosody brings some valuable information that can not be captured by the lexical models alone. Therefore, in this section we combine lexical and prosodic classifiers to recognize DAs. The main contribution of this work concerns the use and comparison of different kinds of combination methods for automatic DA recognition. As described previously, there are a large number of combination methods, but we use and compare only a few of them, that we consider as the most interesting ones for our application, both unsupervised and supervised.

3.5.1 Normalization into Posterior Probability

The outputs of our classifiers are $P(W|C)$ for the lexical model and $P(F|C)$ for the prosodic one, where C is the dialogue act class, W is the words sequence of the utterance and F represents the prosodic features of the utterance.

Posterior probabilities are easier to compare than raw likelihoods, which depend on the observation probability. Hence, we first normalize the classifier likelihoods to compute the a posteriori class probabilities:

$$P(C = c|W) = \frac{P(W|C = c).P(C = c)}{\sum_{i=1}^{N} P(W|C = i).P(C = i)} \tag{3.16}$$

where c and i represent DAs classes, N is the number of DAs and $P(C)$ is the *prior* probability of class C. We assume that all classes are equi-probable. A similar equation is applied to the prosodic model.

After normalization, it is guaranteed that all outputs of both classifiers are in interval $[0.0; 1.0]$ and the sum of all probabilities of each model is 1, i.e. $\sum_{i=1}^{N} P(W|C = i) = 1$ and $\sum_{i=1}^{N} P(F|C = i) = 1$.

3.5.2 Unsupervised Approaches

The main advantage of unsupervised combination methods is their simplicity. They do not need any parameter training before combination, which usually gives them good generalization properties. Conversely, they are usually less efficient than the supervised ones. It is not possible to use the simple voting (the output class is the class chosen by the majority of the individual classifiers) technique, because there are two individual classifiers only. This is why the scores of the output classes are combined, instead of the class labels themselves, as in majority vote.

Product Combination

The first combination (*product*) is chosen, because it is frequently used as a baseline approach in the DA recognition domain, as described in Section 2.9.6. When we assume

independence of lexical and prosodic information, this combination is only the product of their posterior probabilities:

$$
\begin{aligned}
P(C|W,F) &= \frac{P(W,F|C)P(C)}{P(W,F)} = \frac{P(W|C)P(F|C)P(C)}{P(W)P(F)} \\
&= \frac{P(W|C)P(C)}{P(W)}\frac{P(F|C)P(C)}{P(F)}\frac{1}{P(C)} \\
&= P(C|W).P(C|F)\frac{1}{P(C)}
\end{aligned}
\tag{3.17}
$$

Methods based on Order Statistics

These methods are chosen, because they combine the simplicity of averaging and the generality of meta-learners [118]. These methods are very efficient when there are important variations between component classifiers in certain parts of the joint input-output space. They are less efficient, when the partial training sets cannot be considered as random samples from a common universal data set.

We chose the next three combination methods based on order statistics: *maximum, minimum* and *median.* For each class, the *a posteriori* probabilities returned by both classifiers are ordered, and the final score of each class is respectively the greatest, smallest, and average a posteriori probability for that class as described in Section 2.9.3.

3.5.3 Supervised Approaches

Supervised approaches can be interpreted as training a single meta-classifier from the outputs of the individual classifiers. This classifier usually better models the relation between the outputs of the base classifiers and the reference class than the previous unsupervised simple combination techniques, even though it is often more dependent on the task and corpus.

Weighted Linear Combination

This method is chosen because it combines the simplicity of unsupervised approaches with the performance of supervised methods. This combination computes a weighted linear combination of the *a posteriori probabilities* as:

$$
P(C|W,F) \simeq (g).P(C|W) + (1-g).P(C|F)
\tag{3.18}
$$

where the weight g is optimized via a grid-search on a development corpus.

Combination by an MLP

The last algorithm combines the *a posteriori* probabilities with an MLP. An MLP is used to model a non linear function between the outputs of the base classifiers and the correct DA class as:

$$P(C|W, F) \simeq f(P(C|W), P(C|F)) \qquad (3.19)$$

where $P(C|W)$ and $P(C|F)$ are respectively the *a posteriori* class probabilities of the lexical and prosodic models.

The function $f(\cdot)$ is the mapping function of the neural network. Its output can be interpreted as *a posteriori*, which explains the left-hand term of the equation.

The recognized class is simply given by:

$$\hat{C} = \arg \max_{C} f(P(C|W), P(C|F)) \qquad (3.20)$$

Note that these last two supervised approaches require a development corpus.

3.6 Main Contributions

The most important contributions of my research described in this chapter are summarized below:

- Proposition of three new dialogue act recognition approaches based on lexical information and word position within the utterance:
  - *multiscale position*,
  - *non-linear merging* and
  - *best position approach*
- Proposition of a new dialogue act recognition model, *clustered unigram model*, based on word clustering.
- Analysis and comparison of several methods of combination of classifier in order to improve the recognition of individual classifiers.

3.7 Conclusions

In this chapter, we proposed three approaches that consider information about global word position to improve recognition accuracy. The first one, *multiscale position* approach, exploits a description of the sentence at several levels to smooth the probabilities across these levels. The second one, *non-linear merging* method, models the dependencies between words in the sentence W and their positions P by a non-linear function implemented

as an MLP. The third one, *best position* approach, assumes that words in the sentence W are independent of their positions P, which allows to reliably train the joint probability of the words and positions.

We further proposed the *clustered unigram model*, which has also been designed to specifically address the issue of the lack of training data. The words in the sentence are clustered into several groups based on maximization of mutual information between two neighbor word classes. These groups replace single words during recognition. The number of free model parameters is thus greatly reduced, which makes this method efficient for small DA corpora.

We finally studied several methods of combination of classifiers, and compared their theoretical advantages and drawbacks. Our objective is to improve the quality of DA recognition models by combining different knowledge sources, such as the lexical and prosodic ones.

Chapter 4

Evaluation

4.1 Introduction

This chapter deals with experimental validation of the proposed approaches. The methods that are evaluated respectively exploit lexical information (with and without sentence structure), prosody and a combination of both approaches.

The transcription of utterances into words is computed in two different ways:

- From a manual transcription

- From a speech recognizer

The advantage of the first case is that only the DA recognition methods are evaluated, without the influence of the speech recognizer errors. However, this only provides an upper-bound of the performances of our system in real conditions.

The second type of evaluation aims at evaluating the system in real conditions. These experiments require to set-up a functional large-vocabulary speech recognizer. The accuracy of word recognition influences the quality of dialogue act recognition.

The difference between both cases is discussed in this chapter. This gives a second evaluation of the proposed system, not in terms of DA recognition accuracy, but rather in terms of robustness of the proposed approach to speech recognition errors.

This chapter is organized as follows: first, the speech recognizer used to estimate word transcriptions is described. The next section deals with the LNKnet tool, which is used in several of our experiments, and especially in the prosodic ones. In Section 4.4, we describe our dialogue act corpus. The following section describes the experimental setup for the evaluation of the methods based on sentence structure. The evaluation of the proposed approaches with word clusters is realized in Section 4.6. The following sections deal with experiments with prosody. Then, the combined approaches are tested and compared. The last section of this chapter discuses and compares the performances of our proposed methods. Future directions of research are also proposed here.

4.2 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).

4.2.1 Neural Network Acoustic Model

According to many authors (see e.g. [13]) 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:

$$P(O|S_j) = \frac{P(S_j|O) \cdot P(O)}{P(S_j)}. \tag{4.1}$$

where S_j denotes the jth 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.

4.2.2 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 4.4). The chosen solution has been to merge words into classes and train an n-gram model based on those classes. This should compensate for 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 sentences. The Maximization of Mutual Information (MMI) algorithm (as described in Section 3.3.2) is used for this purpose. It begins by assigning each word to a 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.

4.3 LNKnet Tool

LNKnet [72] is a pattern classification software developed at the MIT Lincoln Laboratory. It contains more than 22 neural network, statistical, machine learning classification, clustering, and feature selection algorithms. Several interesting algorithms proposed from LNKnet are shown in Table 4.1.

LNKnet is originally developed under Sun Microsystem's Solaris 2.5.1 UNIX operating system, but is currently being ported to Red Hat Linux and to Cygwin to run under the Windows operating system. The source code of LNKnet is also distributed. The three following principal interaction possibilities of LNKnet are used in the experiments: the LNKnet graphical user interface, the call to LNKnet commands from shell scripts, and

---

[1]O represents the observation, i.e. in this case the feature vector

Algorithm	Training Type		
	Supervised	Semi-supervised	Unsupervised
NNs	Back-Propagation (BP) Adaptive Stepsize BP Cross-Entropy BP Hypersphere Classifier	Radial Basis Function (RBF) Incremental RBF (IRBF) Learning Vector Quantizer Nearest-Cluster Classifier	Leader Clustering
Pattern Classif.	Gaussian Linear Discriminant Quadratic Gaussian K-Nearest Neighbor Binary Tree Naive Bayes Classifier Support Vector Machine	Gaussian Mixture Model (GMM) Diagonal/Full Covariance GMM Tied/Per-Class Centers GMM	K-Means Clustering E&M Clustering
Feature Selection	Canonical Linear Discriminant Analysis Forward and Backward Search, using N-fold Cross Validation		Principal Components Analysis

Table 4.1: LNKnet algorithms summary.

the control of LNKnet from C programs. In our experiments, we use LNKnet from shell scripts to model a Back-Propagation MLP and a Gaussian Mixture Model (GMM).

4.4 Dialogue Acts Corpus

The Czech Railways corpus, which contains human-human dialogues, is used to validate the proposed methods. It was created at the University of West Bohemia mainly by members of the Department of Computer Science and Engineering in the context of a train ticket reservation application. For the next experiments, it has been labeled manually with the following set of dialogue acts: statements (s), orders[1] (o), yes/no questions (qy) and other questions (q). This DA tag set is based on the reduced tag-set considered in Section 2.3.10, which have been further simplified with regard to the specificities of this corpus.

The number of utterances of this corpus is shown in Table 4.2 with examples.

The LASER recognizer is trained on 6234 sentences (c.f. first part of Table 4.2), while 2173 sentences pronounced by different speakers (c.f. second part of Table 4.2) are used for testing. The word transcriptions given by the LASER recognizer are used to compare the performances of DA recognition with and without manual word transcription.

All experiments of DA 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 following accuracy results have thus a confidence interval of about $\pm$ 1%.

---

[1]Class *order* is based on the *action motivators* class as described in Section 2.3. It contains the utterances, which impose to anybody perform anything. Function of orders are similar as *commands*, but orders contains only the utterances in imperative form.

DA	No.	Example	English translation
1. Training part			
Utter.	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!
qy	282	Řekl byste nám další spojení?	Do you say next connection?
q	1200	Jak se dostanu do Šumperka?	How can I go to Šumperk?
Utter.	2173		

Table 4.2: Composition of the Czech Railways corpus.

4.5 Sentence Structure

We describe in this section our experiments that exploit lexical information and word position in the utterance to recognize DAs.

4.5.1 Multiscale Position

The multiscale position approach (as described in Section 3.2.1) trains a conditional unigram model of $P(w_i|C, p)$ at different scales, as shown in Figure 3.2. Recognition is then performed based on Equation 3.4.

Figure 4.1 shows the recognition accuracy of this method in function of the minimum number of word occurrences 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 maximum depth of the tree used in this experiment is 3, which defines 8 segments. We do not expect better results with a larger number of leaves in the tree because of the small size of our corpus. The unigram model recognition accuracy is also reported on Figure 4.1 for comparison.

The recognition accuracy of each class is shown in the second section of Table 4.3.

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 lack of training data issue.

Non-Linear Merging

The *Non-Linear* model merges lexical and position information with 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
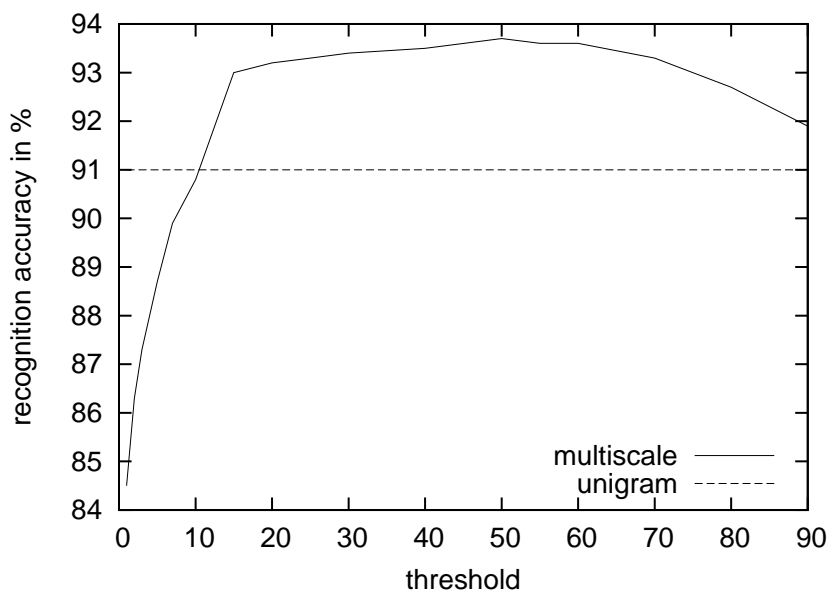
Figure 4.1: Dialogue act recognition accuracy of the multiscale position tree system. The X-axis represents the minimum number of words in the tree, and the Y-axis plots the DA recognition accuracy.

and 4 output neurons, which encode the *a posteriori* class probability. The dialogue act class is given by Equation 3.5.

The recognition results of this method is also shown in the second part of Table 4.3, along with the results obtained with the baseline unigram model. The global recognition accuracy of this model is 94.7%.

Best Position Approach

The third position-based proposed approach is the *Best Position* method, which recognizes dialogue acts based on Equation 3.10. 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 4.2.

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

$$P(p = 1|C) = 1 \text{ and } P(p > 1|C) = 0 \text{ for all } C$$
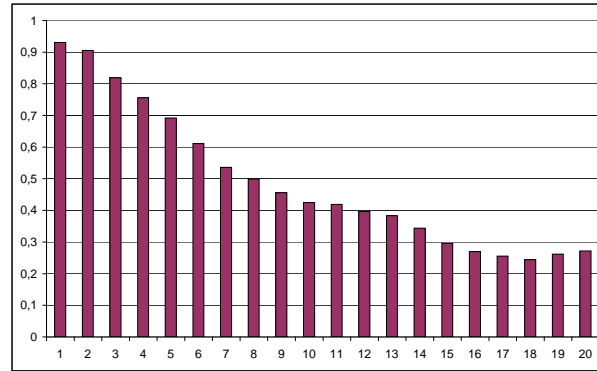
Figure 4.2: DA recognition accuracy on the development corpus when only a single position is considered.

After the gradient descent algorithm, the resulting weights are shown in Figure 4.3.
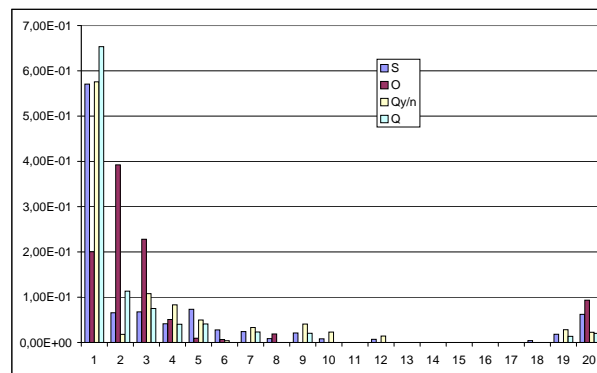
Figure 4.3: Weights obtained after the gradient-descent algorithm.

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 confirm our intuition.

Then, using the weights shown in Figure 4.3, recognition is performed on the test corpus. The results are given in the last line of Table 4.3.

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

Approach/ Classifier	accuracy in [%]				
	s	o	qy	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**
2.3 Best position	93.6	95.2	97.2	94.3	**95.8**

Table 4.3: Dialogue act recognition accuracy for different sentence structure approaches and different classifiers with manual word transcription.

4.6 Clustered Unigram Model

In the following experiments, classes of words are used instead of words to train the lexical models. Two different types of clustering are tested. In the first one, words are clustered independently of their DA class. Hence, word clusters are the same for all DA classes. The main advantage of this option is that the number of word occurrences within every word cluster is larger. A drawback is that word clusters do not take into account the specificities of each DA class. In the second implementation, a word cluster is created for each DA class. The unigram statistics are not estimated as robustly as in the previous solution, but they should be more accurate.

The optimal number of word clusters depends on the corpus characteristics. In our experiments, it is found empirically using a cross validation procedure on the development corpus. Table 4.4 shows the recognition accuracy of both variants of clustered unigram model. The baseline unigram model recognition accuracy is reported in the first row of this table. The global recognition accuracy of the DA-independent clustered unigram model is 91.1%, which is comparable with the unigram model. The DA-dependent clustered unigram model gives 92.1% recognition accuracy, which slightly outperforms the unigram model, our baseline approach. The DA error rate is thus reduced by 12%.

Approach/ Classifier	accuracy in [%]				
	s	o	qy	q	Global
1. Lexical information					
1 Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Word clusters					
2.1 Common clusters	94	65.6	93.6	91.8	**91.1**
2.2 Clusters per DA	92.5	76	92.5	93.8	**92.1**

Table 4.4: Dialogue act recognition accuracy for different clustered unigram model in %.

4.7 Prosody

In the following experiments, we investigate the possibility to recognize dialogue acts (statements and questions) in Czech from prosodic features only. The objective of these experiments is to study the importance of prosody (namely F0 and energy) to automatically recognize three DA classes: statements, yes/no questions and other questions (wh-question, etc.).

The DA tag-set of these experiment is a subset of our DA tag-set with four DAs which can be (as presented in Section 2.5.2) characterized solely by prosody as:

- Statements are usually characterizes by a falling intonation.
- Questions (particularly declarative and yes/no) are often characterized by a rising F0 contour.

The class *questions* has been divided into two classes: yes/no and other questions, which differ with regard to their prosodic characteristics. These experiments shall complement the study realized by [110] about the importance of prosody in DA recognition.

4.7.1 Analysis of Fundamental Frequency

We first compare the F0 curves for every DA. The mean and variance of the F0 values are computed for each of the twenty F0 features per utterance. The mean values of F0 are shown in Figure 4.4. The variances are very small (in the interval [0; 0.02]) and are not plotted to make the curves as visible as possible.

In the first part of the segment the F0 curves are very similar for all DAs. The last third of the segment is the most discriminating. The F0 slope of yes/no questions (qy) is clearly increasing, decreasing for statements (s) and almost horizontal for other questions (q). This corresponds to the prosodic characteristics described previously. Note that these are the average curves, and that individual exceptions might occur in the corpus.

This first visual analysis is completed next by a more detailed analysis of the mean and variance of the F0 slope at the end of the utterance. Four values of F0 are computed on the last 1.5 seconds of the utterance by an autocorrelation function. A linear regression is then performed on these four values to approximate the final slope. Table 4.5 shows the number of utterances in function of this slope. The "/" and "\" symbols respectively represent an increasing and decreasing final F0 slope. Most of the yes/no questions are actually characterized by a positive F0 slope. Conversely, the number of "other questions" with a positive and negative F0 slope is well-balanced.

In the second part of Table 4.5, the slope values are splitted into 3 segments, depending on whether the slope is strictly less than -0.03, between -0.03 and 0.03, and above 0.03. The lower range should be characteristic of statements, and the higher range of yes/no questions. This analysis shows that most of the DAs (about 80%) have a flat F0 slope and
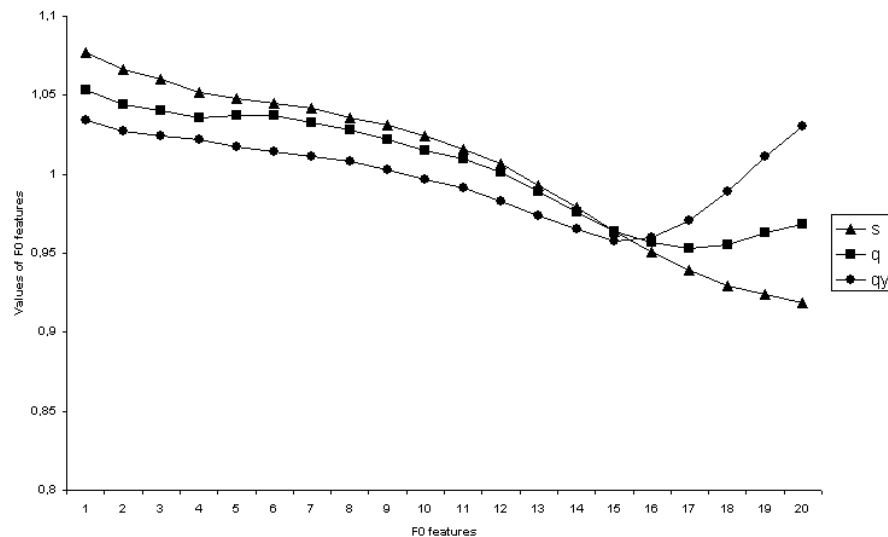
Figure 4.4: F0 curves for three types of DAs: *s* curve for statements, *q* curve for other questions, *qy* curve for yes/no questions.

thus are not discriminating. Only about 30% of yes/no questions have a clear increasing F0 slope, which can be useful information to recognize these DAs.

Class	\	/	< -0.03	[-0.03; 0.03]	0.03 <
s	69.7	30.3	9.1	84.9	6.0
q	46.9	53.1	4.2	85.4	10.4
qy	26.8	73.2	3.5	66.9	29.6

Table 4.5: Analysis of the F0 slope at the end of sentences for the three DA classes.

4.7.2 DA Recognition with F0

We use in this experiment only the fundamental frequency features that are computed on the final segments of each DA. A GMM classifier with 5 Gaussian mixtures is trained on these features. This number of Gaussians is reasonable with respect to the size of the training corpus. This classifier models $P(F|C)$. Table 4.6 shows the confusion matrix obtained by this model. The global accuracy is 42%, which is significantly above the random guessing accuracy (33%).

4.7.3 Analysis of Energy

We analyze next the energy curve in function of the DA. As previously, the mean and variance are computed for all of the twenty energy features. The mean energy curves per

	Recognized class in [%]		
Pronounced class	s	q	qy
s	**41.5**	38.9	19.6
q	40.0	**36.2**	23.8
qy	26.1	28.8	**45.1**

Table 4.6: GMM confusion matrix for recognition of three DA classes solely by fundamental frequency in %.

DA are shown in Figure 4.5. One can observe on this figure that the statement (s) curve is less variable, i.e. closer from a straight line than both question curves. This difference is the largest near the end of sentences. However, the variances of the energy features are globally very high (up to 0.13), with an important overlap between all DA classes. Therefore, in our experimental setup, we can conclude that energy features are not very discriminant between DAs.
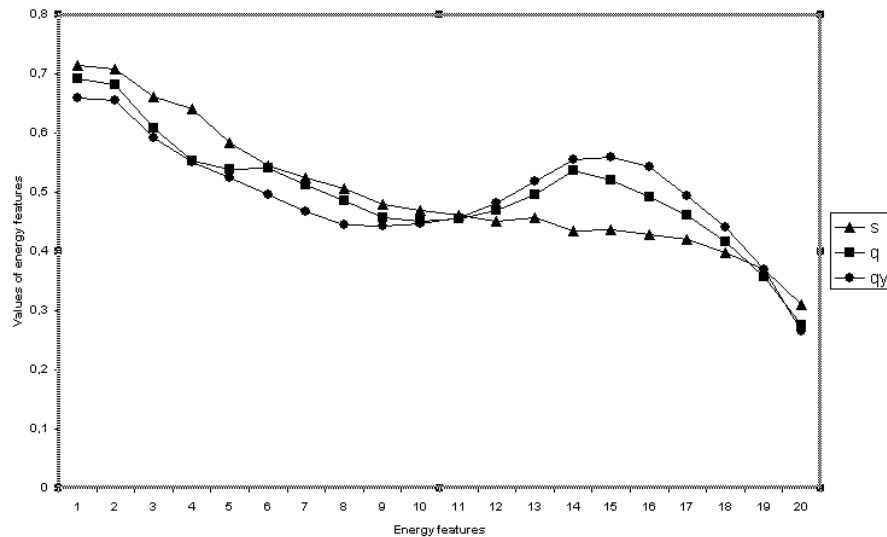
Figure 4.5: Energy curves for three types of DAs: *s* curve for statements, *q* curve for other questions, *qy* curve for yes/no questions.

4.7.4 DA Recognition with Energy

We use 20 energy features to train a GMM classifier (as in the previous case). Table 4.7 shows the confusion matrix obtained in this experiment. The best recognition accuracy 40% is obtained with 3 Gaussian mixtures. Increasing the number of Gaussians decreases the accuracy, because of the lack of training data.

	Recognized class in [%]		
Pronounced class	s	q	qy
s	**41.1**	26.1	32.8
q	28.9	**33.8**	37.3
qy	25.7	28.8	**45.5**

Table 4.7: GMM's confusion matrix for recognition of three DA classes from the energy only in %.

4.7.5 Discussion

A remarkable conclusion that can be drawn from the previous experiments is that prosodic features, such as the F0 and energy, are less discriminant at the beginning of utterances than at their end. This confirms the results of several previous studies [115, 127, 128]. Furthermore, the energy is less discriminant than the F0, probably because of its high variance. This is confirmed by a slightly higher DA recognition score of F0, which may be explained by the relatively different F0 slopes of the three chosen DAs, as shown in Table 4.5. Yes/no questions are quite different from the two other DAs, which is probably related to the fact that this class obtains the best recognition accuracy (45%). F0 slopes of statements and other questions are similar, which causes some confusion between these DA classes during recognition.

The global recognition scores of both DA recognition experiments are significant higher than random guessing (about 10% in absolute). Hence, prosody brings some relevant clues for DA recognition, although it can not be used alone.

4.7.6 DA recognition with F0 and energy

Other studies [110] as well as our own previous prosodic experiments suggest that prosodic features are not sufficient to recognize all DAs with a good accuracy. Therefore, we combine them next with lexis (and syntax). In this experiment, four DAs (statements (s), orders (o), yes/no questions (qy) and other questions(q)) are recognized. Both basic prosodic features, F0 and energy, are used.

Table 4.8 shows the recognition accuracy with two different classifiers that use only prosody: a GMM and an MLP. The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The best recognition accuracy is obtained with the 3-mixtures GMM. It is difficult to use more Gaussians, because of the lack of training data, mainly for class o.

These recognition scores are still much lower than the ones obtained with lexical information, but we will show next that prosody may nevertheless bring some relevant clues that might not be extracted from words sequence.

Approach/ Classifier	accuracy in [%]				
	s	o	qy	q	Global
1. Lexical information					
1 Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Prosodic information					
2.1 GMM	47.7	43.2	40.8	44.3	**44.7**
2.2 MLP	38.7	49.6	52.6	34.0	**43.5**

Table 4.8: Dialogue act recognition accuracy in % for prosodic classifiers compared to our baseline, an unigram model.

4.8 Combination of Prosodic and Sentence Structure Approaches

We first study the correlation matrix of both lexical and prosodic GMM classifiers in Table 4.9: this matrix shows the proportion of examples that are classified correctly and incorrectly by both classifiers. For example, 40.04% of the examples are classified correctly by both classifiers while 5.57% of the examples are not recognized by any classifier. An interesting remark from this table is that 2.12% of the examples are recognized by the prosodic classifier, but not by the lexical one. This suggests that there is a small but significant potential improvement that can be obtained by considering prosodic information as well.

	lexical correct	lexical incorrect
prosodic correct	40.04	2.12
prosodic incorrect	52.28	5.57

Table 4.9: Correlation of classification error rate of both classifiers in %.

4.8.1 Evaluation of Combination Methods

In these experiments, we evaluate several combination approaches described in Sections 2.9 and 3.5. We use here our baseline lexical unigram model for its simplicity and the prosodic GMM for its performance in DA recognition task.

Table 4.10 compares the recognition accuracies of both independent lexical unigram and prosodic GMM models with their combination, which is realized using several combination methods.

We can note that, amongst order statistics combiners, the minimum and median ones are better than the maximum one. But we can also observe that every unsupervised combination gives a lower accuracy than the lexical classifier alone. This can be explained by the fact that we combine only two classifiers, and most importantly because of the big difference between each individual classifier recognition accuracy. Indeed, this is confirmed

by the *weighted linear* combination, which optimal weight is 0.97 in favor of the lexical approach.

The best recognition accuracy is obtained with the MLP combination, which reduces the lexical word error rate by an absolute 2%. This figure can be compared with the 2.12% shown in Table 4.9. Therefore, this combination method is chosen for the following experiments.

Approach/ Classifier	accuracy in [%]				
	s	o	qy	q	Global
1. Lexical information					
1 Unigram	93.5	77.6	96.5	89.9	**91.0**
2. Prosodic information					
2 GMM	47.7	43.2	40.8	44.3	**44.7**
3. Unsupervised Combination					
3.1 Product	81.1	76.8	86.2	64.4	**72.3**
3.2 Maximum	81.8	81.6	88.3	57.9	**69.4**
3.3 Minimum	80.0	73.6	84.8	64.6	**71.7**
3.4. Median	81.3	81.6	88.3	63.2	**72.2**
4. Supervised Combination					
4.1 Weighted Linear	88.5	90.4	92.9	94.2	**92.3**
4.2 MLP	90.3	88.0	92.9	97.3	**94.3**

Table 4.10: Dialogue act recognition accuracy for individual lexical and prosodic classifiers and their combination in %.

4.8.2   Combination of Sentence Structure Model and Prosody

The *Non-linear merging* scheme is evaluated in this experiment. In this approach, an MLP encodes both lexical and position information, as described in Section 3.2.2, while prosody is modeled by a GMM.

The combination of both models is realized with another MLP, which takes as input the best results presented in the previous section.

The winning dialogue act class is given by Equation 3.20. The MLP is composed of three layers as follows: 4 (for each DA class) times 2 (two classifiers to combine) input neurons, 9 neurons in the hidden layer and 4 output neurons, which encode the *a posteriori* class probability.

Table 4.11 shows the recognition accuracy of this experiment.

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.

Approach/ Classifier	accuracy in [%]				
	s	o	qy	q	Global
1. Sentence structure					
1 Non-linear	90.3	83.2	91.1	98.8	**94.7**
2. Prosodic information					
2 GMM	47.7	43.2	40.8	44.3	**44.7**
3. Combination with a MLP					
3 MLP	91.5	85.6	94.0	98.7	**95.7**

Table 4.11: Dialogue act recognition accuracy of combination of Non-linear merging and prosodic GMM models in %.

4.9 Recognition with LASER Recognizer

Table 4.12 shows the same DA recognition scores as before, but with an automatic word transcription obtained by the LASER recognizer instead of a manual transcription. 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%.

The first section of Table 4.12 shows the DA recognition accuracy of an unigram model, our baseline. The second section shows the accuracy of our proposed approaches based on the word position within the utterance. The third section shows the recognition accuracy of our main prosodic experiments. The recognition accuracy of the combination of prosodic and *Non-linear* merging approaches are shown in the last part of this table. The topology of all classifiers are the same as in the previous case, when the word transcription was labeled manually.

Approach/ Classifier	accuracy in [%]				
	s	o	qy	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**
2.3 Best position	92.1	86.4	95.3	92.2	**93.6**
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					
4 MLP	88.5	77.6	90.4	97.3	**93.0**

Table 4.12: Dialogue act recognition accuracy for different approaches/classifiers and their combination with word transcriptions obtained from the LASER recognizer.

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.

4.10   Conclusions

In this chapter, we have evaluated several new methods for automatic DA recognition.

First, we have studied the influence of word positions in a dialogue act recognition task. Three proposed approaches 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 common situation in dialogue act recognition.

Next, we have presented two variants of a new method for automatic dialogue act recognition based on word clusters. Words in the utterance have been replaced by word clusters. In the first method, words are clustered independently of their DA class. In the second implementation, a word cluster is created for each DA class. The recognition accuracy of the best method, a *DA-dependent clustered unigram model*, is 92.1%. Compared to the baseline system, the dialogue acts error rate is reduced by 12%. We show that it is possible to replace words in the utterance by word clusters and handle thus the issue of the small corpus size, where the number of words per DA class would not be large enough to reliably estimate word probabilities.

We have shown in the prosodic experiments that is is not possible to recognize all DAs with a good accuracy only by prosodic features, but that prosody can help to recognize several particular DAs. We have thus proposed to combine sentence structure and prosodic methods.

In Section 4.8, we have studied and compared different methods to combine lexical and prosodic information in the context of automatic dialogue act recognition. We have shown that it is possible to improve our baseline result by combining two lexical and prosodic classifiers with a MLP. A statistically significant 2% absolute improvement is then obtained, which is actually very close to the potential improvement derived from the correlation matrix between both classifiers. This confirms that prosodic clues are *complementary* to the lexical ones, as it has been already suggested in other studies such as [110, 127]. All the other combination schemes, and in particular the unsupervised ones, do not reach the level of the lexical classifier alone. This shows the importance to fine-tune the combiner on a development corpus in our experimental set-up. This might result from the large difference in the performances of both classifiers, and also from the small number of experts that are combined.

One of the systems that used both lexical and position information has then been enhanced by further considering prosodic information. The supervised combination with an MLP 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 LASER 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.

One 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 contribution of the statistical and prosodic features, 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.

Chapter 5

Semi-automatic Labeling

5.1 Introduction

Automatic dialogue act recognition is mainly used in dialogue systems. Generally, different dialogue systems use a different set of DAs, depending on their application domain. A specific dialogue act corpus may thus be required for each application.

One of the main issue in the domain of automatic dialogue act recognition concerns the design of a fast and cheap method to label new corpora. Manually labeling corpus is very time consuming. This represents an important part of the project costs. Conversely, completely unsupervised methods are less efficient. We study in this chapter the use of a semi-automatic approach for DA corpus creation.

The resulting corpus will be further used to validate our previous recognition approaches described in Chapter 3 and presented in [64, 65, 71, 70, 68, 67]. These methods were tested on a Czech corpus with four DAs only, and will next be evaluated on another language (French) and with a larger set of DAs.

The following section presents a state of the art of semi-automatic learning approaches. A general view in this domain is given in the first subsection. The second subsection concerns the application of these methods in dialogue act annotation. Section 5.4 describes the process of corpus preparation, which is composed of the following main steps: definition of our DA tag-set, development of the software tool that is used for manual labeling, and creation of the initial corpus. The proposed approaches, which are based on the Expectation Maximization (EM) [35] algorithm and confidence measures, are described in Section 5.5. Section 5.6 evaluates our methods and compares them with the baseline EM procedure. In the last section, we discuss the research results and propose some future research directions.

5.2 General Methods for Semi-supervised Training

Semi-supervised learning is a special case of training where the classifier or model is trained on labeled and unlabeled data. Manually labeled data are usually difficult and expensive to obtain, while row data are relatively easy to get. Semi-supervised training starts from an existing classifier, usually trained on a small corpus of manually labeled data, and iteratively trains new versions of this classifier on both the small labeled corpus and a large amount of unlabeled data. The main semi-supervised approaches are described next.

5.2.1 Expectation Maximization

The Expectation Maximization (EM) algorithm is the most popular method used in semi-supervised training. It is used to train models in the maximum likelihood sense with hidden variables. The EM algorithm is composed of two main steps:

1. Expectation: computes the observation likelihood, with given evidence for hidden variables equal to their expected value;

2. Maximization: computes the maximum likelihood estimates of the parameters by maximizing the likelihood computed in step (1).

The "new" parameters estimated in step (2) are used in the next iteration at step (1). This process iterates until convergence.

When the hidden variable represents the labels of a part of the training corpus, the EM algorithm can then be used to achieve semi-supervised training. This method is closely related to the self-training principle described in [129].

This basic EM algorithm is very successful when data conform to the generative assumption of the model. But when this assumption is not satisfied, the performance of the algorithm might degrade.

The EM algorithm is used in several domains. Nigam et al. exploit in [92] the EM algorithm with a naive Bayes classifier. They use a small number of labeled and a large amount of unlabeled data for text classification. Two improvements of the EM algorithm are also presented: the use of a factor to weight the importance of the unlabeled data, and the use of a many-to-one correspondence between GMMs and classes (unlike the usual one-to-one correspondence). Their algorithm is evaluated on three different corpora: UseNet news articles (20 Newsgroups [55]), web pages (WebKB [31]), and newswire articles from Reuters. The authors show that the use of unlabeled data reduces the classification error by up to 30%.

Lamel et al. use EM in [73] for semi-supervised training of acoustic models. The basic idea is to use a speech recognizer trained on a very small corpus to automatically transcribe unlabeled audio data on the DARPA DTD-2 corpus [23]. They show that the annotated audio data is not very important to train the acoustic models. The acoustics models are thus initialized only with 10 minutes of manually annotated data. The recognition results

of supervised and semi-supervised training are very closed: the Word Error Rate (WER) for supervised training on 50h of training data is about 21%, which is to compare with about 24% for semi-supervised training. The authors also show that it is possible to estimate the models on automatically annotated data without filtering potentially incorrect words in automatic transcription.

Lauritzen describes in [75] how to apply EM to Bayesian networks. The EM algorithm is used to found maximum likelihood estimates or penalized maximum likelihood estimates with hierarchical log-linear models and recursive models for contingency tables with missing data. Several experiments show that the likelihood function has a number of local maxima, and a direct maximization of likelihood might give erroneous results with missing data. The authors mention that the use of penalized likelihoods, where penalty may be computed for example from a prior density as explained in [45], give better results.

5.2.2 Transductive Support Vector Machines

Standard Support Vector Machines (SVMs) [28] approaches are trained only on labeled data. Their main goal is to find a maximum margin linear boundary in the Reproducing Kernel Hilbert Space [82]. Transductive Support Vector Machines (TSVMs) [91] is an extension of standard SVMs on unlabeled data. The goal is to find a labeling for the unlabeled data that maximizes the margin on the labeled and unlabeled data.

Figure 5.1 compares the linear separators of the SVMs and transductive SVMs on an example. Symbols "+" and "#" mark two types of labeled data. Unlabeled data are represented by symbol "o". The inductive SVMs solution is represented by the full line and the transductive ones by the dashed line.
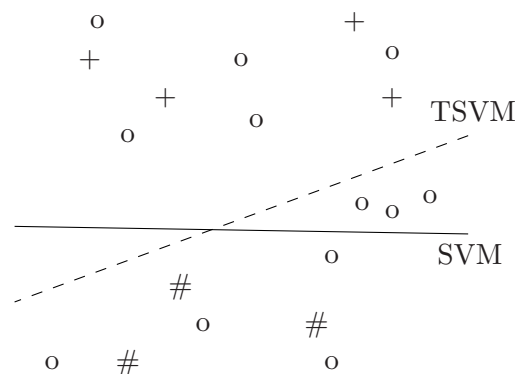
Figure 5.1: SVMs and TSVMs on labeled and unlabeled data.

The advantages and drawbacks of transductive SVMs are respectively described in [56] and [130].

5.2.3 Other Semi-supervised Approaches

Several semi-supervised approaches iteratively increases the size of the initial labeled data set. The large labeled corpus obtained at the end of these iterations is used to train the final classifier. A common method to increase the size of that data set is to exploit the EM algorithm at each iteration to estimate the labels of a new part of the unlabeled corpus. A confidence measure is often used to filter out the recognized examples that are likely to be wrong (*co-training*, *self-training*, etc.). In the case of *active learning*, those probably erroneous examples may be shown to the user, who can correct their labels. The process described above is iteratively repeated until there is no unlabeled data left, or until a predefined number of iteration is reached.

Co-training

Co-training (as described by Blum et al. in [12]) assumes that the features that describe the data are redundant. The feature space X is divided into two subspaces $X = X_1 \times X_2$, where X_1 and X_2 correspond to two different "views" of an example. It is assumed that each view is sufficient for correct classification. Two classifiers are respectively trained on these partitions. Both classifiers are then used to classify the unlabeled examples in the corpus. The most confident predictions of each classifier are selected, labeled and included into the training pool. The authors show in the best experiment that co-training outperforms the experiment with only labeled data by up to 6% in absolute accuracy. These experiments are performed on a corpus composed of 1051 web pages collected by four universities.

Active Learning

Active learning [25] is also an iterative algorithm that starts from an initial labeled training set and expands it at each iteration. Unlabeled samples are automatically labeled, and some of these examples are also presented to the user who can correct their labels. The core of active learning is to design an effective strategy to choose the examples that shall be presented to the user. Several solutions have been proposed in the literature.

A common method proposed by several researchers and described in [80] is *selective sampling*. This method is based on the observation that the distribution of the training data is not uniform and some examples are more representative than others. By manually labeling only a small number of these representative examples, the performances obtained by the final classifier are comparable to the case where many unlabeled examples are randomly chosen.

Another selection strategy, *partition sampling*, is presented by Souvannavong et al. in [111]. Their selection sequence is given by a greedy maximization of the error reduction, when the ground-truth of the corpus is known. They compute the improvement of the accuracy obtained after adding each sample into the labeled data set (all labels are known) at each iteration. The sample that maximizes the decrease of the classification error is

then selected. The authors compare the performances of this approach with those of *random sampling* on the TRECVID database [78]. Partition sampling outperforms random sampling and almost reaches its optimal learning sequence, i.e. the sequence of examples that maximizes the recognition accuracy.

For more information about active learning, please refer for example to [7] or [116].

5.3 Semi-Automatic Training Methods for Dialogue Act Labeling

The semi-automatic training methods that have been used in the particular task of DA labeling are described next.

5.3.1 Lexical Information and the EM Algorithm

Venkataraman et al. use in [122] semi-supervised training for segmentation and classification of DAs on the Speech in Noisy Environments (SPINE) corpus [39]. They use a small hand-labeled training set and a larger unlabeled data. DAs are modeled by HMMs where states represent DAs. The utterances correspond to the observations generated by the states. This system can be described by the following equation:

$$P(C, W) = \prod_{i=1}^{N} P(C_i|C_{ih})P(W_i|C_i) \tag{5.1}$$

where $C = C_1, .., C_N$ is a random variable that represents the DA sequence. The observations are the N utterances $W = W_1, .., W_N$ in the dialogue. Each utterance W_i is composed of a words sequence $(w_1, .., w_n)$. C_{ih} is the dialogue history of the ith DA. In the paper, the Markov assumption is used.

The EM algorithm is used to maximize the likelihood of the system on the training data. Only the most probable DAs are used during the maximization step. Several size of bootstrap data, several number of iterations and two language models, unigram (without DA context) and 3-gram (context of two previous DAs), are tested. The experiments show that this method works very well when the local context of DAs is considered (3-gram), and when only a small initial labeled data set is used.

5.3.2 Prosody and EM

The system previously described is improved in [121] by further considering prosodic information. HMMs now integrate lexical, prosodic and contextual information. Equation 5.1 now becomes:

$$P(C, W, F) = \prod_{i=1}^{N} P(C_i|C_{ih})P(W_i, F_i|C_i) \qquad (5.2)$$

where $F = F_1, .., F_N$ is a sequence of observed prosodic features.

The reported experiments show that the combination of these information sources significantly improves the recognition accuracy compared to when lexical information is used alone. The DA recognition error rate thus decreases relatively of about 15%.

5.3.3 Active Learning

Active learning is successfully used in [123] to increase the amount of training data. The authors use the same methods as in the first case, i.e. HMMs with n-gram models, but in this case the EM algorithm is enhanced by active learning.

Furthermore, they propose an alternative model based on maximum entropy. The role of the entropy is to represent the correlation between features, such as the identity of the first two and the last two words, a bigram of the first two words, etc.

The maximum entropy model is then used to discriminate between the data that is likely to be classified correctly (with large entropy values) and the data that should be labeled manually (with small values of entropy). The experiments show that the maximum entropy approach improves the classification accuracy from one iteration to another, which is not the case for the baseline HMM classifier.

More details about semi-automatic training in the domain of dialogue acts are also given in [54].

5.4 Initial Corpus Preparation

Every semi-automatic labeling approach that has been described previously needs a small labeled corpus to initialize the models. In this section, the chosen corpus and dialogue act tag-set are described and discussed. Then, we describe the process and tools designed to build our initial small training and testing corpora. The composition of both corpora is also detailed in this section.

5.4.1 Choice of the Source Corpus

The French broadcast news corpus ESTER [34] has been chosen, because:

1. It is in French language.
2. It is available in the Parole team.
3. An important part of the corpus (80h) is transcribed into words.

4. It contains natural human-human speech, which makes it suitable for real-world applications.

The transcription files are in xml format [14]. The speech files are in wave format [27].

5.4.2   Baseline DA Tag-set

The set of DAs must be defined before realizing manual labeling. Our initial tag-set is based on the SWBD-DAMSL and MRDA systems. SWBD-DAMSL contains 42 clustered DA classes and MRDA 11 general DA tags and 39 specific DA tags. The DAs from SWBD-DAMSL and MRDA are described in Section 2.2. For ESTER, we have chosen a subset of DAs from these baseline tag-sets that has been completed with additional specific radio-oriented DAs. All the chosen DAs are described next.

5.4.3   Specific Dialogue Acts in ESTER

First, we describe the DAs that are neither contained in the DAMSL nor MRDA tag-set.

Radio Info

The DA "Radio Info" corresponds to a special statement in broadcast news, which provides information about the radio the user is currently listening to, and/or about the actual time. The examples are: "France Inter, il est 5 heures" (*France Inter, it's 5 o'clock*) or "France Info à Marseille" (*France Info in Marseille*). The Radio Info DA usually occurs at the beginning or at the end of every news section. Thus, an additional information that can be used to recognize this DA is the current position in the news stream. This information is not used yet, but we suppose to use it later.

Person Speaking

The function of this DA is to give information about the current speaker of the news. Its position in the news is usually close (it often immediately precedes or follows) to the Radio Info DA. Examples of Person speaking are: "Joël Collado" or "Présenté par Hervé Guillemot" (*Presented by Hervé Guillemot*).

Dialogue Subject

The Dialogue Subject DA is a particular statement that is specific to broadcast news. It usually gives information about the topic of the following news: "Météo France" (*France, weather forecast*), "Sport football" (*sport, football*) or "La politique ce matin" (*Politics, this morning*). This DA may bring relevant cues concerning the place of the reportage "À Londres la contre-offensive des avocats de Pinochet" (*In London, the contra-offensive*

of the Pinochet's advocates). The position of a Dialogue Subject DA is usually closed to both Radio Info and Person Speaking DAs.

5.4.4 Dialogue Acts from SWBD-DAMSL and MRDA Tag-sets

We review next the DAs that have been used in our work, which come from the SWBD-DAMSL and MRDA tag-sets. We describe some special properties of the French language and the simplifying assumptions considered in the following work.

Yes/No Question

Three groups of French yes/no questions can be derived: the DAs in the first group have syntactically the same structure as statements and can be recognized from prosody only. For instance, the French sentence "Tu peux ouvrir la fenêtre?" (*You can open the window?*) could be classified without prosody as a statement or as a yes/no question.

The second group of yes/no questions is characterised in French by the inversion of the couple subject-verb, such as: "Peux-tu ouvrir la fenêtre?" (*Can you open the window?*).

The third group of French yes/no question is identified by the interrogative form "Est-ce que" at the beginning of the utterance, such as: "Est-ce que tu peux ouvrir la fenêtre?".

Prosodic information is not included in the first version of our system. Hence, for this first version, the "prosodic" yes/no questions are added into the statement class. Also, the inversion of subject-verb can not be detected with a unigram model, because the position of each word is not considered. Therefore, yes/no questions with an inversion of subject-verb are removed from the corpus. Our yes/no question class thus contains the third group of yes/no question only.

Order

This DA class is based on *action motivators* class as described in Section 2.3. It contains utterances that imply for the listener to perform an action. The function of orders is similar to that of *commands*, but the form may differ. We consider that only utterances with an imperative form can be considered as orders. Incitative questions are thus excluded from this DA class. The written form finish usually by the "!" mark. Examples are: "Udělej to rychle!" (*Do it quickly!*) or "N'oublis pas ce livre demain!" (*Don't forgot this book tomorrow!*).

Interruptions

This DA class, described in Section 2.3.9, is divided into two sub-classes in our application: *interruption-begin* and *interruption-end*.

The reason for this division is that the respective lexical structures of both sub-classes are different and it may be possible to model them with two different models. This property

is not used in this work, but we hope to investigate it in a future work.

The **Interruption-begin** tag marks the first part of interrupted utterances. When an interrupted speaker continues his speech, then the last part of his utterance, after the interruption, is marked as **Interruption-end**.

The interrupted utterance "Je voudrais ... **hum** ... rester anonyme." (*I would like ...* **huh** *... stay anonymous.*) is thus labeled as: interruption-begin, backchannel, interruption-end. The instants of interruptions are market by a "..." symbol.

Indecipherables

We divided this DA class, which is described in Section 2.3.9, into two sub-classes: *indecipherable* and *indecipherable without prosody*. This is mainly due to technical and historical reasons. The first version of our system did not include prosody and could not thus classify correctly the DAs in this class (e.g. questions and statements with a similar grammatical form). We then removed this DA class from the corpus for the first version of our semi-automatic labeling system, and added it back in the last version.

5.4.5   Initial DA Tag-set

Table 5.1 summarizes our initial tag-set for the ESTER corpus. This DA tag-set is composed of 21 dialogue acts.

5.4.6   Reduction of the Initial Tag-set

Several DA classes of this initial DA tag-set occur only very rarely, which makes them difficult to model. Furthermore, several other DAs are not very important for our application. Therefore, the initial DA tag-set is reduced into a few broad classes. This is realized by removing the DA classes that are not needed by our application, and by grouping together DA classes that do not occur enough times.

The final grouped DA tag-set used for semi-automatic labeling is shown in the first part of Table 5.2. It is based on the reduced SWBD-DAMSL tag-set shown in Table 2.3.

5.4.7   DA Label Tool

Two corpus need to be manually labeled: a first small corpus is needed to initialize the semi-supervised training algorithm, and another one is required to test our approaches.

To achieve this task, we have developed the *DA Label* software, which is a tool dedicated to manual corpus labeling with DAs. DA Label is a system independent software developed in the java programming language. This tool contains a graphical user interface, which is controlled by a combination of keyboard and mouse. The mouse is used to select a dialogue

---

[1]There are actually several types of other questions. For instance, the *Or question* is used when there is at least two possible answers or options to choose from.

DA Type	Tag	Example	Translation
Statement non-opinion	sd	J'ai quatorze ans.	I am fourteen.
Statement opinion	sv	Je pense que c'est bien.	I think it is right.
Yes/No question	qy	Est-ce que tu aimes Eve?	Do you love Eve?
Wh-question	qw	Quelle heure est-il?	What time is it?
Others questions[1]	qo	Pensez-vous ça ou pas?	Are you think that or not?
Order	e	Fermez la porte!	Close the door!
Conventional-opening	o	Bonjour!	Hello!
Conventional-closing	c	Au revoir.	Good bye.
Accept	aa	Oui, volontiers!	Yes, with pleasure!
Reject	n	Non, pas du tout!	No, not at all!
Backchannel	b	Eh-hum	Uh-huh
Floor holder (hesitation)	h	Euh-euh	Uh-uh
Thanks	t	Merci beaucoup!	Many thanks!
Interuption-begin	ib	Je pense que ...	I think that ...
Interuption-end	ie	... sont satisfaisant.	... are satisfactory.
Radio Info	g	France Inter	France Inter
Person speaking	p	Pascal Dervieux	Pascal Dervieux
Dialogue subject	d	sport football	sport football
Indecipherable	z	<mumbled, muffled, ...>	
Indecipherable without prosody	zz	On peut commencer(./?)	We can start(./?)
Other	v	<examples that do not belong to the previous classes>	

Table 5.1: 21 dialogue acts from the French ESTER corpus with corresponding examples.

act to label, while predefined shortcut keys are used to add the corresponding DA label marks (one mark at the beginning, one mark at the end) into the *trs* files.

DA Label operates in two view modes. The default mode shows dialogues without XML tags and the advanced mode shows also the XML tags of the *trs* file. The screen of the DA Label in the advanced mode is shown in Figure 5.2.

The user manual is available at `http://home.zcu.cz/~pkral/manual.doc` and the free version of the DA Label software (limited functionality) in the zip archive at the `http://home.zcu.cz/~pkral/dalabel.zip`.

A dedicated window of the graphical interface displays the current line of the transcription file, the eventual error messages and the application help. The help contains the list of DAs along with the corresponding shortcut keys and DA labels, and all the other functions of the tool, such as the "Undo" and "Change display" functions. Figure 5.3 lists a section of the transcription file labeled with some DAs.

No.	Clustered DA Type	Tag	DA Types	Tag
Dialogue acts used for semi-automatic labeling				
1.	Statement	gs	Statement non-opinion	sd
			Statement opinion	sv
2.	Yes/No question	qy	Yes/No question	qy
3.	Other question	gq	Wh-question	qw
			Other question	qo
4.	Dialogue delimitation	goc	Conventional-opening	o
			Conventional-closing	c
5.	Accept	ga	Accept	aa
			Backchannel	b
6.	Floor holder	h	Floor holder	h
7.	Radio specific DA	gg	Radio Info	g
			Person speaking	p
			Dialogue subject, domain	d
Remaining dialogue acts				
8.	Order	e	Order	e
9.	Reject	n	Reject	n
10.	Thanks	t	Thanks	t
11.	Interuption	gi	Interuption-begin	ib
			Interuption-end	ie
12.	Other	v	Other	v
13.	Indecipherable	gz	Indecipherable without prosody	zz
			Indecipherable	z

Table 5.2: 13 clustered dialogue acts used in the French ESTER corpus: the first 7 DAs are used for semi-automatic labeling, the other DAs are not used.

5.4.8 Initial Corpus Creation Process

Two disjoint subsets, composed of 12 radio programs each, are first selected randomly from all radio sessions. The first subset is the small initial training corpus, while the second one is used for testing.

The ESTER corpus contains both dialogues and monologues. Monologues are not very interesting, because they are mainly composed of statements, and they contain very few other DAs. Dialogues are more important, because they contain a greater variety of DAs.

In order to balance the number of different DAs, we have first to remove most of the monologue sections so that mostly dialogues are manually labeled. We have developed a fully automatic algorithm to select these sections only. This algorithm is based on the two following constraints: (1) the speaker identity must change enough time during the session, (2) any segment with only one speaker must be shorter than some predefined time threshold. This value is set experimentally to 12 seconds.

This algorithm is used to select speech segments that are likely to contain dialogues from the speech training subset. These segments are labeled manually using the DA Label tool. The remaining segments are discarded. The testing corpus is created fully manually.
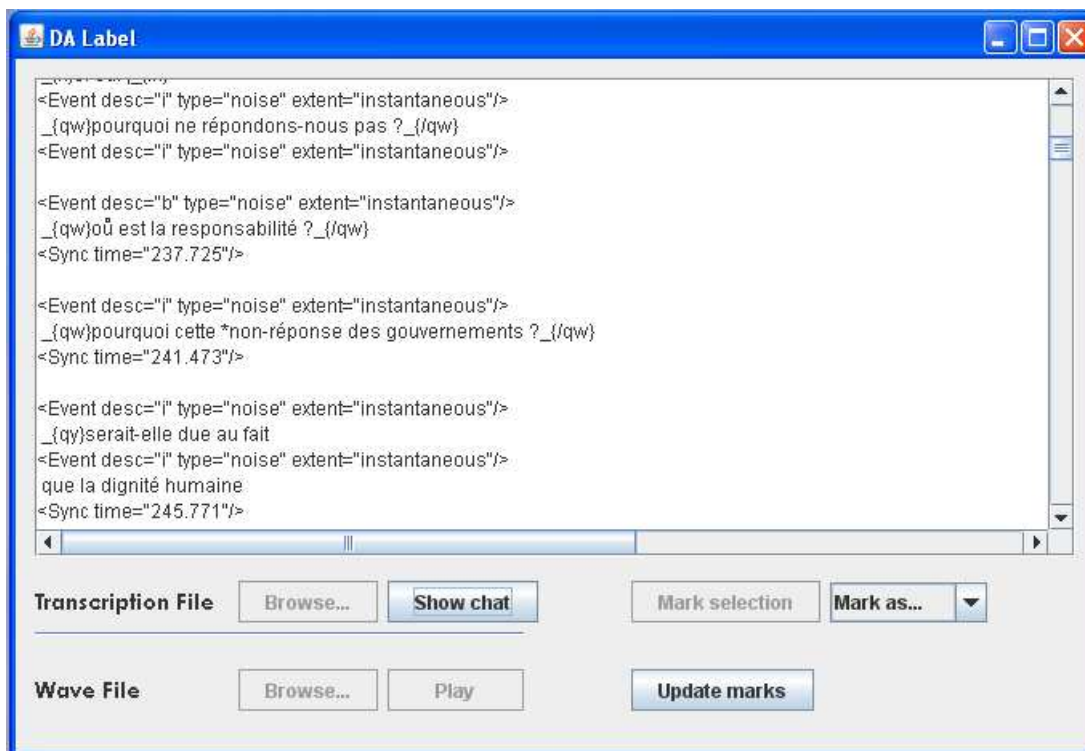
Figure 5.2: Example of DA Label tool screen.

Table 5.3 shows the numbers of DAs in the initial training and testing corpora after the manual labeling task. Note that the prefix "g" in the DA class names is removed to simplify notations.

No.	Clustered DA Type	Tag	Training	Testing
1.	Statements	s	251	609
2.	Yes/No questions	qy	24	27
3.	Other questions	q	39	72
4.	Dialog delimitations	oc	55	23
5.	Accepts	a	44	34
6.	Floor holders	h	46	71
7.	Radio specific DAs	g	130	93
Tot.	All DAs		589	929

Table 5.3: Structure of the manually created corpora for semi-automatic labeling.

This initial manual corpus can not be used to train statistical models, because it is not large enough to reliably estimate the model parameters, apart from the Statements and Radio specific DAs.

Several solutions exist. We describe and discuss next the three most interesting ones:

- Randomly select more radio programs and label them manually.

<Turn speaker="spk26" startTime="1036.08" endTime="1039.359" fidelity="high" channel="studio">
<Sync time="1036.08"/>
{o} bonjour Patricia ! {/o} {sd} je vous parlais la semaine dernière, Patricia, d'un spectacle d'Alfredo Arriaz {/sd}
</Turn>
<Turn speaker="spk2 spk26" startTime="1039.359" endTime="1042.022" fidelity="high" channel="studio">
<Sync time="1039.359"/>
<Who nb="1"/>
{aa} oui ... {/aa}
<Who nb="2"/>
{qy} vous vous souvenez, à Bobigny qui s'appelait Aimer sa mère ?{/qy}
</Turn>
<Turn speaker="spk26 spk2" startTime="1042.022" endTime="1046.446" fidelity="high" channel="studio">
<Sync time="1042.022"/>
<Who nb="1"/>
{qy} vous vous souvenez pas du tout ?{/qy} {sd} c'est dramatique {/sd}
<Event desc="rire" type="noise" extent="instantaneous"/>
<Who nb="2"/>
{sd} Gérard Zenoni ! {/sd}
<Who nb="1"/>
{t} merci ...{/t}
</Turn>

Figure 5.3: Example of dialogue with corresponding DA labels and XML tags in the *trs* file (Transcriber format): the XML tags are identified by the "<" and by the "/ >" signs, DA labels are represented by the sign "{}" at the beginning of the DA and by the sign "{/}" at its end.

- Use another corpus (or part of it) that is already annotated with DAs.
- Define and use rules, based on the general characteristics of the French language, to automatically label some DAs from the ESTER corpus.

Manual Labeling

The first solution provides the best labeling quality amongst the three options proposed. But it requires a lot of work and a lot of time, and it can not be the only option, because of its cost in terms of human efforts.

Use of Another Corpora

We have at our disposal two other small dialogue corpora:

MEDIA contains textual transcripts of phone dialogues in the reservation domain. One part of this corpus is labeled with seven DAs: statement, query[1], accept, reject, open dialogue and close dialogue. It contains about one thousand of labeled DAs.

ECOLE MASSY is a corpus of the VALORIA laboratory of South-Brittany University. It contains 31 simulated human-human oral dialogues (audio files with transcription) in the tourist information domain. There are 20 speakers: 19 children and 1 teacher. The corpus contains 5300 words, which represents 45 min of audio recording. It is labeled with two DA tags only: statements and questions.

The main advantage of this method is its low cost and efficiency. However, it has some limitations because of the incompatibility of the DA tag-sets used in both cases. Another problem is that after a small manual analysis of the labeled DAs, some mistakes in the DA classes have been found. Moreover, the dialogue domains are different, and the sentence structure and the words used are thus different. For instance, reservation sentences are usually much shorter than in the radio domain. Also, the vocabulary used in the reservation domain is much more constrained, which is not the case for broadcast news. The last issue is that the corresponding speech files for the MEDIA corpus are sometimes missing.

Lexical Rules

A small set of lexical rules is defined manually, based on general properties of the French language. The rules are defined by keywords and their corresponding position in the utterance. Examples of rules are:

- Every utterance starting with "est-ce que" is a *yes/no question.*

- Every utterance starting with a wh-word (such as "comment", "combien", ..) is a *wh-question.*

The complete list of rules is available at `http://home.zcu.cz/~pkral/rules.txt`.

The main advantage of this method is its quickness. Another positive aspect is that the rules are defined with respect to the chosen DA tag-set, and the resulting labeling is thus totally compliant with the original labeling.

On the other hand, these rules do encode only a very small portion of possible utterances of each DA, and can not capture the whole variety of the DAs extension. Furthermore, spontaneous speech may sometimes break these rules, which then produce erroneous labels. This is the main drawback of this fully automatic method. This issue may be partly tackled by manually verifying some of the automatically generated DAs, which is not as much time consuming as it is to label utterances by hand from scratch.

---

[1]This DA tag corresponds to the tag *question* from DAMSL.

Labeling using Rules

We have chosen the third proposed solution in our case, i.e. to automatically label the remaining training corpus with lexical rules.

Table 5.4 shows the composition of our initial corpora. The 1652 dialogue acts in the initial training corpus are thus composed of 589 DAs labeled manually plus 1063 DAs labeled automatically using rules.

No.	Clustered DA Type	Tag	Training			Testing
			Man.	Rules	Init.	Manual
1.	Statements	s	251	0	251	609
2.	Yes/No questions	qy	24	0	24	27
3.	Other questions	q	39	488	527	72
4.	Dialog delimitations	oc	55	411	446	23
5.	Accepts	a	44	21	65	34
6.	Floor holders	h	46	102	148	71
7.	Radio specific DAs	g	130	61	191	93
Tot.	All DAs		589	1063	1652	929

Table 5.4: Structure of the initial corpus for semi-automatic labeling created both manually and with rules.

We did not found any rule for the statement class, because its structure and content are too much variable. Another surprising fact is that, in our training corpus, no question with the interrogative form "est-ce que" at the beginning of the utterance has been found. This may be due to the fact that this form of yes/no question is more likely used in spontaneous speech than in more structured speech like broadcast news. Other questions characterized by the subject-verb inversion are removed from the corpus, and those characterized solely by prosody are considered as statements (c.f. Section 5.4.4).

The unlabeled part of corpus is composed of 5230 utterances.

5.5 Semi-automatic Labeling of Dialogue Acts with Confidence Measure

We respectively describe next the dialogue act models, the semi-supervised training algorithm and the proposed confidence measures, which role is to filter out erroneous training examples.

5.5.1 Dialogue Act Modeling

Each dialogue act is represented by a unique state in the ergodic HMM shown in Figure 5.4. Each state computes the observation log-likelihood $P(w_i|C)$ in the unigram model

described in Equation 5.3.

$$P(w_1, \cdots, w_T | C) = \prod_{i=1}^{T} P(w_i | C) \qquad (5.3)$$

where C encodes the dialogue act class and w_i represents the ith word of the current utterance.

Transitions between states encode transition probabilities between subsequent dialogue acts. In the following experiments, these transition probabilities are not trained, but are rather set manually, with the same values for every state: the loop probability models the average duration of any DA of the training corpus, while the out-going transitions are set equiprobable for all destination states.
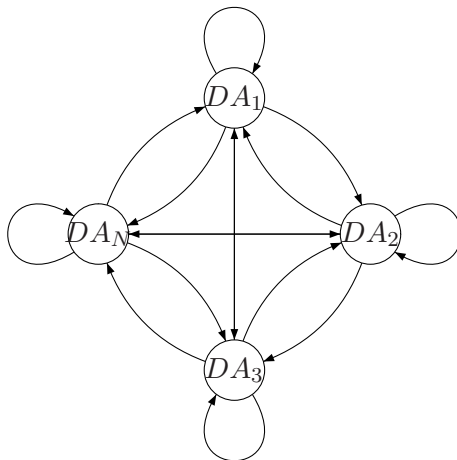
Figure 5.4: Dialogue act model: each node of the ergodic HMM represents one DA class.

Unlike our previous works in automatic DA recognition, prosodic information is not included in the feature vector: DA models exploit lexical features only. This choice has been made to first test our semi-supervised training procedure with a low-dimensional feature space, and because our previous experiments have shown that the contribution of prosody is anyway quite limited compared to lexical information. But prosody shall be considered in a future work. Furthermore, because of the small size of the initial corpus, only unigram statistics are computed. Our expectation is that once a larger part of the corpus has been semi-automatically labeled, this simplified framework could be advantageously replaced by more complex models, with prosodic features and longer temporal dependencies for example. But we investigate next the most critical part of the corpus creation process, which is likely to be just after initialization.

5.5.2 Semi-supervised Training

The structure of our initial corpus is summarized in Table 5.4: it is composed of a small part labeled manually, another part labeled automatically with rules, and a third part without any labels, which contains 5230 utterances. On this unlabeled part of the corpus,

we assume that the labels (the DA classes) are instances of an hidden random variable C. This variable is estimated by the classical Expectation Maximization algorithm, as follows:

1. Initialization: let Ω be the whole training corpus, and $\mathcal{D} \subset \Omega$ the small labeled training corpus; Initially, at $t = 0$, a classifier $\mathcal{C}_0$ is trained on $\mathcal{D}$.

2. The DAs of the unlabeled corpus $\Omega - \mathcal{D}$ are inferred (and segmented) by the current classifier $\mathcal{C}_t$.

3. The classifier $\mathcal{C}_{t+1}$ is re-trained on Ω.

4. The procedure is iterated from step 2 until a given number of iterations is reached.

In this algorithm, all DAs are classified and are further used to train a new classifier at the next iteration of the EM algorithm. However, the classifier is not perfect and make errors that impair the next training phase. This algorithm is also highly sensitive to the quality of the initial training corpus, especially with regard to its coverage property. There are two principal problems to solve:

1. How to select the "correct" labels for the next training iteration?

2. How to eventually select the few ambiguous examples that can be labeled manually?

The solution of the first challenge is the use of confidence measures, which give the probability that a recognized DA is correct or not. Any example is included in the next training corpus if and only if this probability is greater than a given threshold. The proposed algorithm based on the EM procedure is summarized next:

1. Initialization: let Ω be the whole training corpus, and $\mathcal{D}_0 \subset \Omega$ the small labeled training corpus; let $t = 0$.

2. The classifier $\mathcal{C}_t$ is trained on $\mathcal{D}_t$.

3. The DAs of the unlabeled corpus $\Omega - \mathcal{D}_t$ are inferred (and segmented) by the current classifier $\mathcal{C}_t$.

4. For each recognized DA, a confidence measure is computed to assess its reliability; let $\mathcal{M}_t$ be the most reliable DAs.

5. The most reliable examples are included into the training corpus: $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{M}_t$.

6. t is incremented, and the procedure is iterated from step 2 until a given number of iterations is reached.

The second issue, which concerns the selection of some DAs to be labeled manually, can also be solved with a confidence measure. The idea is to select only the DAs that bring the maximum information into the labeled corpus, which may correspond for instance to the most ambiguous DAs. This approach is known as active learning, and it is one of the perspectives of my work.

5.5.3   Dialogue Act Recognition

The performance of the classifier is evaluated at each iteration on the test corpus, which has been manually segmented and labeled. Recognition is realized with the ergodic HMM of Figure 5.4 and the Viterbi algorithm, which outputs both the DA labels and their temporal limits. The recognition rate is computed for each word by comparing the recognized and correct labels.

5.5.4   Confidence Measure

Like many confidence measures used in speech recognition [81], our first confidence measure for DA recognition is an estimate of the *a posteriori* class probability. The output of our lexical classifier is $P(W|C)$, where C is the dialogue act class and W is the words sequence in the DA. The likelihoods $P(W|C)$ are normalized to compute the *a posteriori* class probabilities:

$$P(C|W) = \frac{P(W|C).P(C)}{\sum_{D \in \mathcal{DA}} P(W|D).P(D)} \tag{5.4}$$

$\mathcal{DA}$ is the set of all DAs and $P(C)$ is the *prior* probability of the DA class C.

In the first version of our training algorithm, called *maximum a posteriori probability* method, only the DAs $\hat{C}$ so that

$$\hat{C} = \arg\max_C (P(C|W))$$

and

$$P(\hat{C}|W) > T$$

are included into the training corpus.

In the second version, called *a posteriori probability difference* method, the difference between the *best* hypothesis and the *second best* one is computed by the following equation:

$$\Delta P \;=\; P(\hat{C}|W) - \max_{C \neq \hat{C}}(P(C|W)) \tag{5.5}$$

Only the DAs with $\Delta P > T$ are included into the training corpus. This second approach aims at identifying the DAs that "dominate" all the other candidates, which is not always well captured by the first measure.

T is in both cases an acceptation threshold and its optimal value is found experimentally.

5.6   Experiments

In the following experiments, the unigram probabilities $P(w_i|C)$ with less than 6 examples in the training corpus are smoothed to the class-independent backoff *prior* $P(w_i)$. Furthermore, all DA *priors* are set equiprobable, because the training corpus is partly generated

from hand-crafted rules that bias the estimates of these *priors*.

5.6.1 Maximum *a Posteriori* Probability

Figure 5.5 plots the DA recognition rate on the manually labeled test corpus, with the Maximum *a posteriori* probability method, in function of the number of EM iterations and for different values of T. The results obtained without any confidence measure (or equivalently for $T = 0$) are also shown with the label "EM". We can note that the performance of this EM-only curve degrades, which justifies the use of confidence measures to filter out incorrectly recognized DAs.

Figure 5.5: Performance of the maximum *a posteriori* probability method: the X-axis represents the number of EM iterations and the Y-axis plots the DA recognition rate.

After three iterations, the recognition rate tends to stabilize, with a maximum of 80% for threshold 0.999 and at the third iteration. The improvement due to our semi-supervised training algorithm represents a decrease of 30% of the recognition errors. The evolution of the size of the training corpus is shown in Figure 5.6.

Table 5.5 shows the recognition rate per DA at different iterations with $T = 0.999$. One can observe that most of the individual DA rates increase. Only the score of yes/no questions is decreasing, which is probably due to the lack of training data for this class in the initial manual corpus.

The confusions between the DA classes at the best recognition rate (third iteration with $T = 0.999$) are shown in Table 5.6. Most of these errors occur with the *statement* class. This can be explained by the lexical characteristics of each class. Every DA class except statements contains at least one word that is strongly correlated to the DA class. Moreover, their vocabularies are usually less variable and more discriminant than for the *statement* class. For example, a wh-word is characteristic of *wh-questions* and rarely occurs in other DA classes, the word "Bonjour" (*Good-day*) contains *conventional-opening* DA class only,
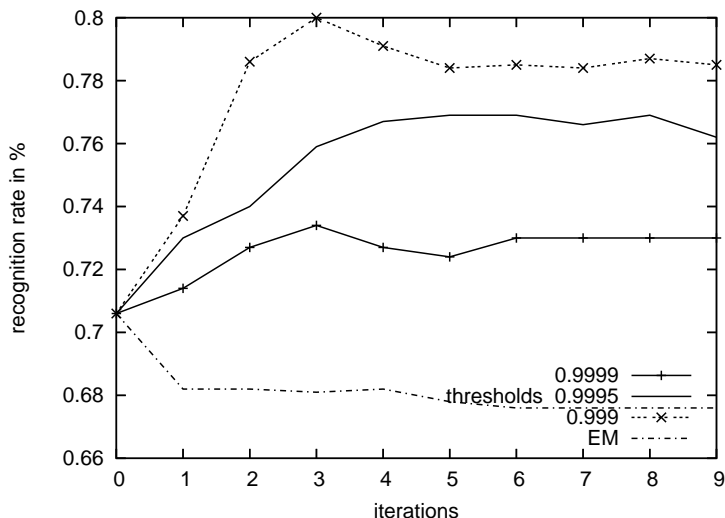
Figure 5.6: Performance of the maximum *a posteriori* probability method: the X-axis represents the number of EM iterations and the Y-axis plots the DA corpus size.

Iter.	Recognition rate in [%]							
	s	qy	q	oc	a	h	g	glob.
0	72.4	70.3	62.9	66.1	51.4	100	41.6	**70.6**
1	76.4	58.6	62.9	66.1	51.4	100	42.7	**73.7**
2	81.8	58.0	62.5	66.1	65.3	100	45.7	**78.6**
3	83.8	52.3	65.5	66.1	65.3	100	41.0	**80.0**
4	82.6	51.1	66.5	66.1	62.5	100	43.1	**79.1**
5	81.9	47.1	68.2	66.1	62.5	100	43.1	**78.4**
6	81.8	51.1	68.2	66.1	62.5	100	43.1	**78.5**
7	81.8	46.8	68.2	66.1	62.5	100	43.1	**78.4**
8	82.2	46.8	68.8	66.1	62.5	100	43.1	**78.7**
9	81.9	46.8	68.8	66.1	62.5	100	43.1	**78.5**

Table 5.5: Performance of the maximum *a posteriori* probability method: dialogue act recognition rate in % at different iterations with probability threshold 0.999.

etc. The *radio specific DAs* class is the class that has the worst accuracy, which is probably due to its lexical characteristics that are largely common with the *statement* class.

5.6.2 *A posteriori* Probability Difference

Figure 5.7 shows the DA recognition rate in function of the number of EM iterations. The corresponding corpus sizes are shown in Figure 5.8.

Table 5.7 shows the recognition rate per DA at different iterations with $T = 0.9995$. Like in the previous experiment, most of the DA recognition rates increase. Only the score of yes/no questions does not increase, which is probably due to the lack of training data for

DA	Recognized DA class (in [%])						
	s	qy	q	oc	a	h	g
s	**83.8**	2.7	10.6	0.2	0.7	0.0	2.0
qy	39.9	**52.3**	4.2	0.0	0.6	0.0	3.0
q	31.8	2.7	**65.5**	0.0	0.0	0.0	0.0
oc	33.9	0.0	0.0	**66.1**	0.0	0.0	0.0
a	20.8	0.0	9.7	4.2	**65.3**	0.0	0.0
h	0.0	0.0	0.0	0.0	0.0	**100**	0.0
g	52.9	3.0	1.9	1.3	0.0	0.0	**41.0**

Table 5.6: Confusion matrix of the maximum *a posteriori* probability method for the best DA recognition rate (third iteration and probability threshold 0.999).

Figure 5.7: Performance of the *a posteriori* probability difference method: the X-axis represents the number of EM iterations and the Y-axis plots the DA recognition rate.

this class in the initial manual corpus. The results stabilize after the seventh iteration, with a maximum of 78% for threshold 0.9995: this represents a decrease of 27% of the recognition errors.

The confusions between the DA classes at the best recognition rate (seventh iteration with $T = 0.9995$) are shown in Table 5.8. The confusion matrix is almost the same as in the previous experiment (see Table 5.6).

The difference between the Maximum *a posteriori* probability and the *A posteriori* probability difference methods is quite small.
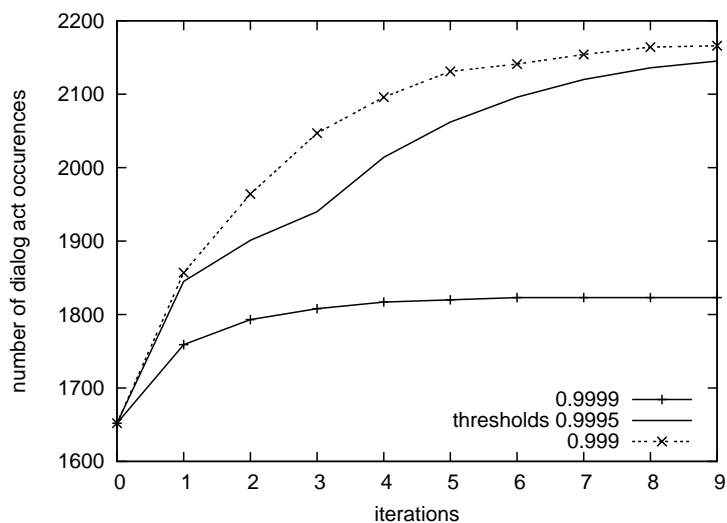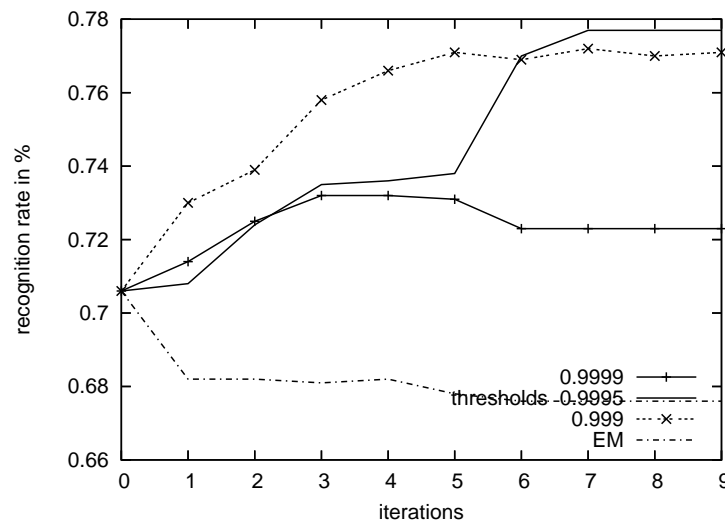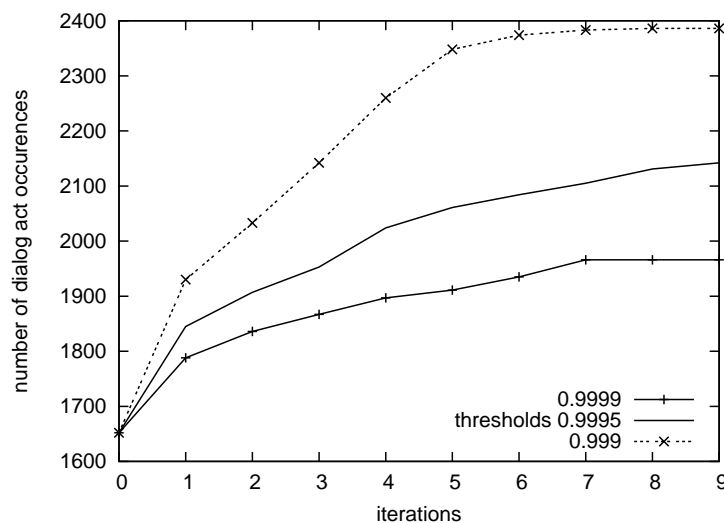
Figure 5.8: Performance of the maximum *a posteriori* probability method: the X-axis represents the number of EM iterations and the Y-axis plots the DA corpus size.

Iter.	Recognition rate in [%]							
	s	qy	q	oc	a	h	g	glob.
0	72.4	70.3	62.9	66.1	51.4	100.0	41.6	**70.6**
1	72.8	64.3	61.4	66.1	51.4	100.0	43.8	**70.8**
2	74.4	64.3	62.9	66.1	51.4	100.0	47.6	**72.4**
3	75.5	64.3	64.3	66.1	65.3	100.0	46.7	**73.5**
4	75.9	65.8	60.0	66.1	65.3	100.0	46.7	**73.6**
5	75.9	69.1	61.0	66.1	65.3	100.0	46.7	**73.8**
6	80.1	59.2	60.0	66.1	65.3	100.0	45.9	**77.0**
7	80.8	57.1	62.0	66.1	65.3	100.0	45.0	**77.7**
8	80.8	57.1	62.0	66.1	65.3	100.0	45.0	**77.7**
9	80.8	57.1	62.0	66.1	65.3	100.0	45.0	**77.7**

Table 5.7: Performance of the *a posteriori* probability difference method: dialogue act recognition rate in % at different iterations with probability threshold 0.9995.

5.7   Main Contributions

The most important contributions of my research described in this chapter are summarized below:

- Proposition of a new DA tag-set for the ESTER radio corpus based on DAMSL and MRDA projects.

- Proposition and implementation of two confidence measures methods: Maximum *a posteriori* probability and *A posteriori* probability difference.

- Use of these confidence measures to improve the performance of the EM-based semi-supervised training in the dialogue act recognition domain.

	Recognized DA class (in [%])						
DA	s	qy	q	oc	a	h	g
s	**80.8**	7.2	9.5	0.2	0.6	0.0	1.7
qy	33.3	**57.1**	6.0	0.0	0.6	0.0	3.0
q	29.4	7.3	**62.0**	0.0	1.4	0.0	0.0
oc	29.0	4.8	0.0	**66.1**	0.0	0.0	0.0
a	20.8	0.0	9.7	4.2	**65.3**	0.0	0.0
h	0.0	0.0	0.0	0.0	0.0	**100**	0.0
g	46.7	4.7	2.3	1.3	0.0	0.0	**45.0**

Table 5.8: Confusion matrix of the *a posteriori* probability difference method for the best DA recognition rate (seventh iteration and probability threshold 0.9995).

- Semi-automatic creation of a new French DA corpus based on the ESTER corpus.

Some other contributions of my work described next are summarized as follows:

- Design and implementation of the DA Label, a user friendly tool for manual DA labeling of corpora.
- Proposition and implementation of an automatic method for DA labeling of corpora based on pre-defined rules.
- Implementation of the EM algorithm for semi-automatic labeling in the dialogue act domain.

5.8   Conclusions

In this chapter we have proposed a new DA tag-set that is based on the DAMSL and MRDA systems and that is adapted to the broadcast news domain. This tag-set is used to label the French ESTER corpus.

The main contribution of the work described in this chapter is to instantiate the general EM procedure to the task of creating semi-supervised corpora labeled with different sets of dialogue acts and in different languages at a low cost. We show that confidence measures are required to filter out incorrect examples, and we evaluate two such measures on this task.

The recognition score as well as the size of our DA corpus is increasing during this iterative process. However, the corpus size does not increase very much (see for example Figure 5.6) and a large part of the training data is never considered as correctly recognized by the confidence measure, and is thus never included into our labeled corpus. This is probably due to the large variety of the vocabulary used in broadcast news: the initial manually labeled corpus is too small and most of the words that are in the unlabeled corpus have never been seen and are not accurately modeled.

The solution might be to use the *Active learning* approach, which selects a few examples in the unlabeled part at each iteration, and asks the user to manually label them. These examples are chosen at each step in order to increase the models coverage and accuracy.

We have also described how our dialogue act recognition system, which was previously developed for a Czech reservation application, can be retrained and successfully adapted to a new language (French), a new type of corpus (broadcast news) and a different set of dialogue acts.

The principal contribution described in this chapter was presented at the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2007) [66].

Chapter 6

Conclusions and Perspectives

This thesis deals with automatic dialogue act recognition in Czech and in French.

The information given by the sequence of dialogue acts is an important clue to understand spontaneous dialogues, and numerous applications may benefit from this knowledge. During the development of the DA recognition system, we had two particular applications in mind: the first one is a dialogue system that detects the communicative intention of the speaker, e.g. whether the user gives a command that the system must process immediately, or whether the user is rather talking to his friend and does not expect any clear reaction from the system. The second target application is to augment the visual feedback of a talking head by animating it with respect to the current dialogue act (for instance, raising eyebrows when asking questions).

State-of-the-art DA recognition approaches are usually based on n-gram models, which only model the *local* structure of utterances. An important aspect of our work is the different solutions that we have proposed and studied to further model the *global* sentence structure. This has been realized by considering additional information about the words position in the sentence. We have thus proposed three approaches to model this information: the first one, the *multiscale position* approach, exploits a description of the sentence at several levels and smoothes the probabilities across these levels. The second one, the *non-linear merging* method, models the dependency between the words in the sentence and their position with an MLP. The third one, the *best position* approach, exploits the Bayesian framework and assumes conditional independence between the words and their position to infer the probability of the dialogue act.

All the proposed approaches have been described and compared in terms of their respective theoretical advantages and drawbacks, and also experimentally on the Czech corpus. It has been demonstrated that the global position of the words in the sentences is an important information that improves the 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.

Another common issue in DA recognition systems concerns the lack of training data, which limits the complexity of the model. We have also studied this problem and proposed some

solutions. In the first one, a *clustered unigram model* is developed, which clusters the words in the sentences into several groups by maximizing the mutual information between two neighbor word classes. We have shown that this method is especially efficient when the DA corpus is small.

We have also analyzed the importance of prosody, represented by the fundamental frequency and the energy in the DA recognition task. We have shown that prosody alone is not sufficient to perform DA recognition, but that it nevertheless brings useful information that may solve ambiguous cases with the lexicon alone. In particular, we have shown that prosody is especially useful for yes/no questions detection.

All lexical, prosodic and sentence structure features are finally combined. We have thus studied several classifier combination methods and compared all of them theoretically and experimentally. We have shown that the combination of these different knowledge sources improves the recognition score over each individual method.

The proposed approaches have been evaluated in two cases: when the manual word transcription is used and when these word sequences are unknown and estimated by a speech recognizer, which is the most common case in a real application. We have shown that the recognition accuracies of both manual and automatic systems are comparable, and that the use of an ASR system only slightly degrades the performances for our corpus and task.

Another main contribution of this work is the development and evaluation of a semi-automatic DA labeling method. This kind of approach is especially important in the DA recognition domain, where the DA tag-set often changes from one target application to another, which would otherwise require a lot of efforts to manually label many different corpora with a variety of different tags. In this context, we have proposed a new DA tag-set based on the DAMSL and MRDA systems for the French broadcast news ESTER corpus. We have also proposed a semi-supervised labeling approach based on the EM algorithm and confidence measures to label this corpus at a low cost. Two confidence measures, namely the *maximum a posteriori probability* and the *a posteriori probability difference*, have been developed and tested in order to improve the performance of the EM algorithm. We have shown that, in the proposed experimental setup, the use of confidence measures to filter out incorrectly recognized DAs is required to obtain satisfactory results. The study of dialogue act recognition approaches in two different languages (Czech and French) and two applications (reservation system and broadcast news) is also an interesting and original contribution over the state-of-the-art.

This work can be improved in several aspects. First, although global words positions have been modeled, local statistical grammars have not been largely exploited, mainly because of the lack of training data. However, these grammars are known to bring relevant information, and it would be quite advantageous to further combine the proposed global models 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 in 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 it is also often dependent on the target application.

Regarding the combination of DA recognition methods, we have shown that several such combinations improve the recognition accuracy, but we have not tested all of them. This comparison could be completed in a future work, but we do not expect a large difference with the results that have been reported in this thesis. Finally, other promising perspectives of this work concern semi-supervised corpus creation. This approach has still to be tested on the outputs of a real automatic speech recognizer system before it can be used to produce complete new corpora for DA recognition. Furthermore, an Active Learning approach should certainly be used in order to bypass the limits of the models trained on a limited initial corpus. More complex dialogue act models can also be used, for instance with prosody and dialogue grammars, as well as the development of better confidence measures and initial dialogue act rules.

List of Acronyms

ASR Automatic Speech Recognition

CART Classification and Regression Tree

DA Dialogue Act

DAMSL Dialogue Act Markup in Several Layers

DFT Discrete Fourier Transform

EM Expectation Maximization

F0 Fundamental Frequency

GMM Gaussian Mixture Model

HMM Hidden Markov Model

k-NN k-Nearest Neighbor

LM Language Model

MMI Maximization of Mutual Information

MBL Memory-Based Learning

MC Monte-Carlo

MLP Multi-Layer Perceptron

MoC Mean number of Conditions

MRDA Meeting Recorder DA

NN Neural Network

RMSE Root Mean Square Energy

SI Specificity Index

SWBD-DAMSL Switchboard DAMSL

SOM Self-Organizing Map

SVM Support Vector Machine

TBL Transformation-Based Learning

TSVM Transductive Support Vector Machine

WER Word Error Rate

Author's Publications

The following papers were published in conference proceedings:

1. P. Kral, C. Cerisara and J. Kleckova, **Importance of Prosody for Dialogue Acts Recognition**, *in SPECOM'07*, Moscow, Russia, October 2007.

2. P. Kral, C. Cerisara and J. Kleckova, **Confidence Measures for Semi-automatic Labeling of Dialog Acts**, *in ICASSP'07*, Honolulu, Hawaii, USA, April 2007, pp. 153-156.

3. P. Kral, J. Kleckova and C. Cerisara, **Automatic Dialog Acts Recognition based on Words Clusters**, *in WESPAC IX 2006*, Seoul, Korea, June 2006.

4. P. Kral, C. Cerisara and J. Kleckova, **Automatic Dialog Acts Recognition based on Sentence Structure**, *in ICASSP'06*, Toulouse, France, May 2006, pp. 61-64.

5. P. Kral, J. Kleckova, T. Pavelka and C. Cerisara, **Sentence Structure for Dialog Act recognition in Czech**, *in ICTTA'06*, Damascus, Syria, April 2006.

6. P. Kral, J. Kleckova and C. Cerisara, **Sentence Modality Recognition in French based on Prosody**, *in Enformatika*, Budapest, Hungary, October 2005, vol. 8, pp. 185-188, International Academy of Sciences.

7. P. Kral, C. Cerisara and J. Kleckova, **Combination of Classifiers for Automatic Recognition of Dialog Acts**, *in Interspeech'2005*, Lisboa, Portugal, September 2005, pp. 825-828, ISCA.

8. P. Kral, J. Kleckova and C. Cerisara, **Analysis of Importance of the prosodic Features for Automatic Sentence Modality Recognition in French in real Conditions**, *in WSEAS Int. Conf. on ELECTRONICS, CONTROL and SIGNAL PROCESSING (ICECS'04)*, Rethymno, Crete, Greece, October 2004, vol. 3, pp. 1820-1824.

9. P. Kral and J. Kleckova, **Speech Recognition and Animation of Talking Head**, *in IWSSIP'03*, Prague, Czech Republic, September 2003, pp. 122-124.

The following paper was published in the scientific journal:

1. P. Kral, C. Cerisara and J. Kleckova, **Lexical Structure for Dialogue Act Recognition**, *in Journal of Multimedia (JMM)*, ISSN : 1796-2048, Volume 2, Issue 3, June 2007, pp. 1-8.

Bibliography

[1] D. W. Aha, D. Kibler, and M. K. Albert. Instance-Based Learning Algorithms. *Machine Learning*, 6(1):37–66, January 1991.

[2] Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. Insights into the Dialogue Processing of VERBMOBIL. Technical Report 191, Saarbrücken, Germany, 1997.

[3] 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.

[4] T. Andernach. A Machine Learning Approach to the Classification of Dialogue Utterances. In *NeMLaP-2*, Ankara, Turkey, July 1996.

[5] T. Andernach, M. Poel, and E. Salomons. Finding Classes of Dialogue Utterances with Kohonen Networks. In *ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pages 85–94, Prague, Czech Republic, April 1997.

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

[7] Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online Choice of Active Learning Algorithms. *The Journal of Machine Learning Research*, 5:255–291, December 2004.

[8] E. G. Bard, C. Sotillo, A. H. Anderson, and M. M. Taylor. The DCIEM Map Task Corpus: Spontaneous Dialogue Under Sleep Deprivation and Drug Treatment. In *ICSLP'96*, volume 3, pages 1958–1961, Philadelphia, USA, 1996.

[9] R. Battiti and A. M. Colla. Democracy in Neural Nets: Voting Schemes for Classification. *Neural Networks*, 7(4):691–707, 1994.

[10] J. O. Berger. Statistical Decision Theory and Bayesian Analysis. *Springer-Verlag, New York*, 1985.

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

[12] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Workshop on Computational Learning Theory, COLT*, pages 92–100. Morgan Kaufmann Publishers, 1998.

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

[14] T. Bray, J. Paoli, and C. M. Sperberg-McQueenc. *Extensible Markup Language (XML) 1.0.* World Wide Web Consortium, http://www.w3.org/TR/1998/REC-xml-19980210, February 1998.

[15] L. Breiman. Stacked regression. Technical Report TR-367, University of California, Berkeley, CA, 1993.

[16] L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees.* Wadsworth and Brooks, Pacific Grove, CA, 1984.

[17] E. Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, Philadelphia, USA, 1993.

[18] H. Bunt. Dialogue Pragmatics and Context Specification. In *Bunt,H. and Black, W. (eds.), Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics. Amsterdam: Benjamins*, volume 1, pages 81–150, 2000.

[19] P. Chan and S. Stolfo. Meta-Learning for Multistrategy and Parallel Learning. In *Second Intl. Workshop on Multistrategy Learning*, pages 150–165, 1993.

[20] P. Chan and S. Stolfo. On the Accuracy of Meta-Learning for Scalable Data Mining. *Journal of Intelligent Information Systems*, 8(1):5–28, 1997.

[21] P. K. Chan and S. J. Stolfo. Experiments in Multistrategy Learning by Meta-Learning. In *2nd International Conference on Information and Knowledge Management*, pages 314–323, Washington, DC, 1993.

[22] P. K. Chan and S. J. Stolfo. Toward Parallel and Distributed Learning by Meta-Learning. In *Working Notes AAAI Work. Knowledge Discovery in Databases*, pages 227–240, 1993.

[23] C. Cieri, D. Graff, and M. Liberman. The TDT-2 Text and Speech Corpus. In *DARPA Broadcast News Workshop*, pages 57–60, 1999. (See also http://morph.ldc.upenn.edu/TDT).

[24] W. Cohen. Learning Trees and Rules with Set-valued Features. In *13th National Conference on Artificial Intelligence (AAAI-96)*, volume 1, pages 709–716, Portland, Oregon, 1996. AAAI Press.

[25] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active Learning with Statistical Models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press, 1995.

[26] R. O. Cornett. Cued Speech. In *American Annals of the Deaf*, volume 112, pages 3–13, 1967.

[27] IBM Corporation and Microsoft Corporation. *Multimedia Programming Interface and Data Specifications 1.0.* http://www.kk.iij4u.or.jp/ kondo/wave/mpidata.txt, August 1991.

[28] C. Cortes and V. N. Vapnik. Support Vector Network. *Machine Learning*, 20:273–297, 1995.

[29] M. Cottrell and J.C. Fort. Theoretical Aspects of the SOM Algorithm. *Neurocomputing*, 21:119–138, 1998.

[30] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. In *IEEE Trans. Inform. Theory*, pages 21–27.

[31] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew K. McCallum, Tom M. Mitchell, Kamal Nigam, and Seán Slattery. Learning to Extract Symbolic Knowledge from the World Wide Web. In *15th Conference of the American Association for Artificial Intelligence, AAAI-98*, pages 509–516, Madison, US, 1998. AAAI Press, Menlo Park, US.

[32] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg Memory-Based Learner. Technical report, Tilburg University, November 2003.

[33] R. Daneš. *Intonace a věta ve spisovné češtině.* Academia, Praha, 1957.

[34] Département Technologies de l'Information et de la Communication Action Technolangue. French ESTER Corpus. In *http://www.recherche.gouv.fr/technolangue/.*

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 1(39):1–38, 1977.

[36] R. Dhillon, Bhagat S., H. Carvey, and Shriberg E. Meeting Recorder Project: Dialog Act Labeling Guide. Technical Report TR-04-002, International Computer Science Institute, February 9 2004.

[37] P. T. Douglas, , and F. N. Jay. Speech Act Profiling: A Probabilistic Method for Analyzing Persistent Conversations and their Participants. In *37th Annual Hawaii International Conference on System Sciences (HICSS'04)*. IEEE, 2004.

[38] A. Elsner. Focus Detection with Additional Information of Phrase Boundaries and Sentence Mode. In *Eurospeech'97*, pages 227–230, Rhodes, Greece, September 1997.

[39] Astrid Schmidt-Nielsen et. all. Speech in Noisy Environments (SPINE) Training Audio. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000S87.

[40] M. Finke, M. Lapata, A. Lavie, L. Levin, L. Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. CLARITY: Inferring Discourse Structure from Speech. Technical Report SS-98-01, Applying Machine Learning to Discourse Processing, AAAI'98. CA: AAAI Press.

[41] P. N. Garner, S. R. Browning, R. K. Moore, and R. J. Russel. A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In *ICSLP'96*, volume 3, pages 1880–1883, Philadelphia, USA, 1996.

[42] H. Gezundhajt. La prosodie. In *http://www.linguistes.com/phonetique/prosodie.html*.

[43] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP'1992*, volume 1, pages 517–520, San Francisco, CA, USA, 23-26 March 1992.

[44] 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)*, pages 495–499, Saint-Petersburg, Russia, September 2004.

[45] Peter J. Green. On Use of the EM Algorithm for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society*, 52(3):443–452, 1990.

[46] H. Harb, L. Chen, and J.-Y. Auloge. Speech/Music/Silence and Gender Detection Algorithm. In *7th International Conference on Distributed Multimedia Systems (DMS'01)*, pages 257–262, Taipei, Taiwan, September 2001.

[47] J.-P. Haton, J.-M. Pierrel, G. Perennou, J. Caelen, and J.-L. Gauvain. *Reconnaissance automatique de la parole*. BORDAS, 1991.

[48] S. Haykin. *Neural Networks: a Comprehesive Foundation*. Prentice Hall, 2nd edition, 1999.

[49] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.

[50] W. Hess, A. Batliner, A. Kießling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt, and V. Strom. Prosodic Modules for Spech Recognition and Understanding in VERBMOBIL. In Yoshinori Sagisaka, Nick Campell, and Norio Higuchi, editors, *Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech*, pages 363–383. Springer-Verlag, New York, 1996.

[51] E. Ivanovic. Dialogue Act Tagging for Instant Messaging Chat Sessions. In *ACL Student Research Workshop*, pages 79–84, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

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

[53] G. Ji and J. Bilmes. Dialog Act Tagging Using Graphical Models. In *ICASSP'05*, volume 1, pages 33–36, Philadelphia, USA, March 2005.

[54] Gang Ji and Jeff Bilmes. Backoff Model Training using Partially Observed Data: Application to Dialog Act Tagging. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL'2006)*, pages 280–287, New York, USA, June 2006.

[55] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Douglas H. Fisher, editor, *14th International Conference on Machine Learning, ICML-97*, pages 143–151, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[56] T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *International Conference on Machine Learning (ICML)*, pages 200–209, Bled, Slovenia, June 1999.

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

[58] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Discourse Relations and Discourse Markers*, pages 114–120, Somerset, New Jersey, 1998. Association for Computational Linguistics.

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

[60] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

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

[62] T. Kohonen. Self-Organizing Maps. *Springer Series in Information Sciences*, 30, 1995.

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

[64] P. Král, C. Cerisara, and J. Klečková. Combination of Classifiers for Automatic Recognition of Dialog Acts. In *Interspeech'2005*, pages 825–828, Lisboa, Portugal, September 2005. ISCA.

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

[66] P. Král, C. Cerisara, and J. Klečková. Confidence Measures for Semi-automatic Labeling of Dialog Acts. In *ICASSP'07*, pages 153–156, Honolulu, Hawaii, USA, April 2007.

[67] P. Král, C. Cerisara, and J. Klečková. Importance of Prosody for Dialogue Acts Recognition. In *SPECOM'07*, Moscow, Russia, October 2007.

[68] P. Král, C. Cerisara, and J. Klečková. Lexical Structure for Dialogue Act Recognition. *Journal of Multimedia (JMM)*, 2(3):1–8, June 2007.

[69] P. Král and J. Klečková. Speech Recognition and Animation of Talking Head. In *IWSSIP'03*, pages 122–124, Prague, Czech Republic, September 2003.

[70] P. Král, J. Klečková, and C. Cerisara. Automatic Dialog Acts Recognition based on Words Clusters. In *WESPAC IX 2006*, Seoul, Korea, June 2006.

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

[72] L Kukolich and R. Lippmann. *LNKnet User's Guide*. MIT Lincoln Laboratory, February 2004.

[73] L. Lamel, J. Gauvain, and G. Adda. Unsupervised Acoustic Model Training. In *ICASSP'2002*, volume 1, pages 877–880, Orlando, Fl, USA, May 2002.

[74] P. Langlais. *Traitement de la prosodie en reconnaissance automatique de la parole*. PhD thesis, Univerité d'Avignon et des pays de Vaucluse, 1995.

[75] Steffen L. Lauritzen. The EM Algorithm for Graphical Association Models with Missing Data. *Computational Statistics & Data Analysis*, 19:191–201, 1995. Edited by S. P. Azen.

[76] P. Lendvai, A. van den Bosch, and Krahmer E. Machine Learning for Shallow Interpretation of User Utterances in Spoken Dialogue Systems. In *EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles Management*, pages 69–78, Budapest, Hungary, 2003.

[77] L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. Domain Specific Speech Acts for Spoken Language Translation. In *4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

[78] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets. In *TRECVID 2003 Workshop*, 2003.

[79] N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. In *IEEE Symposium on Foundations of Computer Science*, pages 256–261, 1989.

[80] Huan Liu, Hiroshi Motoda, and Lei Yu. A Selective Sampling Approach to Active Feature Selection. *Artificial Intelligence*, 159:49–74, May 2004.

[81] E. Lleida and R. C. Rose. Likelihood Ratio Decoding and Confidence Measures for Continuous Speech Recognition. In *ICSLP'96*, volume 1, pages 478–481, Philadelphia, USA, 1996.

[82] M. Loeve. *Probability Theory*. Van Nostrand, Princeton, N. J., 2nd edition, 1960.

[83] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. pages 500-528, MIT Press. Cambridge, MA, May 1999.

[84] P. Martin. L'intonation: Aspects linguistiques et reconnaissance de la parole. In *8th JEP*, volume II, pages 103–108, Aix-en-Provence, France, May 1977.

[85] P. Martin. Problèmes de neutralisation des marques prosodiques - application à la reconnaissance automatique. In *8th Journées d'Etude sur la Parole*, volume IV-2, pages 306–311, Aix-en-Provence, France, 1977.

[86] P. Martin. Mesure de la frquence fondamentale par intercorrlation avec une fonction peigne. In *7th JEP*, pages 221–232, Montréal, 1981.

[87] P. Martin. Comparison of Pitch Detection by Cepstrum and Spectral Comb Analysis. In *Int. Conf. Acoust., Speech, Signal Processing*, pages 180–183, 1982.

[88] P. Martin. Prosodic and Rhythmic Structures in French. In *Linguistics*, volume II, pages 925–949, 1987.

[89] 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*, pages 217–229, 1996.

[90] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. Dysfluency Annotation Stylebook for the Switchboard Corpus. Technical report, Linguistic Data Consortium, Ferbruary 1995. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps, Revised June 1995 by A. Taylor.

[91] Kasabov N. and S. Pang. Transductive Support Vector Machines and Applications in Bioinformatics for Promoter Recognition. *Neural Information Processing - Letters and Reviews*, 3(2):31–38, May 2004.

[92] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[93] A. M. Noll. Cepstrum Pitch Determination. *J. Acoust. Soc. Am.*, 2(41):293–309, 1967.

[94] D. O'Shaughnessy. Timing Patterns in Fluent and Disfluent Spontaneous Speech. In *Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 600–603, Detroit, 1995.

[95] Z. Palková. *Fonetika a fonologie čestiny*. Karolinum - publisher of Charles University in Prague, 1997.

[96] M.P. Perrone and L.N. Cooper. Learning from what's been Learned: Supervised Learning in Multi-Neural Network Systems. In *World Congress on Neural Networks*, volume III, pages 354–357. INNS Press, 1993.

[97] J. Psutka. *Komunikace s počítačem mluvenou řečí*. Academia, 1995.

[98] L. Rabiner and Juang. B. An Introduction to Hidden Markov Models. *ASSP Magazine*, 1986. IEEE 3(1).

[99] L. R. Rabinter, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. A Comparative Performance Study of Several Pitch Detection Algorithms. In *Int. Conf. Acoust., Speech, Signal Processing*, volume 24, pages 399–418, October 1967.

[100] N. Reithinger and M. Klesen. Dialogue Act Classification Using Language Models. In *EuroSpeech'97*, pages 2235–2238, Rhodes, Greece, September 1997.

[101] N. Reithinger and E. Maier. Utilizing Statistical Dialogue Act Processing in VERB-MOBIL. In *33rd annual meeting on Association for Computational Linguistics*, pages 116–121, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[102] K. Ries. HMM and Neural Network Based Speech Act Detection. In *ICASSP'99*, volume 3, pages 497–500, 1999.

[103] M. Romportl. *Studies in Phonetics*. Praha - The Hague, 1973.

[104] M. Rotaru. Dialog Act Tagging using Memory-Based Learning. Technical report, University of Pittsburgh, Spring 2002. Term Project in. Dialog Systems.

[105] N. Sabouret and J.P. Sansonnet. Querying Knowledge about Actions in the Semantic Web. In *citeseer.ist.psu.edu/560769.html*.

[106] K. Samuel, S. Carberry, and K. Vijay-Shanker. Dialogue Act Tagging with Transformation-Based Learning. In *17th international conference on Computational linguistics*, volume 2, pages 1150–1156, Montreal, Quebec, Canada, 10-14 August 1998. Association for Computational Linguistics, Morristown, NJ, USA.

[107] E. Sanchis and M. J. Castro. Dialogue Act Connectionist Detection in a Spoken Dialogue System. In *Second International Conference on Hybrid Intelligent Systems (HIS2002)*, pages 644–651, Santiago de Chile, Chile, 1-4 December 2002. IOS Press.

[108] M. R. Schroeder. Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurements. *J. Acoust. Soc. Am.*, 4(43):829–834, 1968.

[109] J. R. Searle. Speech Acts. *Cambridge University Press, London-New York*, 1969.

[110] E. Shriberg *et al.* Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? In *Language and Speech*, volume 41, pages 439–487, 1998.

[111] F. Souvannavong, B. Merialdo, and B. Huet. Partition Sampling: an Active Learning Selection Strategy for Large Database Annotation. *Vision, Image, and Signal Processing*, 152(3):347–355, June 2005.

[112] A. Stolcke *et al.* Dialog Act Modeling for Conversational Speech. In *AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing*, pages 98–105, 1998.

[113] A. Stolcke *et al.* Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In *Computational Linguistics*, volume 26, pages 339–373, 2000.

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

[115] P. Taylor, S. King, S. Isard, H. Wright, and J. Kowtko. Using Intonation to Constrain Language Models in Speech Recognition. In *Eurospeech'97*, pages 2763–2766, Rhodes, Greece, 1997.

[116] Simon Tong and Daphne Koller. Active Learning for Structure in Bayesian Networks. In *IJCAI*, pages 863–869, 2001.

[117] K. Tumer and J. Ghosh. Order Statistics Combiners for Neural Classifiers. In *World Congress on Neural Networks*, volume I, pages 31–34, Washington D.C., July 1995. INNS Press.

[118] K. Tumer and J. Ghost. Robust combining of disparate classifiers through order statistics. *Computer Science*, May 1999.

[119] K. Tumer and J. Gosh. Theoretical Foundations of Linear and Order Statistics Combiners for Neural Pattern Classifiers. Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin, 1995.

[120] A. Van den Bosch, E. Krahmer, and M. Swerts. Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches. In *39th Meeting of the Association for Computational Linguistics*, pages 499–506, Toulouse, France, 2001.

[121] A. Venkataraman, L. Ferrer, A. Stolcke, and Shriberg E. Training a Prosody-based Dialog Act Tagger from Unlabeled Data. In *ICASSP'03*, volume 1, pages 272–275, Hong Kong, April 2003.

[122] A. Venkataraman, A. Stolcke, and Shriberg E. Automatic Dialog Act Labeling with Minimal Supervision. In *Australian International Conference on Speech Science and Technology*, Melbourne, Australia, December 2002. Australian Speech Science and Technology Association.

[123] Anand Venkataraman, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. Does Active Learning Help Automatic Dialog Act tagging in Meeting Data? In *Interspeech'2005*, pages 2777–2780, Lisbon, Portugal, September 2005.

[124] C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University, 1992.

[125] F. Wioland. *Les structures rythmique du francais*. Slatkine-Champion, Paris, 1985.

[126] J. Woller. *The Basics of Monte Carlo Simulations*. University of Nebraska-Lincoln, Spring 1996. http://www.chem.unl.edu/zeng/joy/mclab/mcintro.html.

[127] H. Wright. Automatic Utterance Type Detection Using Suprasegmental Features. In *ICSLP'98*, volume 4, page 1403, Sydney, Australia, 1998.

[128] 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.

[129] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Meeting of the Association for Computational Linguistics, ACL*, pages 189–196, 1995.

[130] T. Zhang and F. J. Oles. A Probability Analysis on the Value of Unlabeled Data for Classification Problems. In *ICML*, pages 1191–1198, Standord, CA, USA, June 29-July-2 2000.