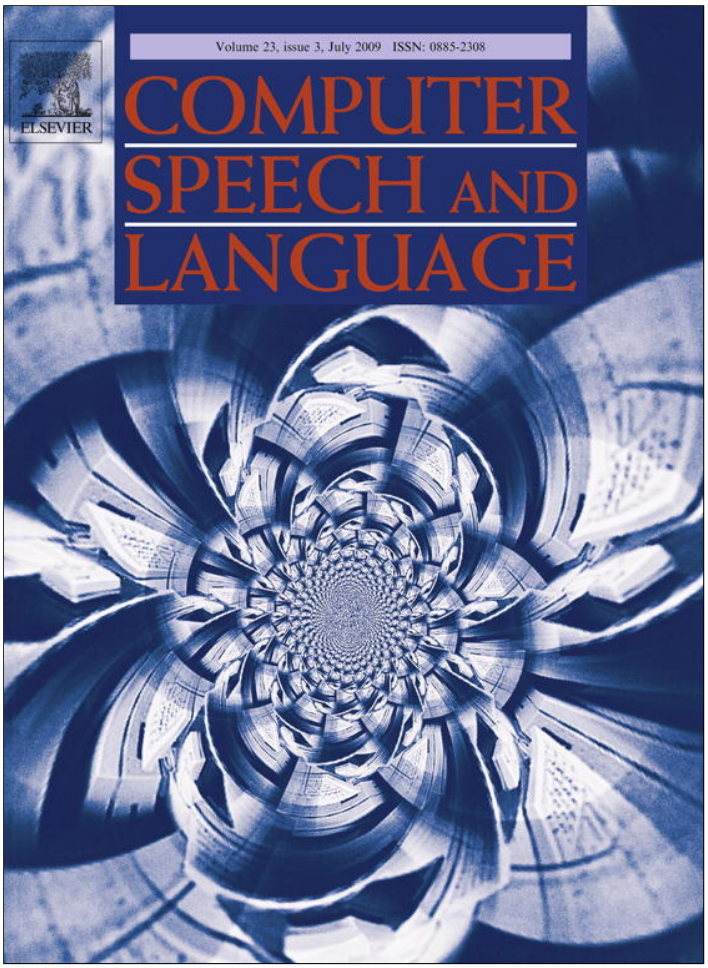

Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 277–310
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

An analysis of prosodic information for the recognition of dialogue acts in a multimodal corpus in Mexican Spanish

Sergio R. Coria [a,b,*], Luis A. Pineda [a]

[a] *Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Depto. de Ciencias Computacionales, Ciudad Universitaria, México, D.F., Mexico*
[b] *Universidad de la Sierra Sur, Dept. of Informatics, Miahuatlán de Porfirio Díaz, Oax., Mexico*

Received 13 August 2007; received in revised form 3 June 2008; accepted 11 June 2008
Available online 2 July 2008

Abstract

This paper presents empirical results of an analysis on the role of prosody in the recognition of dialogue acts and utterance mood in a practical dialogue corpus in Mexican Spanish. The work is configured as a series of machine-learning experimental conditions in which models are created by using intonational and other data as predictors and dialogue act tagging data as targets. We show that utterance mood can be predicted from intonational information, and that this mood information can then be used to recognize the dialogue act.

Keywords: Speech act; Dialogue act; Intonation; INTSINT; DAMSL; DIME-DAMSL

1. Introduction

Spoken language processing in artificial intelligence (AI) presents a large number of research challenges in the first decade of the 21st century. Some of the most demanding are the robustness improvement in automatic speech recognition (ASR) systems and the exploitation of diverse information sources that speakers use when participating in a dialogue.

Since one of the goals of AI is to implement efficient applications to process natural language, this area takes into account knowledge from diverse sciences that study the linguistic phenomena; some of them are linguistics, philosophy of language, psycholinguistics and phonetics.

One of most relevant contributions in philosophy of language is Searle's notion that the minimal unit for linguistic communication is the speech act instead of the symbol, the word or the phoneme. This idea introduced a new approach in the linguistic studies. Linguistics, in turn, suggests that one of the means to express

* Corresponding author. Tel.: +52 951 57 24 100x205; fax: +52 55 56 22 36 20.
E-mail addresses: coria@unsis.edu.mx (S.R. Coria), luis@leibniz.iimas.unam.mx (L.A. Pineda).
URL: http://leibniz.iimas.unam.mx/~luis (L.A. Pineda).

0885-2308/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.csl.2008.06.003

the speech act is utterance intonation. These two claims must be addressed by AI on an empirical basis as they provide a better understanding of spoken language and, as a consequence, elements to implement real-world computational applications.

In computational linguistics, the speech act has been frequently investigated under its notion of dialogue act (DA); i.e. focusing on its communicative function within the dialogue context instead of investigating it as a self-contained unit.

Spoken dialogue is highly complex and its computational analysis and modeling are highly complex too. Thus, computational linguistics studies this phenomenon under controlled conditions and within constrained scopes. For instance, the so-called *practical dialogues* or *task-oriented conversations*, described in (Allen et al., 2000) are those whose participants have established a previous agreement (either implicitly or explicitly) about reaching a goal (or series of goals) that will be reached on a cooperative basis. This kind of dialogues involve a restricted number of speech act types with simpler relations than the speech acts appearing in normal spontaneous conversation, and thus can be modeled computationally more easily.

In addition to its theoretical value, the practical dialogue analysis is also useful to engineering purposes in real-world applications because one of the potential applications of human–computer interaction via spoken language is performing tasks to reach goals on a collaborative basis.

The interpretation of speech acts is a very complex phenomenon involving a large number of information sources; however, it is not completely understood how much and how each source contributes. If this contribution could be estimated and modeled on a formal and statistical basis, it might facilitate the improvement of ASR and automatic dialogue management systems.

The most frequently adopted approach to automatic DA recognition has been analyzing the lexical content acquired from ASR systems. Once an ASR system produces a lexical transcription, this approach applies cue-word search, or language models and, in some implementations automatic planning techniques. Since the 1990s a broader approach has been addressed and the contribution from other sources, such as intonation and the inherent structure of dialogue, has been evaluated. Taking this into account, this work addresses the analysis and modeling of the relation between dialogue act and intonation; in addition, also the contribution of other information sources for DA recognition is modeled and evaluated. The aim is to describe and understand the phenomenon and, in addition, to create models that might provide guidelines for the implementation of efficient systems for spoken dialogue management. The analysis of lexical information is not within the scope of this investigation.

How do intonation interacts with other information sources in the conversation to communicate DAs? How much can the dialogue participants rely on intonation only for the comprehension of DAs? How can a computer be taught to exploit intonational information to enhance its performance and accuracy when interacting with humans via spoken-language interfaces? These are some of the questions that this work attempts to answer. The problem is stated as an empirical evaluation of the contribution of intonational information to DA recognition by using machine-learning algorithms for the construction of classification trees on data from a practical dialogue corpus.

Since the phenomenon needs to be analyzed on a quantitative basis from real speech data, statistical analyses and machine-learning models are important elements of this work. Particularly, machine-learning models might not only describe but also explain to some extent how the diverse information sources interact. The empirical view has also been addressed by previous work in the area, such as Shriberg et al. (1998) for English, the VERBMOBIL project Wahlster (1993) for German and Japanese, and Fernández and Picard (2002) for Spanish.

Most of previous work is for English, and cultural elements might influence the DA phenomenon; e.g. action directives are uttered as imperatives in some languages while other languages or even dialects prefer declaratives or interrogatives for this type of DA. Thus, the findings about the relation between DA and intonation for English might not be representative for other languages. This is a good reason to analyze the phenomenon from the Spanish language perspective. In addition, most of empirical work on Spanish intonation focuses on its relation to utterance mood. Furthermore, the large number of Spanish speakers is another reason demanding deeper investigation on the diverse linguistic phenomena of this particular language from the computational view. The study of the relation between intonation and DA in Spanish is a novel contribution

of this work. The empirical base for this investigation is the DIME Corpus,[1] an audio and video corpus of task-oriented conversations in a design domain (i.e. kitchen design).

Other novelty is the use of INTSINT (Hirst et al., 2000), an intonational annotation scheme in which labels are semi-automatically assigned to inflection points of the intonational contour, producing a discrete representation. On this basis, an intonation contour is represented as a string, which in turn represents the sequence of relative tones along with their respective timestamps. Previous work have used numerical representations mainly; i.e. statistical computations such as averages, standard deviations, maxima, and minima. INTSINT annotations are so reliable that they can be used to produce the synthesized version of an original intonational contour.

Previous theoretical and empirical work on speech act, dialogue act and Spanish intonation are considered as background for the present investigation. Also, annotation schemes are intensively used for DA as well as for intonation. The paper is organized as follows: Section 2 introduces the analysis of the relation between intonation and dialogue act. Section 3 addresses the background and state of the art. Section 4 describes the DIME Corpus, a speech-and-video corpus on practical dialogues. Section 5 describes the DIME-DAMSL scheme to annotate dialogue acts (Pineda et al., 2006a,b; Pineda et al., 2007), which was used in the reported experiments. Section 6 addresses the representation of intonation, including an overview of intonational systems. Section 7 describes the experimental conditions and results, and also a number of possible applications of the results are suggested. Finally, some conclusions are presented in Section 8.

Results show that intonational information plays a role in the recognition of dialogue acts, although this role is not definite and depends on other non-intonational sources. The work is an initial exploratory analysis on a particular corpus, and although a number of questions are answered to some extent, many others need deeper analyses by using larger volumes of empirical data as well as other representation schemes for DA and intonation.

2. Intonation and dialogue acts

In the 1960s, Searle's theory on speech acts states that the production or emission of an utterance instance under certain conditions constitutes a speech act, and speech acts are the basic or minimum units of linguistic communication. The notion of dialogue act is an extension of the former and involves "a speech act in the context of a dialogue" (Bunt, 1994), or an act with internal structure related specifically to its dialogue function, as assumed in (Allen and Core, 1997), or a combination of the speech act and semantic force of an utterance (Bunt, 1995). This work adopts the notion by Allen and Core (1997).

Intonation is one of the sources humans use to recognize dialogue act types in utterances. A number of theoretical and empirical works (e.g. Wilson et al., 1988) have proposed that utterance mood and thus intonation provide instructions to the listener on how to process the lexical content of an utterance; such instructions are highly useful to disambiguation tasks when the dialogue act type cannot be recognized from the lexical content solely. Analyses of the relation between intonation and dialogue act might allow us to create models to improve the performance of dialogue management systems. Intonation is also related to utterance mood, e.g. interrogative, declarative, and imperative. Utterance mood constitutes a characterization of intonational and syntactic patterns in speech and provides a vehicle to the dialogue act, too; that is why dialogue act recognition can be supported by utterance mood recognition.

In Spanish, the utterance intonational pattern is more relevant than syntactic structure for utterance mood classification purposes. The most common utterance mood types in Spanish are: declarative, interrogative, imperative and exclamative. According to most of Spanish intonation descriptions, such as Navarro-Tomás (1948), Sosa (1999) and Quilis et al. (1981), declaratives show a flat intonational contour and the main verb is not in imperative mode. Interrogatives are those in which the final region of the intonational contour is rising or falling-rising. The intonational pattern in exclamatives is usually falling. The pattern of imperatives is comparable to that of declaratives or exclamatives; in addition, the verb is in a particular mode and tense, so their recognition requires a syntactic analysis as well. Rising end is an information source to distinguish between two general classes: interrogatives and non-interrogatives.

[1] http://leibniz.iimas.unam.mx/~luis/DIME/CORPUS-DIME.html.

3. State of the art

The state of the art for this work is determined by three key topics: (1) theories and tagging schemes of dialogue acts, (2) theories and representational schemes of intonation, and (3) studies on the relation between dialogue act, utterance mood and intonation. These topics are presented below.

The notion of dialogue act can be considered a specialization of the notion of speech act, as stated by Searle, in order to be implemented and analyzed on a computational basis. Bunt uses the term dialogue act for referring to functional units used by the speaker to change the context in a dialogue. Allen and Core implement DAMSL (Dialogue Act Markup in Several Layers), an annotation scheme for dialogue acts that has been used in a number of previous work in the area.

DAMSL provides four dimensions of analysis for dialogue act: communicative status, information level, forward-looking and backward-looking function. Every dimension has a tagset. The communicative status describes whether the utterance is intelligible and whether it was successfully completed; the information level classifies the semantic content of an utterance into task, task management or communication management. The forward-looking function includes dialogue acts that constrain the future beliefs and actions of the dialogue participants; e.g. action directive, information request, and affirm. Finally, the backward-looking function includes dialogue acts that relate the current utterance to the previous discourse; e.g. accept, reject, and answer. The complete DAMSL tagset is described and explained in (Allen and Core, 1997).

Another approach for dialogue act annotation is adopted in the Verbmobil project (Wahlster, 1993). A taxonomy of dialogue act types is presented in (Jekat et al., 1995). A key difference between this and DAMSL is that Verbmobil addresses dialogue acts in the specific domain of business-appointment scheduling and the tagset is constrained to this, whereas DAMSL has a broader scope and can be used to annotate practical dialogues in arbitrary domains.

Besides dialogue act theories and their respective annotation schemes, theories on intonation, particularly for Spanish, are a key input for this work. The toneme (*tonema*), introduced by Navarro-Tomás (1948), refers to the final region of the intonational contour of an utterance. Both Navarro-Tomás and Quilis et al. (1981) state that this region carries a large amount of linguistic information to determine the mood of utterances as well as the corresponding speech act type.

One of the best known schemes for intonation annotation from the phonologic view is ToBI, Tone and Break Indices (Beckman, 1997), whose implementation for Spanish, Sp-ToBI, is presented in (Beckman et al., 2002). INTSINT (Hirst et al., 2000) is another scheme to represent intonation; it is based on an intonational contour stylization of inflection points and their tone annotation.

Regarding the research on the relation between intonation, utterance mood and dialogue act, some relevant proposals have been presented by Wilson et al. (1988), Mast et al. (1996), Garrido (1991, 1996) and Shriberg et al. (1998). Wilson and Sperber analyze sentence mood from a semantic view and their claim is that the characteristic linguistic features of declarative, imperative or interrogative form merely encode a rather abstract property of the intended interpretation. Mast et al., in the Verbmobil project, use empirical data to implement recognition models of dialogue act in which features extracted from intonation are used among the input data. Garrido addresses an empirical analysis of intonational patterns in Spanish adopting classical theories as a baseline to analyze utterance mood in empirical data from a speech corpus; his results extend those of his predecessors. Finally, Shriberg et al. use data from a corpus annotated with DAMSL as well as machine learning techniques, particularly classification trees, to produce models for recognition of dialogue act types; they represent intonation as raw signal and their results show that intonation in spoken English can contribute to the recognition of dialogue acts.

In (Jurafsky et al., 1998) the authors investigate the *back-channel*, *acknowledgment* and *yes-answer* types of dialogue acts in multiparty dialogue. They consider four subtypes of *back-channel*: *continuers*, *assessments*, *incipient speakership*, and *agreements*. The analysis considers lexical, prosodic and syntactic information. Lexical knowledge allows the listener to distinguish these dialogue acts under ambiguity conditions; e.g. whith words like *yeah*. Prosodic knowledge contributes to the identification of certain DA types, while cue words may suffice for the remainder. Their results suggest that particular DA types, such as *assessments*, might present a specific microsyntax.

Venkataraman et al. (2003) address the use of unlabeled data for training HMM-based dialog act taggers. Three techniques are succesfully exploited: (1) iterative relabeling and retraining on unlabeled data, (2) a dialog grammar to model dialog act context, and (3) a model of the prosodic correlates of dialog acts. By the combined use of prosodic information and unlabeled data on the SPINE corpus (Navy Research Laboratory, 2001), the tagging error is reduced between 12% and 16% over a baseline in which word information and a diversity of labeled data volumes are available.

Rangarajan et al. (2007) propose two schemes to integrate prosody in DA modelling: (1) syntax-based categorical prosody prediction from an automatic prosody labeler and (2) models of continuous acoustic-prosodic observation sequence as a discrete sequence by using quantization methods. The authors report a relative improvement of 11.8% compared to using lexical and syntactic features alone on the Switchboard-DAMSL corpus (Godfrey et al., 1992). The modeling algorithm is maximum entropy. Accuracy is 84.1%, which is higher than using the lexical and syntactic features from three previous utterances (83.9%); the authors claim such results as some of the best for the task.

Ang et al. (2005) explore the tasks of dialogue act segmentation and classification on data from the ICSI Meeting Corpus (Janin, 2003) by employing simple lexical and prosodic knowledge sources. They contrast results for manually-transcribed versus automatically recognized words. They observe that both tasks are difficult, specially for an entirely automatic system. A complementary prosodic model improves performance over lexical information alone, especially for segmentation. The investigation used very simple lexical and prosodic features. Lexical features are: word *n*-gram information for segmentation and a series of lexical cues for classification. Prosodic features are pause information for segmentation and a set of features for classification, including: pause, duration, pitch, energy, and spectral tilt features, many normalized by speaker-specific statistics and/or phonetic context. Accuracy in the two tasks is impacted by word recognition errors; however, lexical-based segmentation is more degraded than prosody-based when these errors occur. By comparing DA classification results for meeting data to previous results for telephone conversations, meetings show little gain from DA context modeling.

Tur et al. (2006) present a supervised adaptation method for dialog act tagging. The investigation evaluates the model adaptation for dialog act tagging by using out-of-domain data or models. The authors use the ICSI meeting corpus (Janin, 2003) as empirical resource with the MRDA (meeting recognition dialog act) tagset. The DA set is: questions, statements, back-channels, disruptions, and floor grabbers/holders. Controlled adaptation experiments were performed using the Switchboard (SWBD) corpus (Godfrey et al., 1992) with SWBD-DAMSL tags as the out-of-domain corpus. Results show that a better DA tagging can be obtained by an automatic selection of a subset of the corpus. The confidences obtained by both in-domain and out-of-domain models are combined via logistic regression; this is particularly useful when the in-domain data is limited. Their experimental conditions produce the same classification accuracies by using approximately 50% less labeled data.

4. The DIME Corpus

In this work, the approach to analyze intonation and dialogue act is empirical, so data from actual dialogues are required. These dialogues need to satisfy some restrictions to be useful to our objectives: (1) they should be task-oriented conversation, and (2) they should contain spontaneous speech representativity. The first restriction arises because this class of dialogue is one of the most attractive to the implementation of human–computer interaction systems, and also because the number and complexity of dialogue act types in this particular class of dialogue are less than those in other types of conversation, which facilitates analyses under controlled conditions. Second, empirical data, i.e. utterances in dialogues, should mirror as much as possible the behavior of a user speaking as spontaneously as possible while participating in a task-oriented dialogue. These two issues determine the requirement to use dialogues generated by using a protocol similar to the so-called Wizard of Oz (Dahlback et al., 1993).

In the DIME Project (Pineda et al., 2002), a corpus satisfying the requirements described above, the DIME Corpus (Pineda, 2007), is available. The DIME Corpus is a collection of video, audio and annotation files containing 26 dialogues in which participants are grouped into two categories: a speaker, named *the System* or *the Wizard*, acting as if he was the computer, and the other speakers acting as users of the System. The goal in the

Table 1
DIME corpus sample statistics

No. of dialogues	12	
No. of participants (total)	24	
No. of participants (distinct)	13	
No. of dialogues per participant	0.9	
Lengths	Turns	Minutes
Corpus sample	1038	115.8
Dialogue avg.	87	9.7
Largest dialogue	136	20.0
Shortest dialogue	49	4.4

dialogues is to arrange pieces of furniture in a virtual kitchen by using a computer-aided design software tool as specified by a layout on paper given to each user. The System participated in every dialogue by performing the graphical actions, e.g. adding a piece of furniture, moving it, and deleting it as specified by user utterances; the System can talk to the user to provide (or to ask for) information.

Annotation layers were defined taking into account phonetic and phonological features previously showed (or suggested) as correlated to dialogue act or to utterance mood by theoretical and empirical research in the area. Therefore, annotations to be tagged in the corpus were defined as follows: orthographic transcription and segmentation into utterances; on the segmental phonetic level, allophones and phonemes; on the supra-segmental phonetic level, phonetic syllables and intonational contours; on the phonologic layer, words, parts of speech (POS), break indices from the Sp-ToBI model (Beckman et al., 2002) and utterance mood. Finally, a layer for dialogue acts was defined (Pineda et al., 2006a).

Dialogue act annotation with DIME-DAMSL is one of the most important pieces of information for this work because it allows to analyze the phenomenon in question and to represent the target data in the recognition models. This annotation guaranteed a high enough consistency for statistical analyses and machine-learning modelling.

Once the tagging layers were determined, tagging schemes were chosen. Orthographic transcription was annotated with the Spanish alphabet characters. Allophones and phonemes, as well as phonetic syllables and words, were annotated with the MexBet alphabet (Cuétara, 2004). POS level was annotated with a tagset collected and refined by Moreno and Pineda (2006). Break indices were annotated with Sp-ToBI. Intonation was semi-automatically annotated with the International Transcription System for Intonation, INTSINT (Hirst et al., 2000). Utterance mood was manually tagged based on both the acoustic perception of intonation and also on the syntactic structure of utterances.

Regarding the annotation tools, the orthographic transcription was performed using a basic text editor. CSLU Toolkit (OHSU, 2004) was used to annotate most layers: allophones, phonemes, phonetic syllables, words, break indices and POS. Intonation was annotated by using Motif Environment for Speech, MES (Espesser, 1999). Forms in a basic spreadsheet were used for utterance mood and dialogue act annotations. Table 1 presents general statistics of the sample from the DIME corpus that is analyzed in this investigation.

5. DIME-DAMSL theory and application

Spontaneous speech is used for conversational purposes mainly. According to Searle, the minimal unit for communication is the speech act. When a speech act occurs in a conversation, it performs a particular function, the so-called dialogue act. DA taxonomies can be stated by addressing similarities among the functions; for instance, DA types that present, request, and accept or reject information.

The number and complexity of DA types are related to the nature of specific dialogue classes. According to the number of participants and the dialogue purposes, some of the dialogue classes are, for instance, that in which only two individuals interact, multi-party conversation, informal conversation, and task-oriented dialogue.

A DA instance is determined by the conversational context and, in turn, this is determined by the knowledge, beliefs and presuppositions that the interlocutors share. The DAs preceding an utterance are context elements as well.

Since spoken dialogue phenomena can be highly complex, its empirical investigation from the computational linguistics view has to be addressed within controlled conditions. Task oriented or practical dialogues is one class that provides a suitable framework as the number and complexity of their DA types is lower than in other classes and because real-world applications can be implemented from its modeling.

Computational linguistics has proposed a number of annotation schemes for DA investigation. One of the most known is DAMSL (Allen and Core, 1997). DAMSL provides a conceptual frame and a series of rules to annotate dialogue acts; however, these do not suffice to obtain a satisfactory inter-annotator agreement. For instance, Tables 2 and 3 present consistency annotation scores of DAMSL on a number of corpora. In both tables, most *Kappa* (*K*) values (Core, 1997, 1998) are less than 0.7, the value commonly used as threshold to consider annotations consistent. A source of low agreement in DAMSL is the lack of a higher level structure to constrain the possible label(s) an utterance can be assigned to; i.e. the scope of DAMSL rules is restricted to analyze single utterances and perhaps a non-specific number of previous or subsequent utterances but without considering formal stages along the dialogue. This allows a broad space to select and combine labels but, on the other hand, there is a high risk that inter-annotator agreement to be low.

DIME-DAMSL (Pineda et al., 2007), adopts the DAMSL tagset and dimensions and extends them by defining three additional notions, as follows:

(1) two expression planes: the obligations and the common ground;
(2) charge and credit contributions in balanced transactions;
(3) transaction structure.

The obligations and the common ground planes are parallel structures along which dialogue acts flow. A dialogue act might contribute to any (or both) of the two planes.

In DIME-DAMSL, the obligations plane is construed by dialogue acts that generate a responsibility either on the speaker himself or on the listener to perform an action, either verbal or non-verbal; e.g. the obligation to provide some piece of information or to perform a non-verbal action. Dialogue acts that mainly contribute to the obligations plane are: *commit*, *offer* (when it is accepted by the interlocutor), *action directive* and *information request*. For instance, in utterances from dialogues of the DIME corpus, *okay* is a commit (in certain contexts); *can you move the stove to the left?* is an action directive, and *where do you want me to put it?* is an information request. Table 4 presents the complete tagset for obligations (including multi-label tags) and statistics for the DIME corpus sample.

The common ground is the set of dialogue acts that add, reinforce and repair the shared knowledge and beliefs of the interlocutors and preserve and repair the communication flow. DIME-DAMSL defines two sub-planes in the common ground: *agreement* and *understanding*; agreement is the set of dialogue acts that add knowledge or beliefs to be shared between dialogue participants; understanding is defined by acts that keep, reinforce or recreate the communication channel. Dialogue acts that mainly contribute to the agreement sub-plane are: *open option* (e.g. *these are the cupboards we have*), *affirm* (e.g. *because I need a cabinet*), *hold* (e.g.

Table 2
Tagging consistency using DAMSL scheme in three corpora, cited from Core (1998)

	VERBMOBIL			TRIPS			MAPTASK		
	K	*PA*	*PE*	*K*	*PA*	*PE*	*K*	*PA*	*PE*
answer	0.37	0.78	0.64	0.59	0.91	0.78	0.53	0.86	0.70
info-request	0.58	0.85	0.65	0.62	0.88	0.68	0.62	0.90	0.75
influence-on-listener	0.35	0.66	0.48	0.59	0.80	0.51	–	–	–
influence-on-speaker	0.42	0.68	0.44	0.27	0.75	0.66	–	–	–
statement	0.34	0.67	0.50	0.32	0.70	0.56	0.55	0.75	0.46
agreement	0.46	0.71	0.47	0.52	0.77	0.52	0.21	0.72	0.65
understanding	0.15	0.80	0.76	0.40	0.77	0.62	0.29	0.72	0.61
info-level	0.49	0.87	0.74	0.55	0.87	0.71	0.41	0.89	0.82
other-forward-function	0.26	0.94	0.92	–	–	–	–	–	–

K: *Kappa*, *PA*: proportion of times that coders agree, *PE*: proportion of agreement expected by chance.

Table 3
Tagging consistency using DAMSL scheme in TRAINS 91–93 corpus, cited from Core (1997)

	Main forward function labels			Main backward function labels			
	Statement	*I.A.F.*	*Other F.F.*	*Unders.*	*Agreem.*	*Answer*	*Resp-to*
K	0.67	0.71	0.48	0.58	0.43	0.81	0.77
PA	0.82	0.88	0.92	0.83	0.78	0.95	0.83
PE	0.47	0.60	0.85	0.59	0.61	0.72	0.28
Significance level	0.000005	0.000005	0.000005	0.000005	0.000005	0.000005	0.000005

I.A.F.: influence addressee future action, *Other F.F.*: other forward functions, *Resp-to*: utterances collection responded to by an answer.

Table 4
DIME-DAMSL obligations tagset and statistics in the corpus sample

Obligations DA	Count	%
answer	225	21.6
info-request	175	16.8
action-dir	120	11.5
commit	112	10.7
info-request_graph-action	98	9.4
action-dir_answer	16	1.5
info-request_answer	8	0.8
info-request_graph-action_answer	8	0.8
offer	5	0.5
graph-action	1	0.1
offer_info-request	1	0.1
action-dir_offer	1	0.1
no-tag	273	26.2
Total	1043	

do you want me to move this cabinet toward here?), *accept* (e.g. *yes*), *reject* (e.g. *no, there is no design problem*), *accept part*, *reject part* and *maybe*. Dialogue acts on the understanding sub-plane are *acknowledgment* (e.g. *yeah*, *yes*, and *okay*), *repeat-or-rephrase* (e.g. *do you want me to put this stove here?*), and *backchannel* (e.g. *mhum*, *okay*, and *yes*). Table 5 presents the complete tagset for common ground (including multi-label tags) and statistics for the DIME corpus sample.

Charges and credits are the basic mechanism underlying the interaction between pairs of dialogue acts along each of the two expression planes. A charge generated by a dialogue act introduces an imbalance requesting for satisfaction, and a credit is the satisfactor for that charge. Certain pairs of dialogue acts generate a balanced situation; *e.g.* on the obligations plane an action directive makes a charge that can be balanced with a graphical action; on the agreement plane a charge introduced by an open option can be balanced with an accept; and on the understanding plane an affirm creates a charge that can be satisfied with an acknowledgment. These and other additional pairs guide the charge-credit annotation and allow to identify and to annotate the most prominent dialogue acts of the utterance. This annotation of dialogue acts is called Preliminary DIME-DAMSL and supports the completion of the dialogue act tagging in a subsequent stage, the so-called Detailed DIME-DAMSL. The difference between *detailed* and *preliminary* annotations is that detailed use *information level* labels (from DAMSL) while *preliminary* does not.

A transaction is defined by a set of consecutive charge-credit pairs intending to accomplish a sub-goal within a dialogue. A transaction presents two general phases: intention specification and intention satisfaction. In the most common case, the first phase starts with an offer or an action directive, and the second phase starts when the System begins, usually after a commit, to satisfy the action directive. Most of transactions in the corpus are focused on handling one single instance of a piece of furniture by performing one single type of graphical action at the virtual kitchen.

DIME-DAMSL evolved on an incremental basis by refining it along a series of tagging rounds with a team of annotators. Evaluations with *Kappa* statistics (Carletta, 1996) of the tagging data were performed after each

Table 5
DIME-DAMSL common ground tagset and statistics in the corpus sample

Common Gr. DA	Count	%	Common Gr. DA	Count	%
accept	383	36.7	*affirm_display*	3	0.3
no-tag	210	20.1	*accept_hold_ repeat-rephr*	2	0.2
graph-action	105	10.1	*affirm_maybe*	2	0.2
affirm	73	7.0	*offer_accept*	2	0.2
hold_repeat-rephr	54	5.2	*display*	2	0.2
open-option_display	41	3.9	*offer*	2	0.2
ack	24	2.3	*affirm_hold*	2	0.2
accept_part	20	1.9	*affirm_accept-part_exclamation*	1	0.1
reaffirm	16	1.5	*affirm_accept_ exclamation*	1	0.1
hold	13	1.2	*hold_ on_task_mangmt*	1	0.1
n.u.s.	12	1.2	*maybe*	1	0.1
reject	11	1.1	*reaffirm_hold*	1	0.1
affirm_accept	9	0.9	*open-option_accept*	1	0.1
open-option	6	0.6	*other_in_ commongr*	1	0.1
repeat-rephr	5	0.5	*reaffirm_ complementation*	1	0.1
conv-close	5	0.5	*conv-open*	1	0.1
offer_conv-open	5	0.5	*affirm_graph-action*	1	0.1
perform	4	0.4	*affirm_correct*	1	0.1
affirm_reject	4	0.4	*open-option_reject*	1	0.1
reject-part	4	0.4	*affirm_perform_ conv-close*	1	0.1
back-channel	3	0.3	*hold_nus*	1	0.1
graph-action_accept	3	0.3	*affirm_conv-close*	1	0.1
open-option_ display_accept	3	0.3	*Total*	*1043*	

round; annotation discrepancies were discussed among the annotators in order to determine if the reasons were typos, misunderstanding of the annotation rules, or lack of explicit annotations conventions. When necessary, new conventions were defined and included in a tagging manual. At the end, *Kappa* scores show that DIME-DAMSL is a scheme that produces consistent annotation data, which increases probabilities to produce reliable machine-learning models. DIME-DAMSL *Kappa* scores are presented in Pineda et al. (2006a) and Pineda et al. (2007).

Table 6 presents the consistency tagging scores for DA annotations by two expert annotators on the DIME Corpus particular sample that is used for this work. Comparable *Kappa* values, between 0.81 and 0.85 are reported by Shriberg et al. (1998). The DA annotation task was performed in three rounds. In each round,

Table 6
Kappas and agreement percents in dialogue act annotation with Preliminary DIME-DAMSL by two annotators

Dialogue	Nr. of utts.	*K*		*PA %*		*PE %*	
		Obligs.	Comm. Gr.	Obligs.	Comm. Gr.	Obligs.	Comm. Gr.
d01	116	0.9	0.7	91.4	75.9	16.2	10.0
d03	168	0.8	0.6	83.3	65.5	20.6	12.2
d12	117	0.9	0.9	94.0	88.9	20.0	15.7
d13	191	0.9	0.8	92.7	85.9	17.1	20.4
d14	137	0.9	0.8	91.2	85.4	19.1	16.5
d15	90	0.9	0.7	88.9	75.6	15.8	15.7
d17	237	0.8	0.7	84.8	70.9	14.6	13.6
d19	105	0.8	0.7	85.7	74.3	23.0	14.5
d21	69	0.9	0.8	92.8	79.7	18.5	17.1
d22	181	0.8	0.8	86.7	79.6	17.5	18.8
d23	81	0.9	0.9	90.1	87.7	15.3	17.6
d26	210	0.7	0.7	74.3	72.9	14.0	11.6
Avg.	141.8	0.9	0.8	88.0	78.5	17.6	15.3

K: *Kappa*, *PA%*: percent of times that coders agree, *PE%*: agreement percent expected by chance.

every annotator produces the tagging for all utterances in the sample, so two annotation datasets are produced and a *Kappa* value is computed from them. Then, inter-annotator discrepancies are found and discussed by the taggers in order to refine annotation conventions for the following rounds. Values in Table 6 were obtained after the third round.

Table 7 presents a transaction example from one of the DIME Corpus dialogues. *Ch* and *Cdt* columns are the charges and credits, respectively, corresponding to the obligations and common ground planes. The latter is divided into agreement and understanding. The column for DA type, also divided into obligations and common ground, shows the preliminary DIME-DAMSL tagging. The transaction goal is to place a piece of furniture, particularly a sink, on a determined location in the kitchen. The two transactional phases are observed: in the first phase (utts. 18–29) the User interacts with the System to specify both the action and object types and also indicates the location to place the object. The satisfaction phase (utts. 30–31) involves the System performing the graphical action and confirming the User satisfaction at the end.

In the example, the intention specification phase begins as an action directive and ends as a commit; this is highly frequent in the corpus. Since the User asks for an object insertion action, the System interprets that the User needs to specify a particular type of object, thus the furniture catalogue is displayed. The action directive (18) creates a charge on the obligations plane because it creates the obligation on the listener to perform the action; it also creates a charge on the agreement plane because it introduces information that needs to be accepted by the listener, which occurs in 19. Utterance 20 makes an agreement charge because it also introduces information to be accepted in 21–22. These two utterances also provide information to the System, who accepts it in 23. Once the object instance is determined by the User, the target location must be determined as well. The System needs to ask this in 24 because the User did not provide it; the information request creates a charge on the obligations because it urges the listener to inform and he does in 25. His answer introduces information; however, its acceptance is postponed by a new information request along with a hold (26). The hold involves that the individual needs that some information to be clarified prior to accept it. The hold does not create any charge on the agreement; the charge is created by the information request. The location information is vague to the System. An evidence for this is that in 26 the System refers to a location that the User does not accept. The vagueness is solved in 27 and 28; the reject in 27 balances the charge from 26. A

Table 7
Analysis of a transaction with DIME-DAMSL

Utt #	Turn	Utterance	Obligations		Common ground				Dialogue act type	
			Ch	Cdt	AGR		UND		Obligations	Common ground
					Ch	Cdt	Ch	Cdt		
18	U	Mm... ⟨*no-vocal*⟩ I need a sink	18		18				*action-dir*	*action-dir*
19	S	Okay				18				*accept*
20		These are the four types of sink that we have			20					*open-option*, *display*, *point-obj-coord*
21	U	Let's see... this			21	20				*accept*, *affirm*, *visual*, *point-obj*
22		I select this sink with dish-washing machine			21	20				*accept*, *affirm*, *visual*, *point-obj*
23	S	Okay				21				*accept*
24	S	Where do you want me to put it?	24						*info-request*	
25	U	Let's see. At this ⟨*sil*⟩ at this location		24	25				*answer*	*affirm*, *point-zone*
26	S	⟨noise⟩ Do you want me to put ⟨*sil*⟩ this sink ⟨*sil*⟩ next to the stove?	26		26				*info-request*	*hold*, *point-obj*, *point-zone*
27	U	No		26		26			*answer*	*reject*
28	U	Mmm... ⟨*sil*⟩ at the wall with the windows			28					*affirm*
29	S	Okay	29			28, 25			*commit*	*accept*
30	S	⟨*no-vocal*⟩ is it o.k. there ?	30	29, 18	30				*graph-action*, *info-request*	*graph-action*
31	U	Yes, so far		30		30			*answer*	*accept*

Utt #: utterance number, *turn*: speaker turn (System or User), *Ch*: charge, *Cdt*: credit.

more specific location information is presented in 28, and it is accepted in 29. Also, in 29 the System accepts 25 and commits to perform the action. This commitment cannot be stated without the complete specification of the action type, the particular object instance and the location for the object.

The satisfaction phase begins when the System performs the action (30) and ends when the User accepts (31) the action results. The results must not violate domain or design constraints. In a prototypical situation, this phase uses to be short; however, in certain cases, either the User or the System might begin one (or more) refinement sub-phases in which any of the action arguments (i.e. the action type, the object instance or the location) can be changed to some extent.

Since the scope of a transaction is mainly determined by a particular action type, an object instance and, eventually, a location, a canonical transaction is disjoint, non-nested, non-overlapped and clearly distinguished from the previous and the subsequent transactions. Although the DIME-DAMSL theory addresses these situations in general, particular conventions can be adapted for specific annotation projects. For instance, the first effort for DA annotation in the DIME project defined the operational convention that transactions should be annotated as disjoint, non-overlapped and non-nested units, so that the initial boundary of an overlapping or nesting unit should be annotated as the final boundary of the previous transaction and the beginning of the next.

6. Intonation representation

For the purposes of this work, the intonational information representation should allow us to perform both statistical analyses and machine-learning models. It should facilitate the discretization of intonational contours, involving the possibility of abstracting the contour as a simple data, either numeric or nominal. In addition, such representation should provide a simple way to find out regularities; i.e. evidences showing correlations between intonational contours and dialogue act types. The scheme should be as independent as possible of bias introduced by annotators.

6.1. Intonational systems

Models to represent intonation in linguistics and in computational science have evolved along two overlapping tracks: (1) signal processing from the engineering view, and (2) phonetics and phonology. Signal processing provides a merely physical representation without including linguistic knowledge; both fundamental frequency *f0* and amplitude, represented as numeric values series along a time axis are two of the main information sources considered by this approach. Phonetics and phonology representations incorporate perception-oriented as well as linguistic information analyses and also take advantage of the signal-processing view.

6.1.1. Signal-processing approach

The signal-processing approach is the simplest methodology to represent and to analyze intonation; in this approach the speech signal is sampled, quantized and represented as a number of vectors, allowing automatic analysis. Such data series mainly describe *f0* and amplitude variations along the time axis and they are the source to statistical analyses producing descriptive parameters; e.g. averages, maxima, minima, ranges, and standard deviations. Other frequently used parameters are those describing the presence and duration of pauses and others to describe the intonational contour slope. Those describing the slope provide information about the shape type of the contour final region; i.e. rising, falling or flat. This approach does not produce intonational tagging, but instead numeric features describing segments or suprasegments.

Every speaker presents particular *f0* and amplitude ranges and also a particular speech rate. Therefore, if utterances from more than one speaker are present in a speech corpus *f0*, amplitude and duration data must be statistically normalized, involving arithmetic transformations in order to fit into a standardized scale. This way, normalization eliminates individual variations of the speakers and the resulting parameters are statistically comparable. The parameters can be used in a series of automatic analysis and modeling tools. A number of previous work in the area uses the signal-processing approach; e.g. Shriberg et al. (1998) use it to investigate the relation between intonation and dialogue acts in English. Garrido (1991) uses it to analyze the same phenomenon in Spanish, focusing on *f0* and duration. Fernández and Picard (2002) use it for Spanish as well.

In the VERBMOBIL project, intonation is represented by using acoustic-prosodic information (Kompe et al., 1995). Using a time-aligned phoneme transcription as source, a series of parameters are computed for each syllable, assuming both the six previous and the six next syllables as scope. Two hundred and forty two attributes were produced from such parameters, so this is the previous work in the area in which the largest amount of features is implemented.

In (Garrido, 1991), the intonational contour is represented from a series of inflection points of the contour. An inflection point is that where a change in the slope sign occurs; the difference to the previous *f0* inflection must be greater than or equal to 10 Hz.

As the signal-processing approach performs neither a perceptive verification nor a manual annotation, subjectivity in interpreting speech phenomena is avoided and, thus, any potential bias introduced by annotators is avoided as well. On the other hand, an important disadvantage is that there is no certainty about how much the representation resembles the original stimulus.

6.1.2. ToBI

From the phonology approach, one of the most known intonational schemes is ToBI (Tones and Break Indices), by Silverman et al. (1992) and Beckman (1997), originally used for prosodic tagging of American English. ToBI defines a series of tagging levels; the most important are: *words*, *break indices* and *tones*. The tagging process needs to be supported by software to produce and analyze the utterance intonational contour, so that the tagging levels are time-aligned to the utterance signal.

On the words level the utterance is segmented into words and is orthographically transcribed. On the break indices level the degree of phonetic agglutination between words is tagged according to the annotators perception, selecting a label from the tagset as follows: 0 (zero) for syllabic reduction by vowel contact between words, like in synaloephas; 1 (one) for any other ordinary agglutination between words and 4 for intonational phrase (melodic group). Labels 2 and 3 allow the annotator to mark agglutination phenomena with an intermediate degree between 1 and 4, such as intermediate phrase, tonic group or clitic group.

On the tone level, labels are annotated to describe the shape of particular contour regions; the tagset is: *L* (Low) for low tone and *H* (High) for high tone. *L* represents a valley and *H* a peak on the contour. The $*$ symbol, on the right-hand side of a label, represents that such valley or peak are time-aligned with a lexical accent. For some particular languages, ToBI defines pairs of tone labels in order to represent compound tones; e.g. $L^* + H$, $L + H^*$, and $H + L^*$. There are also boundary tone labels to describe intonational phenomena occurring in the beginning or at the end of intonational phrases; the tagset is: *L%* and *H%*, where the former represents a falling of *f0* after the $L + H^*$ compound tone and the latter represents a rising after any tone.

As the tagging criteria to interpret intonational events in ToBI are not sistematic enough, a high level of annotation expertise is required to take advantage of the scheme. Other reason increasing the difficulty to annotate intonation with ToBI is the lack of a software tool to support the task on an automatic or semi-automatic basis. The low consistency in annotations complicates their use in statistical analyses and machine learning modeling. Despite ToBI is the most commonly used scheme for prosodic annotation, inter-annotator agreement is low most of times.

Sp-ToBI (Spanish ToBI), an adaptation of ToBI for Spanish (Beckman et al., 2002; Sosa, 1999, 2003) defines three additional levels: *syllables*, *miscellaneous* and *code*. In the syllables level, syllables are segmented and then phonetically annotated. The miscellaneous level allows us to annotate diverse phenomena that increase the analysis complexity, such as doubting pauses, disfluences, and laughing. The code level aims to annotate the speaker's dialect or sociolect. Sp-ToBI was first conceived as a research tool for Spanish prosody rather than as a complete tagging scheme and it lacks of a universally accepted standard.

6.1.3. INTSINT

The International Transcription System for Intonation, INTSINT (Hirst et al., 2000), is a scheme proposed from the perspective of the Aix-en-Provence school. INTSINT annotation is produced on a semi-automatic basis. The annotation task is developed in four stages: first, the fundamental frequency (*f0*) of the acoustic signal is obtained by three of the most known algorithms for *f0* extraction: AMDF (Average Magnitude Difference Function), autocorrelation, and comb function. Second, the MOMEL algorithm (Campione et al., 2000; Hirst et al., 2000) looks for inflection points (named *targets*) on the intonational contour and produces

a stylized contour of the *f0*. Third, a human annotator performs a perceptive verification by listening to the synthesized version of the stylized contour and by watching it as a graphical representation; the annotator is allowed to manually modify the location of targets in order to adjust the stylized contour so that it is listened as the original contour as much as possible. On the fourth stage, an automatic tool compares the relative position of each target to the prior and the next; the position determines the INTSINT label that the tool selects from the tagset. The task is supported by a software tool: Motif Environment for Speech, MES (Espesser, 1999); a recent version (Hirst, 2007) is implemented as a plug-in for PRAAT software (Boersma et al., 2007).

The INTSINT tagset (Hirst et al., 2000) is construed by eight labels: *T* (top), *B* (bottom), *M* (medium), *H* (higher), *L* (lower), *U* (upstep), *D* (downstep) and *S* (same). *T* and *B* are the highest and the lowest targets, respectively, along the f0 contour; *M* is the target at the medium height; each of these three labels usually appears only once along the contour. *H* and *L* are local maxima or minima, respectively; *U* and *D* are targets located at ascending or descending regions, respectively; *S* is a target located at the same height of its predecessor and describes a plateau; each of these five labels might appear more than once along the contour. Every label is associated to a timestamp (in milliseconds) that specifies the time when the frequency inflection occurs.

For instance, the original *f0* of the utterance *Eh...me puedes mostrar los tipos de muebles que tengo?* (Mmm... can you show me the kinds of furniture that I have?) is presented in Fig. 1. The production of the stylized contour using MOMEL algorithm is presented in Fig. 2. MOMEL cannot guarantee a perfect stylization and might produce a contour different from the original, as can be seen in Fig. 3 (e.g. regions marked with 1, 2, 3 and 4).

A human annotator performs a perceptual verification task in which inflection points could be relocated, eliminated or inserted until the stylized contour is perceived as the original F0 curve as shown in Fig. 4. Finally, INTSINT tags are automatically produced, as can be seen in Fig. 5; these are BSSUHSLHBSUTS.

In addition to these four stages, and for the particular purpose of this experiment, INTSINT strings were cleansed by unifying the S (*same*) tags because these are redundant. This transformation produces simpler strings without reducing the reliability of the representation. The final string for our example is BSUHSLHBSUTS.

Llisterri (1996) presents general analyses on the general efficiency and frequent failure types in data produced by the INTSINT annotation tool on other corpora. Most of errors are located in target points in final rising contours which are not detected by MOMEL and have to be manually added in order to obtain a perceptually good approximation to the original utterance (72.9%). Other situation is the missing of target points in initial position (15.4%). Those errors are linked to beginning and end of utterances in the passages where a pause exists.

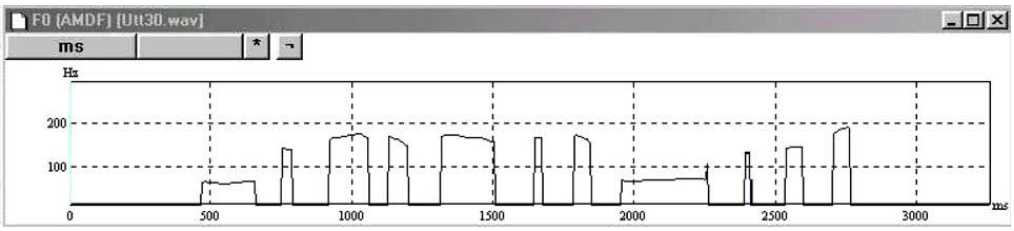

Fig. 1. Original fundamental frequency.

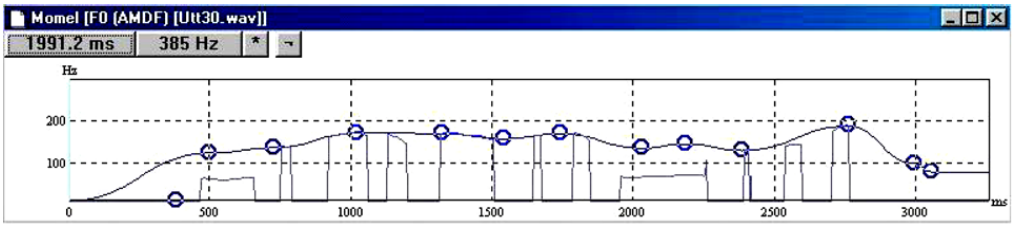

Fig. 2. Stylized *f0* (thickest contour) with inflection points (small circles).

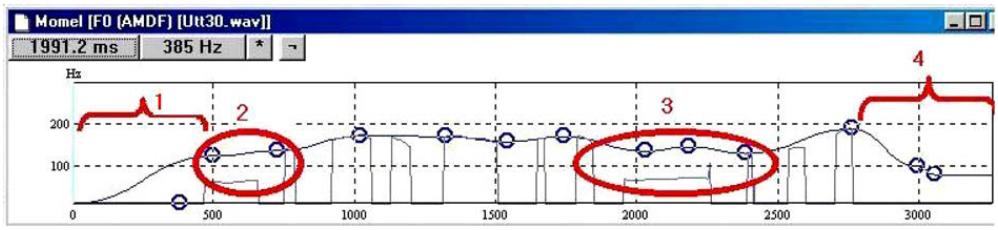

Fig. 3. A stylized contour that requires modification.

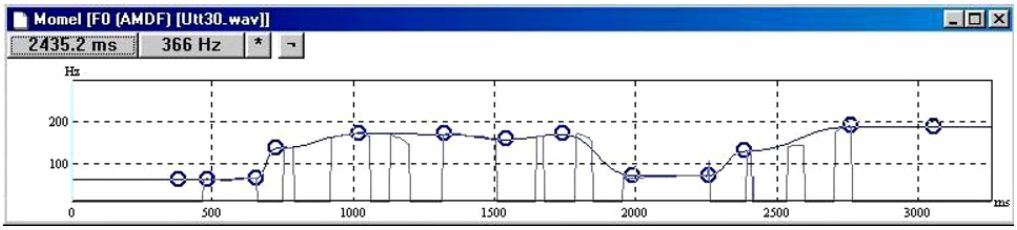

Fig. 4. Stylized *f0* after perceptual verification.

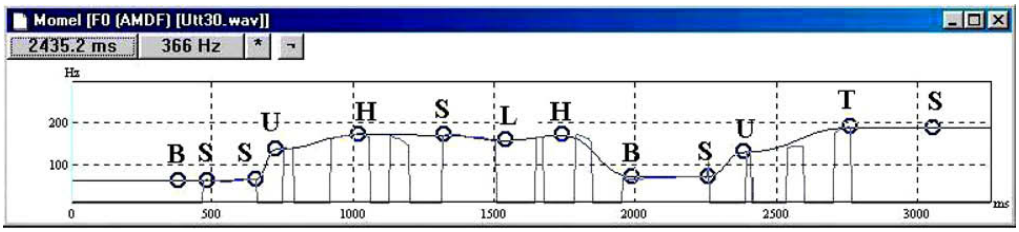

Fig. 5. INTSINT tagging of the inflection points.

A number of theories about the melody, i.e. the *pitch*, of utterances in Spanish state that this is mainly determined by fundamental frequency (*f0*) variations. The most influential authors on Spanish intonation, such as Navarro-Tomás and Quilis, focus on the descriptive information of *f0*. Léon et al. (1970) define it as *melodic variations of the utterance perceived by the listener*. In addition, a series of empirical investigations have focused on *f0* analysis; for instance, Garrido (1991), Rossi et al. (1981), Thorsen (1979, 1980) and 'T Hart et al. (1990). Therefore, this work focuses on the *f0* analysis and representation and does not address other speech signal elements such as amplitude (energy).

The lack of uniform criteria for Sp-ToBI annotation, the utility provided by INTSINT for pitch representation, and the availability of a semi-automatic tool for INTSINT annotation determine that this scheme was selected for this work.

7. Experiments and results

As one of the goals in this work is to find out and represent patterns describing the relation between intonation and dialogue acts, machine-learning experiments were designed as a means to create models able to describe this relation and to recognize the dialogue act type in particular instances of utterances. Management systems of spoken dialogue can take advantage of such models to use dialogue act type as an input in addition to the lexical content of utterances.

A supervised machine-learning algorithm, J48, implemented in the Waikato Environment for Knowledge Analysis, WEKA (Witten and Frank, 2005), is used to create classification and regression trees to recognize dialogue act from intonation, utterance mood and other non-lexical pieces of information. J48 resembles CART (Classification and Regression Trees algorithm) by Breiman et al. (1983). A dataset is required to

create and to test the models; it must contain attributes, i.e. features, representing the phenomenon to be analyzed, arranged into two groups: predictors (usually more than one feature) and target (usually one single feature). Predictors are those that determine the value of a target feature. Previous work in the area shows that dialogue act is correlated to intonation and, in turn, this latter is correlated to utterance mood, thus these are the three main features considered to specify the experimental conditions.

The specific implementation of target features from the dialogue act annotation data is an important issue because it determines what can be recognized and it can also influence the recognition accuracy. Using the detailed DIME-DAMSL tagging data would produce a too heterogeneous dataset because of the large degree of detail in tagging, thus reducing the probabilities to find general patterns. Models produced from this dataset would show a poor capability of generalization and they would not be able enough to describe the phenomenon to a satisfactory extent. Therefore, the preliminary DIME-DAMSL annotation data, such as in Table 7, are used instead of the detailed, thus abstracting the most prominent dialogue acts expressed by utterances.

The DIME corpus layers used to create the predictor and target features in the dataset are: preliminary DIME-DAMSL annotation, INTSINT annotation of intonation, utterance duration, speaker role, utterance mood and number of INTSINT tags in the intonational annotation. A series of alternate datasets with different feature sets were produced from the annotation information; the differences consist in alternate manners to represent utterance mood and dialogue act. The feature definition in the datasets is described below and Table 15 presents the complete feature set.

7.1. Dialogue act features

The DIME-DAMSL annotation information selected is the preliminary DIME-DAMSL of obligations and common ground, which is set on three nominal, i.e. non-numerical, features: obligations, agreement and understanding, each containing the DIME-DAMSL tags of the utterance corresponding to the respective plane. If an utterance is not assigned a tag on some of its planes, the *x* value (meaning *no-tag*) is used in order to avoid a feature to remain empty. Whenever an utterance is annotated with more than one tag on a plane, e.g. *action-dir* and *answer* on the obligations, or *affirm* and *reject* on the agreement, the feature value is represented as a concatenation of the two tags: *action-dir_answer* or *affirm_reject*; this manner is more useful than using two separated features because it allows keeping a fixed number of simple features despite the number of tags assigned.

Since agreement and understanding are two subplanes of the common ground plane, this is considered by implementing two datasets: one that contains common ground tagging as one single feature in which the tags of agreement and understanding are concatenated in one single string; the second dataset contains the two features separated. The two datasets are statistically analyzed to determine the feasibility to create machine-learning models.

7.2. Intonation annotation features

The implementation of intonational information features in the dataset is based on the criteria described as follows. Every speaker presents a particular speech rate (velocity), so the general shape of the intonational contour and the time between two consecutive tones are, in turn, determined by this factor. Differences among speech velocities might produce a series of shapes differing in length for one same utterance spoken by different speakers, which impedes a direct comparison of the shapes. Previous work in the area, such as Garrido (1991), Garrido (1996) and Shriberg et al. (1998) have addressed this issue by computing a normalization of the time information, i.e. adjusting the length of every shape by using a statistical rule that takes into account the maximum, the minimum and the average durations of utterances to produce a representation that keeps the original shape while also constrains its length within standardized limits. Thus, the normalization produces a discretized abstraction of the contour and the discreteness of the representation is valuable to search for patterns relating intonation with dialogue act.

Time normalization is not computed in this work because the representation implemented is an abstraction of the intonational contour at a higher level, described below. The duration of INTSINT tones along a contour, i.e. the time elapsed between pairs of inflection points, as well as the tones themselves, are the

components describing the general shape of the contour. An intonational pattern is the abstraction of this shape discarding the bias introduced by speech rate (velocity) of particular speakers. Like previous work in the area, such as Garrido (1991) and Garrido (1996), this work assumes that only a finite number of categories of intonational contours exist. Other relevant assumption, introduced by the present work, is that these categories can be represented as INTSINT tag sequences without their time information, i.e. without timestamps of the INTSINT tagging. By discarding the timestamps, a higher level view of the contour is preserved; therefore, at the end, the intonationl annotation is represented in the dataset as strings of INTSINT labels.

Garrido (1991, 1996) resume theoretical descriptions and empirical results from previous work in which the significant regions of intonational contours are addressed. The final region of the contour is considered as the most influence in determining the utterance mood and the speech act expressed by the utterance; the initial region also contributes but not as much as the final. Consequently, intonation is represented in the dataset by the INTSINT tagging corresponding to both the final and the initial region of the contour. As the lengths, i.e. the number of tags, of the initial and the final regions are not strictly determined, a series of substrings of the INTSINT annotation are considered: 10 nominal features are produced from this annotation, each as an alphabetic string of up to five labels using the last 5, 4, 3, 2 and 1 labels and the first 1, 2, 3, 4 and 5 of each INTSINT sequence.

Also the total number of INTSINT tags along a contour is considered as predictor and it is implemented as an integer numerical feature in the dataset. It seems evident that this is highly correlated to utterance duration, because the longer the duration, the more number of tones on the contour, and vice versa. This correlation is commented below.

7.3. Utterance duration features

Utterance duration, automatically measured from the speech signal, is evaluated as one of the predictors. It is implemented in the dataset as a floating-point numerical value in milliseconds.

An indirect measurement of the utterance duration is the number of INTSINT tags along the contour. This might be a measurement more reliable than a strictly temporal unit (*e.g.* millisecond) because the speech rate (i.e. speech velocity) might produce different durations of one same sentence uttered by different speakers; however, the number of INTSINT tones would be the same with high probability because the stylization process finds out the inflection points (the so-called *targets*) along the intonational contour and the number of targets is determined by phonologically significant *f0* changes rather than by the speaking rate.

7.4. Speaker role features

Although the role played by the speaker in a dialogue is not an intonational feature, it might contribute to the dialogue act recognition, so its possible contribution is evaluated and it is implemented as a nominal feature whose values can be *System* or *User*.

7.5. Utterance mood features

Garrido (1991, 1996) describe the relation between utterance mood and intonational contour, which shows a strong correlation. Therefore, utterance mood is included in the present dataset: first, as a manually annotated data and then as an automatically tagged feature. Its values are nominal, as follows: *interrogative*, *declarative*, *imperative* or *other*; in an alternate implementation of the dataset, interrogatives are, in turn, divided into *wh-question*, *yes–no-question* and (general) *interrogative*.

7.6. Statistical analyses

A volume of 1043 utterances was selected from the corpus. Prior to machine-learning modeling, statistical analyses are performed to know the distributions of the features and the possible correlations among them. Distributions are analyzed by computing absolute and relative frequencies, averages, histograms and Paretos of each feature in the dataset.

Table 8
Pareto analysis of utterance mood

Utt. mood	%	Accum. %
declarative	67.2	67.2
interrogative	25.2	92.4
other	6.2	98.6
imperative	1.4	100.0

Table 9
Pareto of dialogue act labels

Dialogue act label	%	Accum. %
accept	24.2	24.2
info-request	15.6	39.8
answer	14.5	54.3
action-dir	6.8	61.0
affirm	6.8	67.8
graph-action	6.4	74.2
commit	6.1	80.3
Other labels	19.7	100.0

Some general figures are presented here: the System utters 46.7% and the User 53.3% of utterances. Regarding utterance mood, Table 8 presents its Pareto analysis; *declarative* is the most frequent mood. Most of dialogue act annotations (80.3%) contain one or more of seven DIME-DAMSL tags, as described in Table 9. Statistics show that the occurrence of understanding dialogue acts is quite infrequent (3.5% approximately), so a recognition model for this plane is not feasible with the available data.

INTSINT annotation data presents a statistical distribution as follows: the average number of tones per utterance is 7.1, the maximum is 45, the minimum is 2 and the standard deviation 5.5. Considering that the last 2 INTSINT tags on the annotation are particularly relevant to the utterance mood recognition, an analysis of their combinations is performed and it shows that most of data (80.1%) are construed by 10 combinations: *BS*, *DB*, *UT*, *TS*, *TB*, *BT*, *MB*, *US*, *BH*, and *TL*.

The statistical distribution of utterance durations is: average, 2.1 s; maximum, 31.2; minimum, 0.096 and standard deviation, 2.7.

Correlations are studied by computing absolute and relative frequencies of the following pairs of features: dialogue act versus utterance mood (Figs. 6–8), dialogue act versus speaker role (Tables 10 and 11), utterance

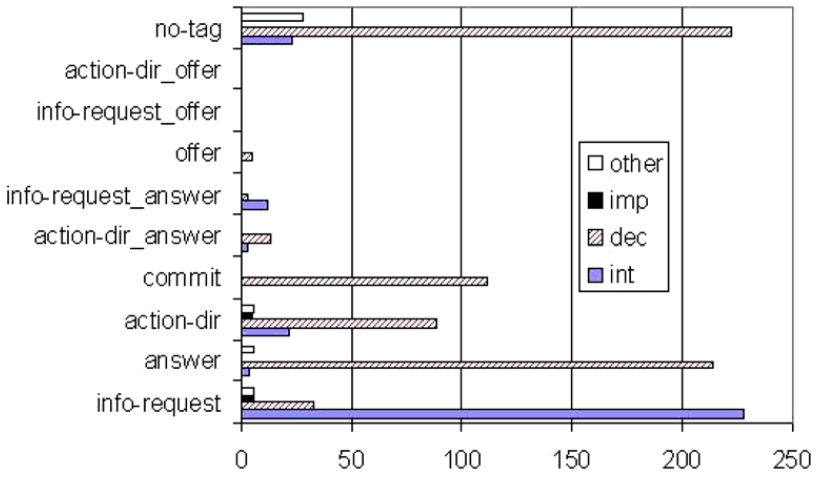

Fig. 6. Obligations dialogue acts per utterance mood.

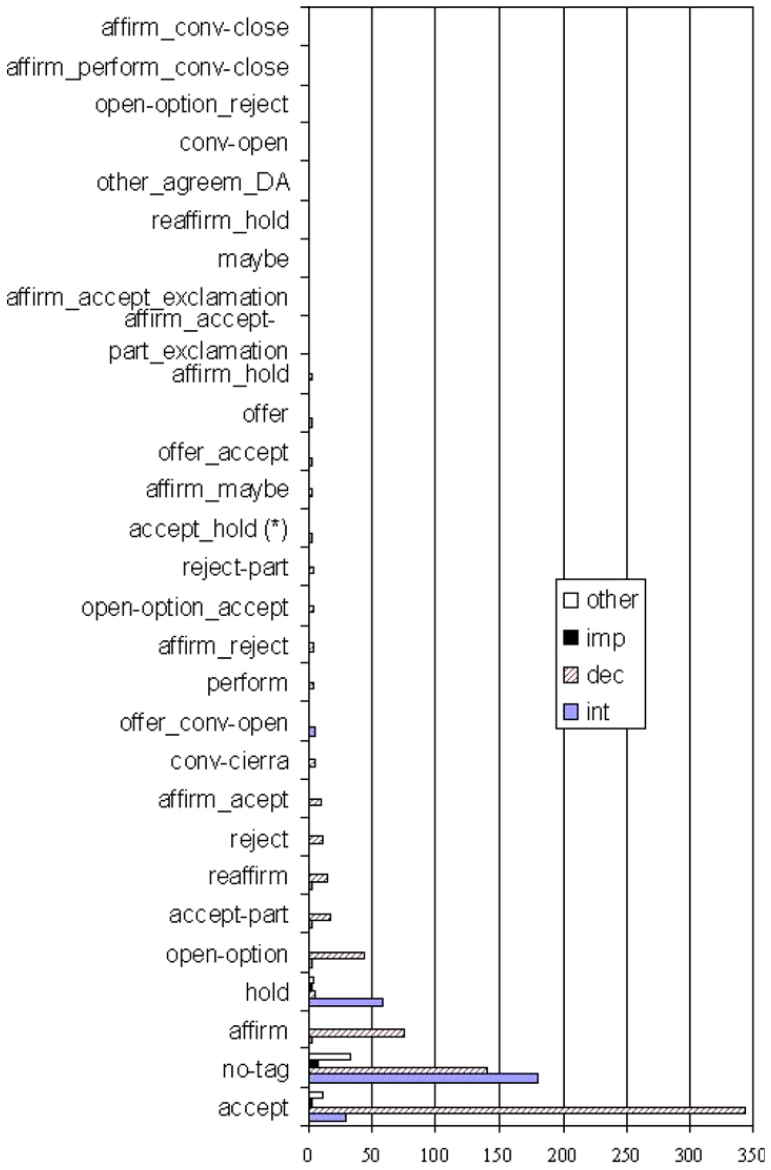

Fig. 7. Agreement dialogue acts per utterance mood.

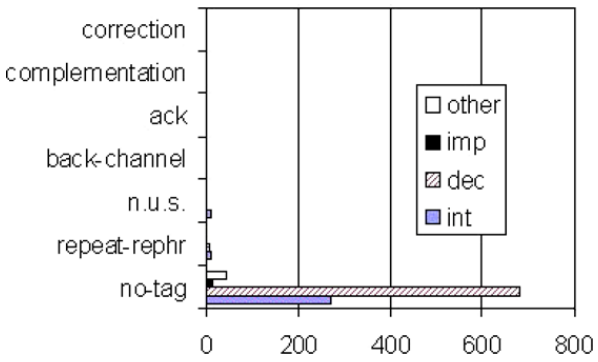

Fig. 8. Understanding dialogue acts per utterance mood.

mood versus speaker role (Table 12), and dialogue act on one plane versus the corresponding on the complementary plane (Tables 13 and 14).

In Figs. 6–8, although *declarative* is the most frequent utterance mood for most of dialogue acts; *interrogative* is the most frequent mood for some particular dialogue acts, such as *info-request* and some of its combinations (*info-request_answer*, *info-request_offer*, etc.), with 75% or higher; *hold* and its combinations (*reaffirm_hold*, *affirm_hold*, etc.), 84% or higher; and *offer* and its combinations (*offer_accept*, *offer_conv-open*, etc.), 100%. Dialogue acts corresponding to classes that do not belong to the agreement plane, i.e. the so-called *no-tag* in the agreement figure, occur as *interrogative* (49.9%) or *declarative* (39.1%). In the understanding plane, particularly for its most frequent label (*no-tag*), utterance moods are *declarative* (67.4%) or *interrogative* (26.8%).

The analysis of dialogue act versus speaker role (Tables 10 and 11) shows that some dialogue acts are more frequently expressed by one of the two roles, and even that some dialogue acts are never expressed by one of them. On the obligations plane, *info-requests* are more frequent in the *System* (72.2%), *answers* in the *User* (73.3%), *action-dirs* are always (100%) expressed by the User, and *commits* by the System (100%). On the agreement plane, *accept* is expressed by the System in 53.1% of utterances and by the User 46.9%; on the other hand, *affirm* is more frequent in the User, i.e. 71.8%; *hold*, in the System, 85.5%, *open-option* is always uttered by the System, 100%, and *offer* by the System, 100%.

By analyzing the correlation between utterance mood and speaker role (Table 12), results show that *declaratives* are used on a similar proportion, i.e. 43.9% by the System and 56.1% by the User; however, the remaining moods present different patterns: *interrogatives* are uttered 74.2% by the System; *imperatives*, 92.9% by the User and the *other* category, 92.9% by the User. This suggests that the role performed by the speaker influences the utterance mood.

As an instance of a dialogue act can occur in the obligations and common ground planes simultaneously, such a dialogue act is annotated with labels for both planes conventionally. Certain dialogue act types on each of the DIME-DAMSL planes show a clear statistical correlation to the corresponding tag on the complementary plane. Tables 13 and 14 present the most clear correlations from obligations to agreement plane and vice versa. For instance, all *offers* and most *commits* are related to *accepts*; also, some combinations of obligations, *e.g. action-dir_answer* is related to *accept-part*. From the agreement plane view, most of *holds* relate to *info-request*, and many *open-options* relate to *answers*. The relation to *no-tag* represents that the DA of the utterance has no presence on such plane.

Statistical results show patterns in the dataset that can be used to guide the implementation of the models and also broadly describe in advance how the features will be interacting in recognition models. Correlation results suggest that speaker role and the dialogue act type of the complementary plane can be useful to recognize the dialogue act type on one determined plane.

7.7. Experimental conditions

The main target feature is dialogue act type and the other features are evaluated as predictors. In addition, utterance mood also needs to be set as a special target data because this is a required input for dialogue act recognition. Therefore, a recognition model for utterance mood is implemented and its output is used as one of the inputs to the dialogue act model.

Since understanding DAs are so infrequent in the dataset, a simplification in this work consists of concatenating the understanding and agreement DA and labeling the composite DA with a single feature.

Table 15 presents the features implemented from the corpus data for experiments. The right-most column specifies if a feature is used as predictor (*P*), target (*T*) or both (*T*/*P*); the *T*/*P* value specifies that such feature is used as target in a particular model and as predictor in other. No lexical information feature is used in the experiments.

The predictors are selected on the basis of diverse linguistic theories and statistical analyses of the empirical data. The machine-learning algorithm evaluates the contribution of every predictor to the recognition task and might eventually discard any of them if it is not useful to the recognition.

Eight experimental conditions (see Table 15) are evaluated. In each condition, two models for dialogue act are created: one for obligations and one for common ground. The reason is that an utterance can simultaneously convey dialogue acts on both planes. Also, it is supposed that certain dialogue acts on a determined

Table 10
Correlation of dialogue act versus speaker role on the obligations plane

Obligations DA	System	User
info-request	72.2	27.8
answer	26.7	73.3
action-dir	0.0	100.0
commit	100.0	0.0

Table 11
Correlation of dialogue act versus speaker role on the agreement plane

Agreement DA	System	User
accept	53.1	46.9
affirm	28.2	71.8
hold	85.5	14.5
open-option	100.0	0.0
offer	100.0	0.0

Table 12
Correlation of utterance mood versus speaker role

Utt. mood	System	User
declarative	43.9	56.1
interrogative	74.2	25.8
imperative	7.1	92.9
other	7.1	92.9

Table 13
Correlation of obligations to agreement DAs

%	Obligations DA	Related to Agrmt. DA
100.0	*offer*	*accept*
97.3	*commit*	*accept*
81.3	*action-dir_answer*	*accept-part*
69.2	*action-dir*	*no-tag*
65.6	*info-request*	*no-tag*
62.5	*info-request_answer*	*no-tag*

Table 14
Correlation of agreement to obligations DAs

%	Agreement DA	Related to Obligs. DA
91.3	*hold*	*info-request*
68.1	*open-option*	*answer*
65.0	*accept-part*	*action-dir_answer*
64.7	*reaffirm*	*answer*

plane are associated to the acts occurring on the other plane. As the common ground plane is structured by the agreement and understanding subplanes, its modeling has to take this into account.

Conditions 1, 2 and 3 assume that certain non-intonational features, such as speaker role and the complementary dialogue act type can contribute to dialogue act recognition.

Conditions 4 and 8 analyze the contribution of the previous dialogue act as one of the predictors; i.e. the dialogue act type of the previous utterance to predict the type corresponding to a current utterance. The

Table 15
Features in the dataset

Feature	Description	Why it is Evaluated	***P*** or ***T***
first_1	The first INTSINT label of an utterance	The initial region of the intonational contour contributes to	*P*
first_2	The first two INTSINT labels of an utterance	utterance mood recognition; each of the three features is evaluated	*P*
first_3	The first three INTSINT labels of an utterance	to determine which is useful	*P*
last_2	The last 2 INTSINT labels of an utterance	Preliminary experiments show that it is highly contributive to utterance mood recognition because it contains the utterance toneme	*P*
utt_mood	The manually annotated utterance mood	It is related to intonational contour and perhaps to dialogue act	*P*
predicted_mood	It is obtained by a complementary recognition model before dialogue act recognition.	An automatically recognized mood instead of the manually annotated is used because this implementation is more similar to a real-world application	*T*/*P*
optimal_pred_mood	It is obtained by a complementary recognition model before dialogue act recognition. It performs better than *predicted_mood* by using other predictors	The recognition rate of dialogue act types is better than using *predicted_mood*	*T*/*P*
utt_duration	Utterance duration in milliseconds (it is *not* normalized)	Preliminary experiments suggest that it might contribute to the recognition of dialogue act type	*P*
number_of_tones	Number of INTSINT labels in the complete intonation tagging of the utterance	It is an abstraction of utterance duration and does not need a normalization process	*P*
speaker_role	Role of the speaker in the dialogue, i.e. *System* or *User*	Statistics suggest that *speaker_role* is correlated to dialogue act type; e.g. *System* to *commit*, *User* to *action directive*	*P*
obligations	Manually annotated tag of dialogue act type on the obligations plane	It is the target in the obligations model and one of the predictors in the common ground model	*T*/*P*
obligations_minus1	Dialogue act tag (manually annotated) of obligations in utterance *n*-1, where *n* is the utterance whose dialogue act type is the target	Its contribution as one of the predictors for dialogue act is evaluated	*P*
commgr	Manually annotated tag of dialogue act on the common ground plane; agreement and understanding tags are concatenated as one single feature	It is the target in the common ground model and one of the predictors in the obligations model	*T*/*P*
commgr_minus1	Dialogue act tag (manually annotated) of common ground in the utterance *n*-1, where *n* is the utterance whose dialogue act is the target	Its contribution as one of the predictors for dialogue act is evaluated	*P*
pred_obligations	Automatic tagging for obligations	It is contrasted to manual tagging	*T*/*P*
pred_commgr	Automatic tagging for comm. gr.	It is contrasted to manual tagging	*T*/*P*
pred_obligations_minus1	*pred_obligations* of *n*-1	It is contrasted to manual tagging	*T*/*P*
pred_commgr_minus1	*pred_commgr* of *n*-1	It is contrasted to manual tagging	*T*/*P*

previous dialogue act is represented as two features: *obligations_minus1* and *commgr_minus1*. The experimental condition assumes that the dialogue act type of one of the two planes is already known by using any other method, such as manual annotation, another machine-learning algorithm, POS-language modeling, etc. Condition 8.1 uses the manually annotated DA. Condition 8.2 evaluates the utility of automatically generated taggings for both obligations and common ground of the previous utterance; such tagging is produced by tree models in 8.1.

Condition 5 focuses on intonational features of the utterance. Theories on Spanish intonation state that it is closely related to utterance mood and that it is also associated to certain dialogue act types, so the real (manually annotated) utterance mood is included as one of the predictors instead of INTSINT taggings in a number of these models. In a real-world application, utterance mood needs to be recognized from the speech signal, so a model is generated to produce an automatically annotated mood and then this output is used instead of the real mood as one of the inputs for dialogue act models.

Condition 6 evaluates the performance if using the complementary plane and the utterance mood and duration. This condition aims to determine the effect of utterance mood and duration over the performance of condition 2, i.e. when using the complementary plane only.

In *Condition 7* the model for *optimal_pred_mood* evaluates the contribution of *speaker_role*; this is justified because of a statistical correlation between *speaker_role* and *utt_mood*. Other model evaluates the contribution of *speaker_role* and also the complementary dialogue act tag. The complementary dialogue act is *commgr* for *obligations* and *obligations* for *commgr*, which are the manually annotated from the corpus.

Most of machine-learning models in this work are created by using 10-fold cross-validation as training-testing method, except for models of: *predicted_mood*, *optimal_pred_mood*, *pred_obligations* and *pred_commgr*, in which models previously created on a cross-validation basis are supplied with data sets arranged to automatically recognize the targets. In the 10-fold cross-validation the input dataset is automatically split into 10 subsets by the machine-learning toolkit to perform 10 tests, as described below: in a first round, a combination of 9 subsets is used to train the model and the remaining subset to test it; in the second round, a second combination of 9 subsets (in which one is different from the first 9) is used to train and other subset (different from the first) to perform the second test; then, a third combination of 9 subsets, etc. until 10 rounds are completed and the 10 subsets are alternately used for training and testing. For every experimental condition, average *accuracy* and average *Kappa* of 10 test results are presented. This method is highly reliable and frequently used to create recognition or classification models.

Machine-learning techniques state a series of evaluation criteria; one of the basic is accuracy. *Accuracy* acceptability must determined for the particular domain in which a model is implemented, so a strict threshold value cannot be stated; for instance, some authors suggest 0.75 or 0.80, but this is not mandatory. *Kappa* statistics Carletta (1996) has been used for the evaluation of coding systems but also in the assessment of machine-learning models. Since *Kappa* measures the annotation's consistency, its application to supervised machine learning assumes that the model output is produced by an annotator and the annotations in the training dataset are produced by other annotator (as they actually are). This way, *Kappa* measures how consistent the model is (or not) with the actual annotations taking into account the contribution of chance, defined by the number of labels available in the tagset, the number of annotated instances and the number of (both manual and automatic) annotators. *Kappa* statistics for supervised machine-learning models is usually computed by a series of machine-learning toolkits such as WEKA (Witten and Frank, 2005). Discussions in the research community have existed about an acceptance threshold for *Kappa*; a frequent opinion is that annotations are consistent if *Kappa* is greater than or equal to 0.7.

Other criterion to evaluate machine-learning models is computing *recalls*, *precisions* and *F* measures of the respective classes on the basis of their corresponding confusion matrices. *Accuracy* is assumed as the percent of instances correctly classified by the model during the test. *Recall* is the ratio of instances belonging to a class which are classified into such class by the model; *precision* is the ratio of instances classified into a class by the model which truly belong to such class. *F* measure is another statistics commonly used in machine learning to evaluate the recognition rate for each class of the target data; it is computed with the formula: $F = 2 * (Precision * Recall)/(Precision + Recall)$.

Once a classification tree is generated and evaluated, the equivalent set of *if–then* rules is extracted; for instance, a rule extracted from an obligations tree is:

`IF commgr=accept AND speaker_role=system AND optimal_pred_mood=declarative, THEN commit`, meaning that if the tag on the common ground is *accept* and the speaker role is System and the utterance mood is *declarative*, then the obligations dialogue act is *commit*. In decision tree models, *if–then* rules are construed by two elements: premise and conclusion; the premise is the set of Boolean conditions associated by *AND* operators before the *THEN*; the conclusion is a value for the target after the *THEN*. A subset of the dataset instances might satisfy the premise and, a sub-subset in such subset might satisfy both the premise and the conclusion. The utility of every rule is weighted by two key figures: *support* and *confidence*; *support* is a ratio computed as the number of instances satisfying the premise divided by the total number of instances in the dataset. *Confidence* is another ratio: the number of instances that satisfy both the premise and the conclusion divided by the number of cases satisfying the premise only. *Support* of the example rule is 182/1043 = 17.4% and *confidence* is 103/182 = 56.6%. *Support* represents how frequently the combination of values of predictor features occurs in the dataset; *confidence* measures how reliable the rule is.

7.8. Machine-learning results

Once the dataset was prepared and formated to be compatible with WEKA, a series of classification trees for dialogue act and for utterance mood was produced. The complete trees along with their complete confusion matrices are available on the web.[2] The results for each experimental condition are presented in Table 16. This includes the percents of majority classes in the manually annotated dataset, which are presented as baselines. Significance tests are performed using McNemar's scores (Everitt, 1977) on the χ^2 distribution. The three rules with highest support in the most interesting trees are presented in a web document.[3]

The predictors presented are those that were identified by J48 algorithm as contributing to recognize the corresponding target; i.e. J48 automatically discards input features that do not contribute to the target recognition.

McNemar's score has been evaluated by a number of authors, such as Dietterich (1996), Salzberg (1997) and Demsar (2006) and their results suggest that it is one of the most suitable methods to compare machine-learning classifiers. It is a non-parametric method used on nominal data to determine whether the marginal frequencies of two data sets are equal. The score allows us to determine if the accuracies of two classifiers are significantly different from each other. It is applied to a contingency table (a 2×2 matrix), whose 4 cells contain the counting of instances misclassified by: (1) both classifiers, (2) by the first but not by the second, (3) by the second but not by the first and (4) by none of the them. For a given pair of classifiers, A and B, the score is computed as $\chi^2 = (|s - f| - 1)^2/(s + f)$, where χ^2 is a chi-squared statistic with 1 degree of freedom, s is the number of instances misclassified by A but not by B, and f is the number of instances misclassified by B but not by A. The number of degrees of freedom is 1 because the contingency table has 2 rows and 2 columns. The result is used to find out the probability (p) associated to the null hypothesis, H_0, on the chi-squared distribution. H_0 supposes that the two classifiers have the same error rate, involving that $s = f$. Thus, if p is less than a threshold value (usually 0.05), then it can be stated that the accuracies of the two classifiers are significantly different from each other. The testing was set by defining pairs of models that present a higher interest for comparison, instead of analyzing all possible pairs. The fact that two models are significantly different to each other, involves that their respective sets of predictor features have different classification capacities.

7.8.1. Conditions 1, 2 and 3

In condition 1, *speaker_role* by itself is a poor predictor for DA, despite the statistical correlation between this feature and particular types of DA. In condition 2, the complementary plane performs significantly better than *speaker_role*. Condition 3 is even better because it takes advantage of the combined contributions of the two features. Although the *accuracies* are higher than baselines in the three experimental conditions, the *Kappas* are low. In conditions 1, 2 and 3, the accuracies of the most comparable models are significantly different from each other, as can be seen in the significance tests in Table 17.

[2] http://www.unsis.edu.mx/~coria/elsevier_appendix_1/.
[3] http://www.unsis.edu.mx/~coria/elsevier_appendix_2/.

Table 16
Experiment results contrasted to majority class percents of manual annotation

Experimental condition	Model	Target	Features	Acc. (%)	K	Baseline (majority class %)	Majority class	Nr. of classes
(1) Speaker role	1.1 A	*obligations*	*speaker_role*	29.9	0.1	26.2	*no-tag*	13
	1.1 B	*commgr*	*speaker_role*	35.4	0.0	36.7	*accept*	45
(2) Complementary plane	2.1 A	*obligations*	*commgr*	53.7	0.4	26.2	*no-tag*	13
	2.1 B	*commgr*	*obligations*	57.4	0.4	36.7	*accept*	45
(3) Speaker role + Compl. Plane	3.1 A	*obligations*	*commgr*, *speaker_role*	64.1	0.6	26.2	*no-tag*	13
	3.1 B	*commgr*	*obligations*, *speaker_role*	66.0	0.6	36.7	*accept*	45
(4) Spkr role + Compl. + Previous turn	4.1 A	*obligations*	*commgr*, *speaker_role*, *obligations_minus1*, *commgr_minus1*	71.1	0.7	26.2	*no-tag*	13
	4.1 B	*commgr*	*obligations*, *speaker_role*, *obligations_minus1*, *commgr_minus1*	69.9	0.6	36.7	*accept*	45
	5.1 A	*obligations*	*first_3*, *last_2*, *utt_duration*	32.6	0.2	26.2	*no-tag*	13
	5.1 B	*commgr*	*utt_duration*	43.7	0.2	36.7	*accept*	45
	5.2 A	*obligations*	*utt_mood*, *utt_duration*	47.9	0.3	26.2	*no-tag*	13
(5) Intonational only	5.2 B	*commgr*	*utt_mood*, *utt_duration*	47.1	0.4	36.7	*accept*	45
	5.3	*predicted_mood*	*first_2*, *last_2*, *utt_duration*	74.3	0.4	67.2	*dec*	4
	5.4 A	*obligations*	*predicted_mood*, *utt_duration*	39.0	0.2	26.2	*no-tag*	13
	5.4 B	*commgr*	*predicted_mood*, *utt_duration*	50.6	0.3	36.7	*accept*	45
(6) Compl. plane + mood + duration	6.1 A	*obligations*	*commgr*, *utt_mood*, *utt_duration*	60.6	0.5	26.2	*no-tag*	13
	6.1 B	*commgr*	*obligations*, *utt_mood*, *utt_duration*	63.1	0.5	36.7	*accept*	45
(7) Intonational + non-intonational	7.1	*optimal_pred_mood*	*last_2*,*first_2*, *speaker_role*, *utt_duration*	77.4	0.5	67.2	*dec*	4
	7.2 A	*obligations*	*commgr*, *speaker_role*, *optimal_pred_mood*, *utt_duration*	66.3	0.6	26.2	*no-tag*	13
	7.2 B	*commgr*	*obligations*, *speaker_role*, *optimal_pred_mood*, *utt_duration*	66.9	0.6	36.7	*accept*	45
	8.1 A	*obligations*	*speaker_role*, *commgr_minus1*, *obligations_minus1*, *commgr*, *optimal_pred_mood*, *first_3*, *number_of_tones*	73.4	0.7	26.2	*no-tag*	13
(8) Intonational + non-intonational + previous turn	8.1 B	*commgr*	*commgr_minus1*, *obligations*, *speaker_role*, *obligations_minus1*, *utt_duration*	70.3	0.6	36.7	*accept*	45
	8.2 A	*obligations*	*speaker_role*, *pred_commgr_minus1*, *pred_commgr*, *optimal_pred_mood*, *utt_duration*, *pred_obligations_minus1*	66.6	0.6	26.2	*no-tag*	13
	8.2 B	*commgr*	*pred_commgr_minus1*, *pred_obligations*, *speaker_role*, *pred_obligations_minus1*, *utt_duration*, *optimal_pred_mood*	71.8	0.6	36.7	*accept*	45

Table 17
Significance test using McNemar's scores

Obligations		Common ground	
Comparing models	Probability	Comparing models	Probability
1.1A *vs.* 2.1.A	$p < 0.05$	1.1B *vs.* 2.1B	$p < 0.05$
2.1A *vs.* 3.1.A	$p < 0.05$	2.1B *vs.* 3.1B	$p < 0.05$
3.1A *vs.* 4.1.A	$p < 0.05$	3.1B *vs.* 4.1B	$p < 0.05$
4.1A *vs.* 5.1.A	$p < 0.05$	4.1B *vs.* 5.1.B	$p < 0.05$
5.1A *vs.* 5.2.A	NOT $p < 0.05$	5.1B *vs.* 5.2.B	$p < 0.05$
5.2A *vs.* 5.4.A	$p < 0.05$	5.2B *vs.* 5.4.B	NOT $p < 0.05$
5.4A *vs.* 6.1.A	$p < 0.05$	5.4B *vs.* 6.1.B	$p < 0.05$
6.1A *vs.* 7.2.A	NOT $p < 0.05$	6.1B *vs.* 7.2.B	NOT $p < 0.05$
7.2A *vs.* 8.1.A	$p < 0.05$	7.2B *vs.* 8.1.B	$p < 0.05$
8.1A *vs.* 8.2.A	$p < 0.05$	8.1B *vs.* 8.2.B	NOT $p < 0.05$

7.8.2. Condition 4

The combination of speaker role, complementary plane and previous turn features produces models in which *accuracies* are approximately 70% and *Kappas* are greater than 0.6. However, since the complementary plane and the previous turn are manually annotated tags, this is only a theoretical condition. The accuracies of models produced in condition 4 are significantly different from those in conditions 3 and 5 (see Table 17).

7.8.3. Condition 5: Intonational-only features as predictors

Results suggest that using intonational annotation and duration features only (models 5.1), produces a low recognition rate of DA. In model 5.1 A, *obligations*, all *F* values are less than 0.4 (13 classes). In 5.1 B, *commgr*, *F* measures are 0.7 for *accept* and 0.5 for *no-tag*; all the remaining 43 classes are less than 0.2. The best recognition rates in this experimental condition are obtained in model 5.2; this uses the actual (manually tagged) utterance mood. However, according to significance tests, models 5.1A and 5.2A are *not* significantly different from each other. A similar situation occurs for 5.2B and 5.4B. Model 5.2 suggests that utterance mood contributes to the task.

A particular analysis was performed for utterance mood recognition, so two models are created and evaluated: in the first, the tagset for utterance mood provides three different labels for interrogatives: *yes-no question*, *wh-question* and (general) *question*; the third label is used for any case in which neither *yes–no* nor *wh-* are suitable. In the second model, one only label, i.e. *interrogative*, is used for any type of interrogatives. Results show that using one single *interrogative* class (*accuracy* = 74.3%, *Kappa* = 0.4) performs better than using three (*accuracy* = 72.3%, *Kappa* = 0.4) and model 5.3 is obtained by using one single interrogative class. The contributing features in model 5.3 are *first_2*, *last_2* and *utt_duration*. The predicted utterance mood is used as one of the inputs to model 5.4. The final region of the intonational contour, i.e. the last 2 INTSINT tags of the intonational representation, is confirmed as highly contributing to the recognition of utterance mood, as stated by theories on Spanish intonation, such as Navarro-Tomás (1948), regarding the rising contour (e.g. *UT* and *TS*) of a number of interrogative moods and the falling or flat contour (e.g. *BS*, *DB* and *TB*) of declaratives. In addition, the recognition task is improved by adding the information of the initial region (the first 2 INTSINT tags), which is also consistent with prior theoretical claims.

Contrasting the Kappa values of conditions 1 (speaker only) and 5, condition 5 seems to perform better. However, the performance of condition 1 is approximately equal to that of models 5.1, in which the features are INTSINT annotations and utterance duration. Interestingly, utterance duration is the only feature needed for common ground recognition in 5.1B; i.e. common ground dialogue acts can be distinguished among each other by just observing their utterance duration (e.g. acknowledgments are shorter than affirms). The main difference between 5.1 and 5.2 is that 5.2 uses utterance mood, while 5.1 does not, so utterance mood is a useful feature for the task. Table 18 presents the *F* measures, recalls and precisions of target classes of model 5.3.

Results for utterance mood recognition are taken into account to implement mood models as well as dialogue act models in conditions 6 and 7.

Table 18
F measures, recalls and precisions of predicted utterance mood (model 5.3)

Utterance mood	*F*	Recall	Precision
dec	0.8	0.9	0.8
int	0.6	0.6	0.6
F = *recall* = *precision* = 0.0 in *imp* and *other*			

7.8.4. Condition 6: Complementary plane + utterance mood + utterance duration

By comparing these results to those of condition 2, these seem to perform more accurately. Improvements are: 6.9% and 5.6% points for accuracy, and 0.1 and 0.1 for *Kappa*, on the obligations and common ground planes, respectively. As *Kappa* increments are approximately equal to 0, the improvement introduced by utterance mood and duration is not relevant. The significance analyses for condition 6 are discussed in the Condition 7 section.

7.8.5. Condition 7: Intonational + non-intonational features

The target data in model 7.1 is the so-called *optimal_pred_mood* and it uses *last_2*, *first_2*, *speaker_role* and *utt_duration* as predictors. Other features are automatically discarded by J48 algorithm. *Accuracy* is 77.4% and *Kappa* is 0.5. Table 19 presents the *F* measures, *recalls* and *precisions*. These results are obtained when considering one single interrogative mood; i.e. one single *interrogative* class is used. On the other hand, if the three interrogative classes are considered, *accuracy* and *Kappa* are lower: 74.6% and 0.4, respectively. The term *optimal* is used to distinguish model 7.1 from the simpler predicted mood in model 5.3, where *accuracy* and *Kappa* are lower.

Table 20 presents the confusion matrix of the model for optimal predicted mood (7.1), where 2 out of 4 classes (*other* and *imp*) cannot be recognized. This might be explained by their low frequency in the dataset. *Dec* is clearly recognized, while *int* is inaccurate and can be confused with *dec*.

In model 7.2, the maximum accuracy for obligation recognition is obtained when using these predictors: *optimal_pred_mood*, *utt_duration*, *commgr* (using agreement and understanding as a concatenation in one single feature) and *speaker_role*. The most accurate model for common ground uses: *optimal_pred_mood*, *utt_duration*, *obligations* and *speaker_role*.

Tables 21 and 22 present the *F* measures, *recalls* and *precisions* of the obligations and common ground DA models.

As statistical analyses show that certain utterance moods and certain dialogue acts are more frequent in one of the two speaker roles, speaker role as one of the predictors improves *accuracies* and *Kappas* of the classification trees for utterance mood as well as for dialogue act. However, its influence is a contingent phenomenon because it is determined by the settings and protocols for the creation of the corpus. Despite this, it is expected that practical dialogues in other domains show a similar relation among speaker role, dialogue act and utterance mood.

The classification trees show that the particular dialogue act tag on any plane might constrain the possible dialogue act tags on the complementary plane; therefore, the tagging of a DIME-DAMSL plane as one of the predictors improves accuracies and *Kappas* of the dialogue act recognition on the complementary plane. This experimental condition involves the assumption that the tagging of one of the two planes has already been determined.

Table 19
F measures, recalls and precisions of optimal predicted utterance mood (model 7.1)

Utterance mood	*F*	Recall	Precision
dec	0.8	0.9	0.8
int	0.6	0.5	0.8
F = *recall* = *precision* = 0.0 in *imp* and *other*			

Table 20
Confusion matrix of optimal predicted mood (7.1)

a	*b*	*c*	*d*	→classified as
158	131	2	0	*a* = *int*
41	649	1	1	*b* = *dec*
4	42	0	0	*c* = *other*
0	14	0	0	*d* = *imp*

Table 21
F measures, recalls and precisions for obligations dialogue acts (model 7.2A)

Obligations	*F*	Recall	Precision
action-dir_answer	0.8	0.8	0.8
info-request	0.8	0.7	0.9
answer	0.7	0.8	0.7
commit	0.7	0.9	0.6
action-dir	0.7	0.8	0.5
no-tag	0.5	0.4	0.6
F = *recall* = *precision* = 0 in the remaining classes			

Table 22
F measures, recalls and precisions for common ground dialogue acts (model 7.2B)

Common ground	*F*	Recall	Precision
accept	0.8	0.8	0.8
no-tag	0.7	0.8	0.7
accept-part	0.7	0.7	0.7
open-option	0.7	0.9	0.5
hold_repeat-rephr	0.6	0.6	0.6
n.u.s.	0.5	0.5	0.6
affirm	0.4	0.4	0.4
F = *recall* = *precision* = 0 in the remaining classes			

Like condition 5, condition 7 suggests that pure intonational information does not contribute highly to DA recognition in the obligations plane; this is demonstrated by comparing to an alternate model where INTSINT annotations are included as predictors in addition to those of the main model, obtaining lower *accuracy* and *Kappa*: 62.7% and 0.5, respectively.

In order to compare the contribution of pure intonational features in the common ground model, an alternate model is created using INTSINT tagging among the predictors, obtaining an accuracy equal to 62.9% and *Kappa* equal to 0.5, which are less than the corresponding to model 7.2B.

As intonational information, especially *last_2*, is the most contributing predictor for *optimal_pred_mood*, it can be stated that this target is an indirect representation of intonation, so intonation can be ranked below the fourth most contributing predictor to dialogue act recognition in condition 7.

A similar experimental condition is addressed in (Coria and Pineda, 2006), where preliminary results are obtained from one single dialogue with only two speakers (one as User and one as System). Results from it and from the present work suggest that intonational information features are highly contributing to utterance mood recognition, as expected on a theoretical basis; however, *accuracies* and *Kappas* for dialogue act recognition in preliminary results are greater than in the present work (12 dialogues). A major reason for this is a larger number of speakers in *User* role, who provide a broader variety of intonational contours, utterance moods and dialogue acts.

Significance analyses for models 7.2 (both obligations and common ground) involve comparing to 6.1 and 8.1. In the former, neither 6.1A is significantly different from 7.2A, nor 6.1B differs from 7.2B. Nevertheless, models 7.2 does differ significantly from models 8.1.

7.8.6. Condition 8: Intonational + non-intonational + previous turn

Condition 8 considers two scenarios: 8.1, where the complementary plane and the previous turn are manually annotated features, and 8.2, where those features are automatically assigned. Both scenarios use *optimal_pred_mood* from model 7.1.

The previous dialogue act is represented by using both obligations and common ground features from the $(n-1)$th utterance, where n is the current; however, 8.2 uses the automatically defined taggings.

In trees 8.1A and B, for obligations and common ground, *accuracies* and *Kappas* are the maximal of all 8 experimental conditions: 73.4% with 0.7 and 70.3 with 0.6, respectively; however, as they are generated by using manually annotated features for complementary DA and previous turn, these values should be interpreted as ideal upper boundaries. Another scenario is presented in models 8.2 A and B, whose results are 66.6% with 0.6 and 71.8% with 0.6, respectively. However, these two models are also theoretical as they are based on models 8.1, which use the manually annotated DA. Taking this into account, a series of considerations for real-world implementations is suggested below. As expected, 8.2 A performs worse than 8.1 A because 8.2 A is biased by inaccuracies in automatic tagging of complementary DA and previous turn. The significance analysis for this pair of models confirms the accuracy difference. On the other hand, 8.2 B performs marginally better than 8.1 B; this might be explained by a reduction in the number of common ground classes generated by the model to produce the automatic *pred_commgr* annotation, where 45 classes are reduced to 18. Such reduction impacts also the *pred_commgr_minus1* feature; thus, an artificial improvement is produced. However, the significance analysis suggests that the accuracies of 8.1B and 8.2B are not different to each other.

Table 23 presents F measures, *recalls* and *precisions* of model 8.1A, where 6 out of 13 classes present $F \geqslant 0.7$. In addition, $0 < F < 0.7$ in 4 classes, and $F = 0$ in 3 classes (*graph-action*, *offer_info-request* and *action-dir_offer*).

Table 24 presents F measures, recalls and precisions of model 8.1B, where 8 out of 45 classes have $F \geqslant 0.7$. Besides, $0 < F < 0.7$ in 9 classes, and $F = 0$ in 28 classes: (1) accept_hold_repeat-rephr, (2) affirm_accept-part_exclamation, (3) repeat-rephr, (4) offer_accept, (5) affirm_accept_exclamation, (6) back-channel, (7) hold_on_task-mngmt, (8) affirm_reject, (9) graph-action_accept, (10) reject, (11) maybe, (12) reaffirm_hold, (13) affirm_display, (14) open-option_display_accept, (15) display, (16) reject-part, (17) open-option_accept, (18) offer, (19) other, (20) reaffirm_complementation, (21) conv-open, (22) affirm_graph-action, (23) affirm_correction, (24) affirm_hold, (25) open-option_reject, (26) affirm_perform_conv-close, (27) hold_NUS and (28) affirm_conv-close. This involves that most of classes cannot be accurately recognized; however, 8 are so. Twenty of the non-recognized classes are multiple-label tagged and they present a very low frequency in the dataset; these two reasons might explain their low recognition rate. The presence of DAs with multiple-label tagging increases the number of classes because each label combination constitutes a class in itself; the complexity of the recognition task depends, in part, on the number of classes to be recognized because the machine-learning algorithm has to learn the pattern describing each class.

Table 25 presents the confusion matrix of model 8.1A. The five most frequent classes in the obligations model include 86.8% of the dataset instances, in which the worst classified class is *action-dir*, frequently confused with *info-request* or *no-tag*.

The complete matrix of 8.1B is too large (45 rows and 45 columns) to be presented in this paper; however, it is available at the web[4] and its most salient aspects are resumed in Table 26. The six most frequent classes include 83% of the instances; *affirm* is the worst classified class, frequently confused with *no-tag*, *accept* or *open-option_display*.

7.9. Application of the results in dialogue management systems

The results can be applied to dialogue management systems as described below: first of all, in a real-world implementation the DA from the complementary plane is not available, so the models suggested in the present work need to be supported by additional resources. A feasible technique is incorporating lexical information

[4] http://www.unsis.edu.mx/~coria/elsevier_appendix_1/condition_8/model_8_1_b.pdf, pp.101–103.

Table 23
F measures, recalls and precisions for obligations (8.1A)

Obligations	*F*	Recall	Precision
info-request_graph-action	0.9	0.9	0.9
info-request_graph-action_answer	0.9	0.9	0.9
answer	0.9	0.9	0.9
commit	0.8	0.9	0.8
offer	0.8	0.8	0.8
action-dir_answer	0.8	0.8	0.7
no-tag	0.7	0.6	0.7
info-request	0.7	0.7	0.6
action-dir	0.5	0.5	0.5
info-request_answer	0.2	0.1	0.5
$F = recall = precision = 0$ in the remaining classes			

Table 24
F measures, recalls and precisions for common ground (8.1B)

Common ground	*F*	Recall	Precision
graph-action	0.9	0.9	0.9
offer_conv-open	0.8	1.0	0.7
accept	0.8	0.8	0.8
open-option_display	0.8	0.9	0.7
reaffirm	0.7	0.6	1.0
accept-part	0.7	0.7	0.7
no-tag	0.7	0.8	0.7
hold_repeat-rephr	0.7	0.9	0.6
affirm_maybe	0.7	0.5	1.0
affirm	0.4	0.4	0.4
perform	0.3	0.3	0.5
conv-close	0.3	0.2	0.3
hold	0.2	0.2	0.5
open-option	0.2	0.2	0.3
affirm_accept	0.2	0.1	0.5
n.u.s.	0.1	0.1	0.3
ack	0.1	0.0	0.1
$F = recall = precision = 0$ in the remaining classes			

Table 25
Confusion matrix of obligations model (8.1A)

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	← classified as
119	3	33	1	17	2	0	0	0	0	0	0	0	*a = info-req*
3	192	1	0	27	0	1	1	0	0	0	0	0	*b = answer*
34	3	64	0	15	0	3	0	1	0	0	0	0	*c = action-dir*
4	0	1	97	10	0	0	0	0	0	0	0	0	*d = commit*
26	21	23	29	174	0	0	0	0	0	0	0	0	*e = no-tag*
2	0	0	0	0	95	0	0	0	0	1	0	0	*f = info-req_graph-action*
0	0	2	0	1	0	13	0	0	0	0	0	0	*g = action-dir_answer*
1	3	0	0	2	0	1	1	0	0	0	0	0	*h = info-req_answer*
0	1	0	0	0	0	0	0	4	0	0	0	0	*i = offer*
0	0	0	0	1	0	0	0	0	0	0	0	0	*j = graph-action*
0	0	0	0	0	1	0	0	0	0	7	0	0	*k = info-req_graph-action_answer*
0	0	0	0	0	0	0	0	0	0	0	0	1	*l = offer_info-req*
0	0	0	0	0	0	0	0	0	0	0	1	0	*m = action-dir_offer*

Table 26
Confusion matrix analysis of the most frequent classes in common ground model (8.1B)

Comm. Gr. DA	Correctly classified (%)	Confused with other classes (%)
graph-action	99.0	*no-tag (1.0)*
open-option_display	85.4	*affirm* (9.8), *accept* (2.4), *affirm_accept* (2.4)
hold_repeat-rephr	85.2	*no-tag* (7.4), *accept* (5.6), other tags (1.8)
accept	83.3	*no-tag* (9.4), *affirm* (2.1), *hold_repeat-rephr* (1.8), other tags (3.4)
no-tag	79.0	*accept* (8.1), *affirm* (5.2), *hold_repeat-rephr* (2.9), other tags (4.8)
affirm	35.6	*no-tag* (24.7), *accept* (24.7), *open-option_dispay* (9.6), *open-option* (2.7), other tags (2.7)

analysis by DA language models (LM), in which every DA class is represented by one or more statistical LM trained on *part-of-speech* (POS) tags. Previous work in the area, such as Shriberg et al. (1998), suggest that, in general, the composite implementation of decision trees and LM performs better than LM alone for DA recognition, so a similar configuration should be useful for this work.

Decision trees can take advantage of features such as intonation, speaker role, complementary DA, DA of the previous turn, utterance mood, and utterance duration, so the trees can increase the accuracy rate for classes in which LMs alone are not so accurate. For the composite implementation, the most accurate tree models can be exploited; i.e. those to recognize obligations and common ground either by using speaker role, previous turn and complementary DA without intonational data, or by using all features including intonational. In turn, only the tree rules for DA classes with highest reliability should be considered; a suitable threshold value for reliability should be determined by empirical analyses.

A real-world system might involve the application of models in two stages: (1) LM and (2) decision trees. In the first stage, LMs produce a DA tagging for each of the two DIME-DAMSL planes; reliability rates of the two labels are known from the prior LM training. The two reliabilities are compared and the most reliable tagging should be selected. Depending on the plane of the most reliable tagging, the *if-then* rule set of the other plane is used in the second stage; i.e. if the LM tagging with higher reliability is for obligations, the common ground *if–then* rules are used in the second stage; otherwise, the opposite applies. The second stage cannot be performed always but only when there exists a rule such that it can use the automatic tagging from the first stage as complementary plane feature.

For some particular DA classes, LM alone might perform better than the composite configuration; but the opposite might occur for other classes. Therefore, in order to implement an optimal dialogue management system, a prior training and evaluation of the composite configuration is needed; i.e. the composite model should be used to produce automatic taggings on the training dataset and reliabilities should be computed for every DA class. Reliabilities of the composite models should be compared to those of LM-only configurations. The most reliable model for each class, either LM-only or composite, is selected to implement the system and the remaining models are discarded.

The DA of a current turn should be temporally stored by the management system because it is used as previous turn for prediction purposes.

8. Conclusions and future work

This paper has presented an initial analysis on a feature set from the DIME corpus to recognize dialogue acts by using machine-learning techniques on intonational information and other sources. Using DIME-DAMSL for DA annotation, and INTSINT for intonation, a series of experimental conditions are evaluated; each condition includes separate models for obligations and common ground dialogue acts. Models for utterance mood are also evaluated because statistical analyses suggest a correlation between this feature and dialogue act.

With baselines equal to 26.2% for obligations, 36.7% for common ground, and 67.2% for utterance mood, the most remarkable results are that utterance mood can be recognized with $accuracy = 77.4\%$ and $Kappa = 0.5$ by using the last 2 and the first 2 INTSINT tones of the intonational tagging along with the speaker role and the utterance duration. One of the two conditions providing the most accurate DA recogni-

tion rates uses no intonational features at all but instead speaker role, complementary DA plane and DA of the previous turn. The other condition uses several types of features, including intonational. The performance of the non-intonational model is comparable to the performance of that with all features: $Accuracy = 70.5$ with $Kappa = 0.6$ (averages) versus 71.9 with 0.6 (averages), respectively. Alternative models with all features for obligations and common ground recognition in which both the complementary plane and the previous turn are semi-automatically assigned instead of using the manual tagging are also evaluated. The semi-automatic setting suggests what would be achievable in a quasi-realistic system; however, as the semi-automatic annotation of the complementary and the previous DA are based on manual annotations, the models should be considered theoretical. The alternative model for obligations performs worse than the primary (-6.8% points in *accuracy* and -0.1 in *Kappa*, approximately), as expected, because it is biased by inaccuracies in automatic tagging. On the other hand, the alternative model for common ground performs marginally better ($+1.5\%$ points in *accuracy* and no change in *Kappa*); this might be explained by a reduction in the number of common ground classes generated by the model to produce the automatic annotation of complementary plane and previous turn, in which 45 common ground classes are reduced to 18.

In conclusion, general results suggest that the contribution of intonational information to DA recognition, without considering lexical information, depends on interactions among a series of elements of the dialogue context, such as the speaker role and the previous dialogue act. Hence, intonational information alone does not guarantee a successful recognition. An explanation for this might be that a requirement to recognize DA from intonational information (without other sources) would be that every DA would have to present only one intonational contour and vice versa; however, this does not occur in reality. According to statistical correlation analyses, a short number of DA types present a typical mood (e.g. information request with interrogative mood, or affirm with declarative mood); in addition, one utterance mood can be used to express any among several DA types. In addition, tree models that use intonational features only provide recognition rates lower than trees that also include non-intonational.

Using only intonational features (i.e. INTSINT annotations from both the beginning and the end of the intonational contour, and utterance duration) for DA recognition, provides accuracy rates less than 50% and *Kappas* less than 0.2. The recognition rates improve using additional non-intonational features. Since statistical analyses suggest a correlation between utterance mood and dialogue act type, the contribution of utterance mood for dialogue act recognition is also addressed, confirming its utility. Adding the real (manually annotated) utterance mood as feature improves the recognition rate; however, this improvement presents two issues: (1) the recognition rate is still low (less than 57%) and (2) in a real-world application the real mood is unknown and it has to be automatically recognized by a model. In order to solve the second problem, models for automatic annotation of utterance mood are also created and evaluated.

The utterance mood models are consistent with theoretical foundations on Spanish intonation, which state that the *tonema* (toneme), i.e. the final region of the intonational contour, determines significantly the utterance mood and that the initial region of the contour provides a marginal contribution. The recognition rate is higher if one single interrogative class is used in the annotation tagset for utterance mood instead of using more labels to annotate diverse classes of interrogatives. In one of the experimental conditions, an optimized model for utterance mood is implemented by using a non-intonational feature (i.e. speaker role). This configuration is based on prior statistical analyses, in which a correlation between speaker role and utterance mood is observed. The output of this tree is used as one of the inputs to feed DA models. Results for DA recognition by using two different mood features are compared: (1) manually annotated mood and (2) automatically annotated mood on the basis of INTSINT tagging and utterance duration. The DA recognition rate from the automatic annotated mood is less than from the manually annotated.

The statistical analyses suggested certain patterns relating speaker role and DA, so this has been included in one of the feature sets, which has increased the recognition rates. Nevertheless, the speaker role contribution to the recognition of both utterance mood and dialogue act is a contingent phenomenon because it is determined by the particular recording conditions of the corpus. Despite this, it can be expected that in other domains certain utterance moods and dialogue acts are more highly expected from one of the speaker roles than from the other.

Statistical correlations between taggings on the two DIME-DAMSL planes have been observed; i.e. certain *obligations/common ground* pairs are more frequent than others in certain stages of the transactions. These

pairs include instances in which a certain component occur without a complementary dialogue act on the other plane (*no-tag*). Such correlations suggest the existence of an interaction phenomenon; i.e. the DA component on obligations influences the common ground component, and vice versa. Although the existence of these correlations is not a solid evidence for causal relation, the complementary DA tagging has been evaluated as one of the features in exploratory models. The underlying hypothesis is that once the tagging on any plane is known, it can support the DA recognition for the other plane. This experimental condition improves the recognition rate over those that do not use the complementary DA.

An experimental condition using speaker role and complementary DA (i.e. common ground tagging as one of the predictors for obligations, and obligations as one of the predictors for common ground), provides *accuracies* greater than 66.2% and *Kappas* greater than 0.5.

A broader scope in the dialogue context is evaluated by including the previous DA as an input to the recognition task. This experimental condition produces the highest recognition rates. For obligations, *accuracy* is 73.4% and *Kappa*, 0.7; for common ground, *accuracy* is 70.3% and *Kappa*, 0.6.

Regarding practical applications, dialogue management systems and DA automatic annotation tools can take advantage of the recognition rules for classes with highest *F* measures in the trees. The rules cannot be used alone but along with part-of-speech language models for specific DA classes, like implemented in previous work in the area. Automatic annotation tools for utterance mood can also exploit the rules for mood classes with highest *F* measures.

This investigation is a first attempt to study the relation between intonation and DA in Spanish using the DIME corpus as an empirical resource. Larger data volumes should be analyzed, including a larger number of speakers. Also, other intonational representations should be investigated. Whenever a larger volume of the DIME corpus annotations is available a new, statistically balanced, dataset could be created and all DA classes could be present in equal or similar amounts.

The interaction phenomena on the two DIME-DAMSL planes should be investigated as well. Some questions to answer are: what are the most frequent *obligations/common ground* pairs? what are the most common contexts in which particular pairs are present? why does this interaction occur?

Finally, future work should evaluate DA recognition by using agreement and understanding tagging as two separated features and targets. Other experimental scenario can be an integration with a complementary model, such as a series of POS-based DA language-models. In such composite configuration, it is expected that intonational and speaker role features would improve recognition rates over that of POS-only models. Furthermore, one of the experimental conditions should use the language models outputs to automatically annotate the *pred_obligations_minus1* and *pred_commgr_minus1* features.

Acknowledgements

This work was supported by the National Council for Science and Technology (CONACyT) of Mexico and the Directorate for Post-Graduate Studies (DGEP) at the National Autonomous University of Mexico (UNAM). The authors thank valuable comments from James Allen (University of Rochester, NY, USA), Joaquim Llisterri (Universitat Autonoma de Barcelona, Spain), Katya Rodriguez (UNAM), Elizabeth Shriberg (International Computer Science Institute, Berkeley, CA, USA) and Christian Lemaitre (Universidad Autónoma Metropolitana, Mexico City). We also thank the collaboration of the DIME Project team, particularly comments by Varinia Estrada.

References

Allen, J.F., Core, M., 1997. Draft of DAMSL: Dialogue Act Markup in Several Layers. Technical Report, The Multiparty Discourse Group. University of Rochester, Rochester, USA.

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A., 2000. An architecture for a generic dialogue shell. Natural Language Engineering 6 (34), 213–228.

Ang, Jeremy, Liu, Yang, Shriberg, Elizabeth, 2005. Automatic dialog act segmentation and classification in multiparty meetings. In: Proceedings of ICASSP.

Beckman, M., Ayers-Elam, G., 1997. Guidelines for ToBI Labelling (version 3.0, March 1997). The Ohio State University Research Foundation.

Beckman, Mary E., Díaz-Campos, Manuel, Tevis-McGory, Julia, Morgan, Terrell A., 2002. Intonation across Spanish, in the Tones and Break Indices framework, Probus 14, Walter de Gruyter, pp. 9–36.

Boersma, Paul, Weenink, David, 2007. PRAAT, a system for doing phonetics by computer (Version 5.0) [Computer Program]. http://www.praat.org/ (retrieved January 15, 2007).

Breiman, Leo, Friedman, H. Jerome, Olshen, R.A., Stone, Charles J., 1983. Classification and Regression Trees. Wadsworth.

Bunt, H., 1994. Context and dialogue control. THINK Quarterly.

Bunt, H., 1995. Dynamic interpretation and dialogue theory. In: Taylor, M.M., Neel, F., Bouwhuis, D.G. (Eds.), The Structure of Multimodal Dialogue. John Benjamins, Amsterdam.

Campione, E., Hirst, D., Véronis, J., 2000. Automatic stylisation and symbolic coding of *f0*: implementations in the INTSINT model. In: Botinis, A. (Ed.), Intonation: Analysis, Modelling and Technology, Text, Speech and Language Technology, vol. 15. Kluwer Academic Publishers, Dordrecht, pp. 185–208.

Carletta, Jean, 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22 (2), 249–254.

Core, Mark, 1997. Addendum to coding dialogs with the DAMSL annotation scheme. In: AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA, USA, pp. 28–35.

Core, Mark, 1998. Homework for DRI Group on Forward and Backward-looking Functions. <http://www.cs.rochester.edu/research/cisd/resources/damsl/results.html>.

Coria, Sergio, Pineda, Luis, 2006. Predicting dialogue acts from prosodic information. In: Seventh International Conference on Intelligent Text Processing and Computational Linguistics, CICLing, Mexico City, Mexico.

Cuétara, Javier O., 2004 Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla. Tesis para obtener el título de Maestro en Lingüística Hispánica. Maestría en Lingüística Hispánica, Posgrado en Lingüística, Universidad Nacional Autónoma de México. México, D.F.

Dahlback, Niels, Jonsson, Arne, Ahrenberg, Lars, 1993. Wizard of Oz studies: why and how. Knowledge-based Systems 6 (4), 258–266.

Demsar, Janez, 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

Dietterich, T., 1996. Statistical Tests for Comparing Supervised Learning Algorithms. Technical Report, Oregon State University, Corvallis, OR.

Espesser, Robert, 1999. MES: Motif Environment for Speech (software). <http://www.lpl.univaix.fr/ext/projects/mes_signaix.htm>.

Everitt, B., 1977. The Analysis of Contingency Tables. Chapman and Hall, London.

Fernández, Raúl, Picard, Rosalind W., 2002. Dialog act classification from prosodic features using support vector machines. In: Proceedings of Speech Prosody 2002. Aix-en-Provence, France.

Garrido, Juan María, 1991. Modelización de patrones melódicos del español para la síntesis y el reconocimiento. Depto. de Filología Española, Universitat Autònoma de Barcelona, Barcelona, Spain.

Garrido, Juan María, 1996. Modelling Spanish Intonation for Text-to-Speech Applications. Doctoral Dissertation, Department of Spanish Filology, Facultat de Lletres, Universitat Autònoma de Barcelona Barcelona, Spain.

Godfrey, J., Holliman, E., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, vol. 1. San Francisco, CA, USA, 1992, pp. 517–520.

Hirst, Daniel, DiCristo, Albert, Robert, Espesser, 2000. Levels of representation and levels of analysis for the description of intonation systems. In: Horne, M. (Ed.), Prosody: Theory and Experiment. Kluwer, Dordrecht.

Hirst, D.J., 2007. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In: Proceedings of the 16th International Congress of Phonetic Sciences, 6–10 August, 2007, Saarbrücken, Germany, pp. 1233–1236.

Janin, A. et al., 2003. The ICSI Meeting Corpus. In Proceedings of ICASSP.

Jekat, Susanne, Klein, Alexandra, Maier Elisabeth, Maleck, Ilona, Mast, Marion, Quantz, J. Joachim, 1995. Dialogue Acts in VERBMOBIL, VM-Report 65.

Jurafsky, D., Shriberg, E., Fox, B., Curl, T., 1998. Lexical, prosodic, and syntactic cues for dialog acts. In: Proceedings of the ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal, Canada, August 1998, pp. 114–120.

Kompe, R., Kiebling, A., Niemann, H., Noth, E., Schukat Talamazzini, E., Zottmann, A., Batliner, A., 1995. Prosodic scoring of word hypotheses graphs. In: Pardo, J.M., Enríquez, E., Ortega, J., Ferreiros, J., Macías, J., Valverde, F.J. (Eds.), Proceedings of the 4th European Conference on Speech Communication and Technology, vol. 2. Madrid, Spain, pp. 1333–1336.

Léon, P.R., Faure, G., Rigault, A. (Eds.), 1970. Systématique des fonctions expressives de l'intonation. Analyse des faits prosodiques. Didier, Ottawa, pp. 57–72.

Llisterri, Joaquim (Ed.), 1996. Prosody Tools Efficiency and Failures. LRE Project 62-050 MULTEXT. WP 4 Corpus. T4.6 Speech Markup and Validation. Deliverable 4.5.2. Final version. October 15, 1996.

Mast, M., Kompe, R., Harbeck, S., Kiealing, A., Niemann, H., Noth, E., 1996. Dialog act classification with the help of prosody, Philadelphia, USA, In: International Conference on Spoken Language Processing, vol. 3, pp. 1728–1731.

Moreno, Iván, Pineda, Luis, 2006. Speech repairs in the DIME corpus. Research in Computing Science 20, 63–74.

Navarro-Tomás, T., 1948. Manual de entonación española. Guadarrama, Madrid, Spain, 1948.

Navy Research Laboratory, 2001. Speech in Noisy Environments, <http://elazar.itd.nrl.navy.mil/spine/>.

Oregon Health & Science University (OHSU), 2004. CSLU Toolkit (software).

Pineda, L., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E., Villaseñor, L., 2002. The Dime Project. In: Proceedings of MICAI 2002, Lecture Notes in Artificial Intelligence, vol. 2313. Springer-Verlag, pp. 166–175, ISSN 0302-9743.

Pineda, L., Castellanos, H., Coria, S., Estrada, V., López, F., López, I., Meza, I., Moreno, I., Pérez, P., Rodríguez, C., 2006a. Balancing transactions in practical dialogues. In: CICLing 2006, Seventh International Conference on Intelligent Text Processing and Computational Linguistics. Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City.

Pineda, L.A., Estrada, V.M., Coria, S.R., 2006b. The obligations and common ground structures of task oriented conversations. In: Proceedings of the Fourth Workshop in Information and Language Technology TIL-2006, in IBERAMIA 06, Brazil.

Pineda, L.A., 2007. Department of Computer Science, Institute of Applied Mathematics and Systems. National Autonomous University of Mexico. Mexico City. <http://leibniz.iimas.unam.mx/luis/DIME/CORPUS-DIME.html>.

Pineda, L.A., Estrada, V., Coria, S., Allen, J., 2007. The obligations and common ground structure of practical dialogues. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial 11 (36), 9–17.

Quilis, Antonio, 1981. Fonética acústica de la lengua española. Gredos (Biblioteca Románica Hispánica, Manuales 49), Madrid, España.

Rangarajan, Vivek, Bangalore, Srinivas, Narayanan, Shrikanth, 2007. Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. In: Proceedings of Interspeech, Antwerp, Belgium, 2007.

Rossi, Mario, DiCristo, Albert, Hirst, Daniel, Martin, P., Nishinuma, Y., 1981. L'intonation, de l'acoustique à la sémantique, París, Klincksieck.

Salzberg, Stefan L., 1997. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Mining and Knowledge Discovery 1. Kluwer Academic Publishers, Boston, USA.

Shriberg, Elizabeth, Bates, Rebecca, Stolcke, Andreas, Taylor, Paul, Jurafsky, Dan, Ries, Klaus, Coccaro, Noah, Martin, Rachel, Meteer, Marie, Van EssDykema, Carol, 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? USA, Language and Speech 41(3–4) 439-487 (Special Issue on Prosody and Conversation).

Silverman, K., Blaauw, E., Spitz, J., Pitrelli, J.F., 1992. Towards using prosody in speech recognition/understanding systems: differences between read and spontaneous speech. In: Proceedings of the Fifth DARPA Workshop on Speech and Natural Language.

Sosa, J.M., 1999. La entonación del español. Su estructura fónica, variabilidad y dialectología. Cátedra (Lingüística), Madrid.

Sosa, J.M., 2003. La notación tonal del español en el modelo Sp-ToBI. In: Prieto, P. (Ed.), Teorías de la entonación, Barcelona: Ariel (Ariel Lingüística). pp. 185–208.

'T. Hart, R., Collier, A., Cohen, A., 1990. Perceptual Study of Intonation An Experimental-Phonetic Approach to Intonation, Cambridge University Press, Cambridge.

Thorsen, Nina, 1979. Interpreting Raw Fundamental-frequency Tracings of Danish. Phonetica 36, 57–78.

Thorsen, Nina, 1980. A study of the perception of sentence intonation – evidence from Danish. Journal of the Acoustical Society of America 67(3), 1014–1030.

Tur, Gokhan, Guz, Umit, Hakkani-Tür, Dilek, 2006. Model adaptation for dialog act tagging. In: Proceedings of SLT 2006, 1st biannual IEEE/ACL Workshop on Spoken Language Technologies, Aruba, December 2006.

Wahlster, W.,1993. VERBMOBIL: translation of spontaneous face-to-face dialogs. In: Proceedings of 3rd EUROSPEECH, Berlin, Germany, pp. 29–38.

Wilson, Deirdre, Sperber, Dan, 1988. Mood and the analysis of non-declarative sentences. In: Dancy, J., Moravcsik, J., Taylor, C. (Eds.), Human agency, Language, Duty and Value. Stanford UP, Stanford CA.

Witten, I., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, CA, USA.

Venkataraman, A., Ferrer, L., Stolcke, A., Shriberg, E., 2003. Training a prosody-based dialog act tagger from unlabeled data. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 1, pp. 272–275.