The *DIME* Project*

L. A. Pineda, A. Massé, I. Meza, M. Salas, E. Schwarz, E. Uraga, L. Villaseñor
Department of Computer Science

Institute for Applied Mathematics and Systems (IIMAS), UNAM

Abstract. In this paper a general description and current state of the project *Diálogos Multimodales Inteligentes en Español (DIME) –Intelligent Multimodal Dialogs in Spanish–* is presented. The purpose of the project is to develop a multimodal conversational agent with spoken input and output facilities in Spanish in a design oriented domain: kitchen design. In this paper, the state of the project, current results, an overview of the prototype system and future work are presented.

1 Introduction

In this paper, the current state of the project *DIME: Diálogos Multimodales Inteligentes en Español –Intelligent Multimodal Dialogs in Spanish–* is presented. The purpose of the project is to develop a multimodal conversational agent with spoken input and output facilities in Spanish in a design oriented domain: kitchen design.

Here, we focus on the three main current tasks of the project: the first consists in the compilation, transcription and tagging of a multimodal *corpus* in a design domain; the second is the development of a Spanish speech recognition system specialized in the language employed in the application domain; the third is the definition of a Spanish grammar and parser for Spanish, comprehensive enough to interpret the lexicon and grammar observed in the *DIME* corpus. In this paper the results of these three modules are presented, and also a brief description of the current state of the prototype system.

This project is developed in a collaboration between the Department of Computer Science, at the Institute for Applied Mathematics and Systems (IIMAS) of the National University of Mexico (UNAM), the ITESM, Campus Morelos, and the Department of Computer Science of the University of Rochester, with the support of the bilateral program between the United States and Mexico NSF/CONACyT for the development of computer science.

2 The *DIME Corpus*

The first task of the project was the collection of a corpus on the application domain: The *DIME* corpus. We required an application domain that, on the one

* A previous version of this paper in Spanish was present at the SLPLT2 workshop that was held in Jaén, Spain, in September, 2001.

hand, was general enough to be accessible and potentially interesting to many people, meriting the use of natural language and, on the other, simple enough to be modelled with current computational technology. In addition, in order to study multimodal communication, it was required to have a domain in which the use of spatial references, either supported by explicit gestures or not, were abundant. Also, we intended to study intentions or speech acts that were expressed through gestures, as commonly happens in multimodal communication, supporting the case for a domain in which it was natural to use an interactive graphics interface. An application domain meeting these requirements seemed to be kitchen design. This is a simple and intuitive task that can be performed by many people, and yet, there are a number of design rules, known by expert designers, that can help to improve the quality and functionality of the designs, and the presence of a design assistant can be justified for this task. In summary, kitchen design was thought to be a kind of task that can be assisted through intelligent multimodal agents, and it was chosen for the project.

As first step we proceeded to develop the *DIME* corpus. It was collected with the help of a Wizard of Oz scenario[12]. In our experiment, subjects were not under the illusion that they were interacting with a real system, but the restrictions placed on the interface had an influence on the language used and the kind of gestures performed by both the wizard and the subject. The wizard, in addition, was instructed to limit his expressivity, and to take the initiative and provide help only in case he noted that some constraints or kitchen design rules were violated or ignored by the user, resembling better the expected behavior of a real system.

To carry on with the experiment, a laboratory with two small chambers was built. This environment allowed the subject and the wizard to solve kitchen design tasks in a collaborative way, yet avoiding visual contact. The task was supported by specialized hardware and software through which all events, both graphical and linguistic, were recorded. This is illustrated in Figure 1.

The graphical interface used in the experiments is shown in Figure 2. For the construction of this environment, a commercial product was used[3]. Further references about this setting can be found in [12].

Each session consisted in an explanation given to the user, a system demonstration and the solution of two tasks; the first was designed mainly to familiarize the user with the environment, and the second was a more comprehensive design task. The experiment was applied to 16 different subjects, in addition to some complementary sessions. From this effort, we obtained 31 dialogs, with a total of 27,459 token words, with an average of 886 words per dialog, 5779 utterance (185 per dialog), 3606 turns (115 per dialog) and 7:10 hours of recordings (14 minutes per dialogue).

The linguistic data obtained through the experiments is used to develop the language models of the speech recognition system, to guide the development of the Spanish lexicon and grammar, and to identify and model the speech acts that occur in the application domain, which will be used in the construction of the pragmatics interpretation and dialogue manager modules of the system.

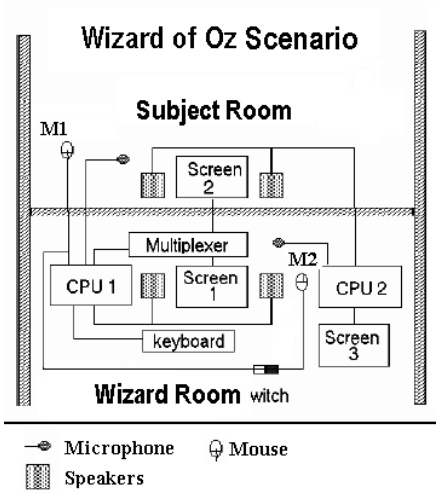

Fig. 1. Setting for the Wizard of Oz experiment

The dialogues were first segmented and transcribed ortographically; however, as the purpose of the project is to build a system for continuous, but not spontaneous speech, the dialogues were simplified eliminating interjections, noises, long pauses, stutters, speech-repairs, simultaneous speech, interruptions, etc. As a result of this process, two versions of the corpus are available: for spontaneous and continuous speech. The corpus is available for the research community [12].

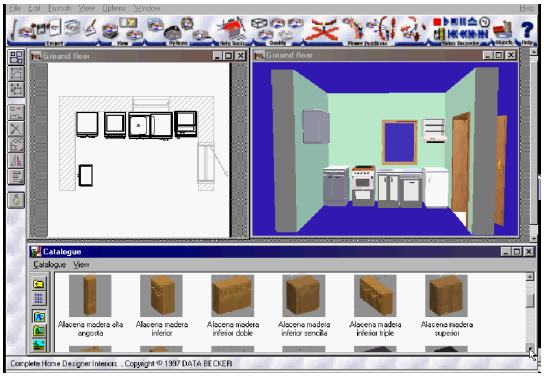

Fig. 2. Wizard of Oz interface

3 Speech recognition

In the context of the project, a module for speech recognition in Spanish focused on the application domain is being developed. In particular, we are developing acoustic-phonetic models, pronunciation dictionaries and language models for the ASR system using the *CSLU Toolkit*[5] and the *HTK*[6] system. With the purpose to create the pronunciation dictionaries, a phonetic alphabet for Mexican Spanish, *Mexbet,* was defined[9]. A set of rules for grapheme to morpheme conversion, including allophonic variation, for automating the process of phonetic transcription and the creation of pronunciation dictionaries, was also defined.

As an initial stage, a set of acoustic models based on neural nets and hidden Markov models was developed. These models were trained with three different architectures, the first two with the CSLU Toolkit and the last with HTK; the first architecture employed a neural net only, the second a hybrid model in which the output nodes of a neural network were initialized with the help of a set of hidden Markov models, and the third was based on hidden Markov models exclusively. As a language model a finite state grammar including word sequences normally found in telephone transfer calls was used. This domain was used to evaluate the first set of models. The performance of these models for a continuous, multilocutor speech, with a vocabulary of 107 words, was satisfactory, as the recognition level for the three architectures were 96.79%, 92.55% and 91.09% respectively.

The acoustic models were later used to recognize speech in the kitchen design domain. For this purpose, the models were evaluated with the *DIME* corpus. In this evaluation, a bigram based language model, trained with the *DIME* corpus for both continuous and spontaneous speech, was used. However, as the *DIME* corpus contains a number of allophonic variation not considered in the original acoustic models, the results of this evaluation were poor. For this reason, a new set of acoustic models was trained using the HTK toolkit and the Tlatoa corpus[8]. The results of the new evaluation were 25.21% and 20.82% of recognition at the word level for the continuous and spontaneous speech respectively. Currently, and with the aim to improve the recognition levels in our system, a new phonetically rich and balanced corpus is being developed; with this new corpus, it will be possible to create acoustic models for the 625 context dependent phonetic combinations (i.e., each phone has a left and right context) that we have identified in Mexican Spanish.

4 Spanish grammar and parser

One of the main objectives in the *DIME* project is to create a Spanish grammar and a parser robust enough to deal with the grammatical phenomena and lexica that appear in the continuous version of the *DIME* corpus. As a theoretical framework for developing this grammar *Head-driven Phrase Structure Grammar* (HPSG)[16] and its associated developing environment LKB[2] were adopted.

Independently of the similarities between English and Spanish, there are several syntactic phenomena particular to Spanish that pose interesting challenges

to computational syntactic theories. In particular, the Spanish clitic pronoun system has no direct counterpart in English; also, although the auxiliary verb systems of Spanish and English are similar in several respects, they are by no means identical, and there are a number of subtle properties of the Spanish grammar that give rise to ambiguities that do not occur in English, but have to be dealt with in computational applications. In addition, Spanish's verb morphology is richer than English, allowing the omission of subjects, a phenomena that does not occur in English. Also the linear word-order of English is much more strict than the word-order of Spanish, making the interpretation of Spanish a very hard task. For all these reasons, the definition of a Spanish grammar and parser was thought to be one of the main objectives of this project.

Currently we have focused on two of these phenomena: the auxiliary verb system and the clitic system. The Spanish auxiliaries pose a number of interesting challenges: *puedes mostrar* (can show) is a verbal phrase formed by an auxiliary verb (*poder*) and an infinitive, where the subject of the auxiliary verb indicates the agent of the non-personal form; however and unlike English, where auxiliary verbs have almost been deprived of their original lexical meanings, in Spanish some forms are still used in their original sense, in addition to their grammatical roles as auxiliaries, producing a number of interesting ambiguities. For instance, in *puedes con las matemáticas (you are able to do mathematics)* the verb *poder (to be able to)* preserves its original meaning of capacity, unlike English where the form *be able to* is preferred to in this latter sense, and it is not an auxiliary, as shown by our analysis[13].

The clitic system is also very important the Spanish language. In general, clitics are atonic grammatical units, mostly function words, that are attracted by tonic units, usually words with semantic content, forming a single lexical unit. An important kind of clitics occur in conjunction with pronouns; for instance, *me* in the word *mostrarme,* is an enclitic pronoun attached to the infinitive form of the verb *mostrar.* These kind of words occur often in periphrastic constructions in conjunction with auxiliary verbs as in *¿puedes mostrarme el catálogo?* (*can you show me the catalog?*). Here, the natural intuition for Spanish speakers, which is reflected in the orthography, is that these two lexical items, of different syntactic categories, form a single word, whose grammatical category is a verbal phrase in which one of the verbal arguments, the indirect object *me*, is already included. However, in the equivalent sentence *¿me puedes mostrar el catálogo?*, the pronoun occurs before the verbal phrase as an independent lexical form. Here, the distance of the proclitic pronoun *me* to the periphrasis is much larger than the one in the enclitic case and it is perceived as an independent lexical unit, as reflected by the Spanish orthography.

In our analysis both sentences receive a similar syntactic and semantic analysis. In the enclitic case, the periphrasis takes the verbal phrase *mostrarme el catálogo* as a complement of *puedes.* In this case it is possible the reading in which *poder* is an auxiliary verb and indicates the possibility of performing the action of showing. Additionally, our analysis permits also the reading in which

poder has the sense of capacity and it is not an auxiliary; in these latter case *puedes* requires an agent, the one who has the capability of showing.

In the proclitic case, *puedes* takes as its complement the verbal phrase *mostrar el catátogo*. In this phrase the verb *mostrar* needs its indirect object, which is represented by the clitic *me*. However, the verb *mostrar* can not take the clitic as its complement, because it is not behind the verb, but in front of the whole periphrasis. Therefore, it is the verb *puedes* which takes the clitic and shares it with *mostrar*. In this latter case both readings, capacity and possibility, are also possible.

Within the context of the *DIME* project, we have developed a systematic analysis for these problems[13, 14], and the interaction of these two grammatical systems, currently being developed, will permit the interpretation of a large number of sentences of the *DIME* corpus.

5 Speech act analysis

The first level of multimodal interpretation outputs the semantic representation of the natural language expressions input by the user, and the output of the graphical parser which interprets the graphical events expressed by the user through the interface. This information is passed to the semantic interpretation process which is responsible for reference resolution, both anaphoric and indexical, and also for determining the intentions expressed by the user; intensions can be simple and can be expressed by a single utterance, or can be complex requiring several utterances, and even several turns, to be expressed and understood along the conversational process. The literal interpretation of *¿me puedes mostrar el catálogo?* is a question asking whether the system has the possibility or the ability to show the catalog, but in the context of our application domain, it is rather an imperative statement commanding the system to show the catalog. To make this inference is part of the job of the pragmatic interpretation component of the system.

One important assumption in the development of this kind of systems is that the set of intentions that can be expressed during task oriented dialogues is finite and small, and can be characterized through a task analysis. In kitchen design, for instance, users can express the intention of including an object in a particular design, change the properties or relations of a number of objects, or simply to remove an object from the design; however, these simple intentions are normally expressed in the context of more complex intentional structures. In our dialogues, for instance, it is common to observe that when a user asks for the inclusion of a piece of furniture, the assistant requires to clarify the desired position before the action takes place; these two intentions form intentional units. The identification of such intentional structures is the purpose of the task analysis, and the result of this process is the *intentional structure* of the application domain. In *DIME*, the intentional structure is specified as the representation of the primitive intentions of the domain with their structural relations; the identification of such primitives and structural relations are obtained empirically out of the corpus.

With the purpose to analyze and characterize the intentional structure of our application domain, the tagging scheme for task oriented dialogs *DAMSL* was adopted[15]. In this scheme, speech acts are analyzed in four orthogonal dimensions, namely, the communication level, the information level, and the forward and backward relations of each expression. However, as the continuous version of the *DIME* corpus is used for this task, it is assumed that there are not communication problems and different information levels; consequently, the analysis for only the last two dimensions of DAMSL was required. Also, the scheme was modified and extended to deal with the peculiarities of the domain, and the multimodal aspect of the project. The analysis of a small segment of the corpus for just one conversational turn, with the forward and backward relations of the user and system utterances, is shown in Table 1.

In the context of the project, some graphical actions are performed with the purpose to express an intention and hence are considered speech acts. For this reason, *DAMSL* was extended considering that graphical actions performed by the system with a communicative intent, as a response to intentions expressed by users, are considered speech acts. In Table 1, for instance, utt29 expresses a forward *information request* speech act that needs to be attended to by the system. In utt30, the system establishes two backward and one forward functions; the *accept* and *acknowledgment* backwards speech acts take notice and accept the user request, and the forward *commit* speech acts is a charge to the discourse model that needs to be discharged before the intentional cycle is terminated. The *answer* speech act in the backward function of utt31 discharge the commit of *utt30,* and the *assert* forward function, which is realized graphically, satisfied the information request initiating the intentional transaction. In the example at hand, the system chose to strengthen its response by providing a *reassert* speech act, showing the information requested also in a textual form. This last expression closes the intentional cycle. The dialogue is modelled as a sequence of conversational cycles of this kind, as shown in the rest of Table 1. A formal model for this kind of interaction is currently being developed.

In order to test the adequacy of the tagging scheme, the *Kappa* coefficients[1] that measure the agreement of different taggers working on the same corpus, were used in a pilot test. According to the first results, the tagging definitions and procedures in *DAMSL* manual[15] were modified, and a new version, similar to the one reported in Chiba[4], but one in which graphical speech acts were contemplated, was defined.

In our current proposal, dialogues are also analyzed at three levels of granularity: the micro, messo and macro level. Table 1 shows the micro level in which the dialogue is segmented into speaker turns and utterance tokens within turns. The messo level consists in a dialogue structure in which so-called common ground units (CGU) are identified. A CGU contains all and only the tokens needed to add some bit of content to the common ground between the conversational participants. CGUs require some initiating material by one participant presenting new content and some feedback or acknowledgment by the other participant. If a token adds new content and there is no accessible ungrounded CGU

Forward	Backward	U#	S	Utterance (Spanish)	Utterance (English)
info-req		29	u:	¿me puedes mostrar el catálogo?	can you show me the catalog?
commit	accept ack	30	s:	ok	ok
assert	answer	31	s:	graphical action	
reassert	answer	32	s:	hay un catálogo de alacenas, uno de estantes, uno de estufas y extractores, uno de fregaderos y máquinas lavatrates	there is a catalog of cupboards, one of shelves, one of stoves, one of extractors and one of sinks and washing machines
info-request		33	u:	¿me puedes mostrar el catálogo de fregaderos y máquinas lavatrates?	can you show me the catalog of sinks and washing machines?
commit	accept ack	34	s:	ok	ok
assert	answer	35	s:	graphical action	
reassert	answer	36	s:	éste es el catálogo de fregaderos y máquinas lavatrastes	this is the catalog of sinks and washing machines
info-request		37	s:	¿quieres alguno en particular?	do you want a particular one?
info-request	hold	38	u:	¿a ver cuál es la diferencia entre el tercero y el cuarto?	let's see, what is the difference between the third and the fourth one?
check	accept-part	39	s:	¿entre éste y éste?	between this and this?
assert	answer	40	u:	si	yes
assert	answer	41	s:	éste tiene una superficie más larga que sale del mueble y éste abarca solamente hasta donde llega la máquina lavatrastes	this has a surface that gets off the furniture and this only covers up to where the dish washing machine gets to
action-dir		42	u:	¿me puedes poner el tercero junto a la estufa?	can you place the third one next to the stove?
check	accept-part	43	s:	¿éste junto a la estufa?	this one next to the stove?
assert	answer	44	u:	sí por favor	yes please
commit	accept ack	45	s:	ok	ok
assert		46	s:	graphical action	
check		47	s:	¿así está bien?	is it all right?
assert	answer	48	u:	sí	yes

Table 1. Corpus segment

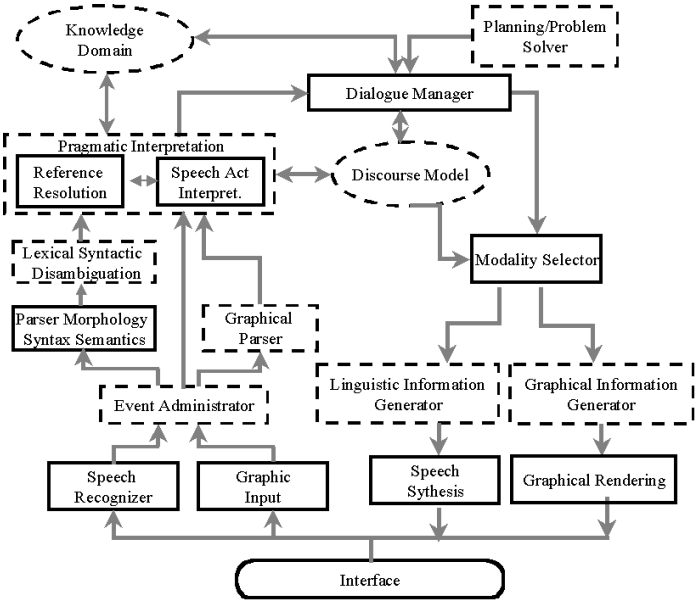

Fig. 3. Architecture of the prototype system

whose contents could be acknowledged with the current token, then a new CGU is created, and the token is included in it. However, if there is an accessible CGU for which the current token acknowledges the content, repairs the content, cancels the CGU, continues the content in such a way that all content could be grounded together, then this content is added to the current CGU. To carry on with the dialogue, all CGUs must eventually be grounded. Finally, the macro level consists in a level in which the dialogue is organized in terms of informational and/or intentional units. In this level, the main intentions expressed by the user during the conversation are represented, and to carry on with the conversation, units at this level of representation need to be completed also.

6 The prototype system

For the implementation of the project's prototype we are using *Open Agent Architecture* (*OAA*)[7]. This is a development environment for distributed systems which offers several advantages over competing technologies like *CORBA* and Microsoft *DCOM*, due, mainly, to its flexibility for handling multimodal asynchronous events. This flexibility permits to develop the prototype system incrementally, adding new modules and enriching the existing ones. The architecture of the *DIME* prototype is shown in Figure 3, where continuous lines indicate the modules that have been partially implemented, and the dashed lines indicate the modules that will be addressed in the future.

For the multimodal reference resolution the model proposed in Pineda and Garza [11] will be used, and for the rest of the system strategies similar to those

followed in *TRIPS*[10] will be employed.

7 Acknowledgments

We gratefully thank James Allen and his group at Rochester University; also to Enrique Sucar, to Eduard Hovy and to the anonymous reviewers of this paper. This project is being developed within the context of the bilateral initiative for the development of computer science between USA and Mexico, NSF/CONACyT with the support of CONACyT grant C092A; we also acknowledge the support of CONACyT grants 27948-A and 31128A.

References

1. J. Carletta. Assessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254, 1996.
2. A. Copestake. The LKB system. Technical report, Stanford University, 2001. http://www-csli.stanford.edu/~aac/lkb.html.
3. Alpha Software Corporation and Data Becker GmbH & Co KG. *Complete Home Designer Interiors, Users Guide 1998.* 1998.
4. M. Core et. al. 3rd. workshop of the discourse resource initiative. Technical report, Dept. of Cognitive and Information Sciences. Chiba University. Japan, 1999. No. 3 (CC-TR-99-1).
5. S. Sutton et. al. Universal speech tools: the CSLU toolkit. In *Proc. of the I'ntl Conf. on Spoken Language Processing (ICSLP)*, ICSLP Conferences, pages 3221–3224, Sidney, Australia, November 1998.
6. S. Young et. al. *The HTK Book.* Entropic Cambridge Research Laboratory, Cambridge University, England, 1997.
7. L. D. Martin. *The Open Agent Architecture: A framework for building distributed software systems.* http://www.ai.sri.com/~oaa/, 1999.
8. Grupo Tlatoa. *Speech Technology Research Group: TLATOA.* http://info.pue.udlap.mx/~sistemas/ tlatoa/, 2001.
9. E. Uraga. *Mexbet: Conjunto de símbolos fonéticos para el español.* IIMAS-UNAM, México, http://cic2.iimas.unam.mx/multimod/dime/doctos/espectrogramas/tablafonemas.html, 2001.
10. J. Allen; G. Ferguson y A. Stent. An architecture for more realistic conversational systems. In *Proceedings of IUI-2001*, pages 1–8, Santa Fe, NM, January 2001.
11. L. A. Pineda y G. Garza. A model for multimodal reference resolution. *Computational Linguistics*, 26(2):139–194, June 2000.
12. L. Villaseor; A. Massé y L. A. Pineda. The DIME corpus. In *ENC01, 3er Encuentro Internacional de Ciencias de la Computacin*, Aguascalientes, México, 2001. SMCC-INEGI.
13. Ivan Meza y Luis A. Pineda. The spanish auxiliary verb system in HPSG. In *Proceedings of CICLing-2002.* Springer-Verlag, LNCS (to be published), 2002.
14. Ivan Meza; Erik Schwarz y Luis Pineda. The clitic pronoun system in HPSG. Material report, Dept. of Computer Science, IIMAS, UNAM, (to appear).
15. J. Allen y M. Core. Draft of DAMSL: Dialog act markup in several layers. Technical report, U. of Rochester, 1997.
16. I. Sag y T. Wasow. *Syntactic Theory: A Formal Introduction.* CSLI Publications, Stanford, 1999.