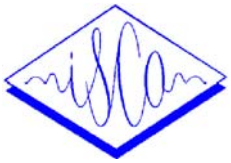

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

A DISCOURSE CODING SCHEME FOR CONVERSATIONAL SPANISH

Lori Levin[†], *Ann Thymé-Gobbel*[‡†], *Alon Lavie*[†], *Klaus Ries*[†], *Klaus Zechner*[†]

[†] Language Technologies Institute,
Carnegie Mellon University
[‡] Natural Speech Technologies
email : lsl@cs.cmu.edu

ABSTRACT

This paper describes a 3-level manual discourse coding scheme that we have devised for manual tagging of the CallHome Spanish (CHS) and CallFriend Spanish (CFS) databases used in the CLARITY project. The goal of CLARITY is to explore the use of discourse structure in understanding conversational speech. The project combines empirical methods for dialogue processing with state-of-the art LVCSR (using the JANUS recognizer). The three levels of the coding scheme are (1) a speech act level consisting of a tag set extended from DAMSL and Switchboard; (2) dialogue game level defined by initiative and intention; and (3) an activity level defined within topic units. The manually tagged dialogues are used to train automatic classifiers. We present preliminary results for statement categorization, and give an in-progress report of automatic speech act classification and topic boundary identification.

1. INTRODUCTION

The goal of the Clarity project [6] is to explore the use of discourse structure in understanding conversational speech. The project combines empirical methods for dialogue processing with state-of-the art LVCSR using the JANUS recognizer [5, 7]. We are currently working with the CallHome Spanish (CHS) and CallFriend Spanish (CFS) databases of unrestricted telephone conversation.

The particular understanding task that we are currently pursuing is *functional activity identification* — classifying segments of a dialogue as representing one or more of the following: informing, instructing, inquiring, planning, convincing, negotiating, gossiping, arguing, managing the conversation, and greeting/closing. The functional activity identifier will take as input aspects of discourse structure and prosodic information. Three levels of discourse structure are thought to be relevant: speech acts [13], dialogue games [3, 2], and topic segments. We are currently training automatic classifiers for these levels. This paper describes the discourse coding scheme that we use for manual tagging of the training data for these classifiers.

Our coding scheme divides discourse structure into three levels tagged separately. The three levels of the coding scheme, from lowest to highest, are (1) a speech act level consisting of a tag set extended from DAMSL and Switchboard; (2) dialogue game level defined by initiative and intention; and (3) an activity level defined by topic boundaries. Figure 4 shows a fragment of a tagged dialogue. Each of the three levels of tagging are discussed in the following sections. To facilitate the tagging process, we have written a detailed manual [14] containing decision trees, tag descriptions, examples, and helpful hints.

2. THE SPEECH ACT CODING SCHEME

The lowest level of tagging is the speech act level. Our coding scheme is heavily based on DAMSL (from the Discourse Resource Initiative [1]) and on the Switchboard SWBD-DAMSL tags (from the Johns Hopkins LVCSR summer workshop 1997 [9]). In this paper we will be referring to two speech act codings of CHS. The preliminary coding differed from SWBD-DAMSL in only minor ways. The entire CHS corpus was tagged with the preliminary coding scheme, inter- and intra-coder consistency was checked, and a speech act classifier was trained. We then revised the coding scheme in order to improve consistency, further refine large categories, and support the task of functional activity identification. Wherever we use a SWBD-DAMSL tag, it retains its original definition from the SWBD-DAMSL manual. We have made up new tag names for speech acts whose definition differs from the definition in the SWBD-DAMSL manual. The remainder of this section describes the revised coding scheme.

Our speech act tags fall into the categories Questions, Answers, Agreement/Disagreement, Discourse Markers (including Backchannels), Forward Functions, Control Acts, Statements, and Other. The greatest changes with respect to SWBD-DAMSL involve an expansion of control acts, the addition of coarse emotional tags to backchannels, a collapse of statements and opinions, and the addition of semantic features to statements.

Questions: The main question speech acts are `yes-no`, `wh-`, `alternative 'or'`, `open ended`, `repetition request`, `verification request`, and `rhetorical`.

Answers: Answers to yes-no questions can be `descriptive affirm`, `dispreferred`, `descriptive negative`, `no`, `don't know`, and `yes`.

Agreement/Disagreement: This category includes `accept` and

reject.

Discourse Markers and Backchannels: To distinguish between uses of backchannels that may be relevant at the dialog game and activity levels, we added three emotion/attitude indicators (positive/happy, negative/sad, and surprise). The main discourse marker speech acts are backchannel, link, verbal pause, and hedge [4].

Control Acts: Control acts, utterances that involve an expected action by the hearer or the speaker, include commands, requests, prohibitions, offers, promises, etc. Instead of using the SWBD-DAMSL ad tag (action directive), we base our expanded set of control act tags on Lampert & Ervin-Tripp [11]. The control acts are commit, permission request, offer, suggestion/permission-statement, directive/request, prohibition, and ownership claim.

Statements: Special attention was given to splitting descriptive statements and opinions (sd and sv in SWBD-DAMSL) into subcategories. Statements and opinion statements make up 48% of the speech acts and cover 71% of the words in our earlier tagged CHS database. Statements and opinions also displayed some inter-coder confusion in the preliminary tagging. One result of the LVCSR summer workshop was that one cannot improve the speech recognition accuracy significantly if there is a huge major category – the language model constraints placed on that category will be weak and therefore the improvement has to be small. [12] focused on subsegments of statements in a technique that has already shown significant word accuracy improvements. During the LVCSR summer workshop an initial study [10] showed that different types of statements are identifiable by their discourse context. For these reasons, we are actively pursuing further subcategorization of statements within project Clarity (Table 1).

In our coding scheme, all statements and opinions are coded with a single tag together with an added dimension of seven semantic features. A statement can be tagged with none, one or several of these semantic features. We are currently tagging the following features: (1) mental state: speaker expresses his/her own emotional state or psychological state; includes expressed emotion, preferences, psychological states, wants, tastes, likes, wishes, and desires; e.g. "I worry about you;" (2) reality: speaker expresses a claim about a hypothetical world; includes hypotheticals, conditionals, some wishes; e.g. "If I'd had time, I would have gone;" (3) value judgment/attitude: speaker expresses an attitude or value judgment, positive or negative, about state, situation, or people; includes some evaluatives (name calling), and some explicit opinions; e.g. "This tastes great," "He's so obnoxious;" (4) obligation: speaker expresses an obligation involving self; e.g. "I have to be back tomorrow;" (5) tense: speaker makes statement about something that has not yet happened; e.g. "We leave tomorrow;" (6) certainty: speaker expresses certainty or uncertainty about accuracy of his/her statement; e.g. "He's coming back tomorrow or maybe Tuesday," "He's back, yes siree;" and (7) joke/sarcasm: speaker makes joke or sarcastic comment; e.g. "Oh she'll just LOVE that." For all but two of the features we tag only the marked case. For the features attitude and certainty, we distinguish two marked cases, + and –, from the unmarked case. Thus, attitude can be positive or negative, and certainty can be explicitly certain or uncertain.

Statement type	Frequency
plain statement of information	3877
future	401
plain statement in indirect speech	235
positive value judgment	218
explicit uncertainty	216
mental	162
negative value judgment	145
statement with a question tag	84
hypothetical	73
future hypothetical	39
obligation	28
correction	26
plain statement to a third party	17
explicit uncertainty & future	15
plain statement, mimicing the other speaker	15
introduction statement	13

Table 1: Distribution of most frequent statement types according to revised coding scheme

3. MANUAL TAGGING OF DIALOGUE GAMES

The middle level of tagging is based on the idea of a dialogue game [2] [3] and work on illocutionary acts by Searle [13] and others. The focus of this level of tagging is on how turns interact, i.e. how utterances from two dialog participants relate to each other. Game boundaries are determined by changes in who has the initiative and changes in speaker intention, for example changing from informing to questioning. Carletta et al. describe a conversational game as "a set of utterances starting with an initiation and encompassing all utterances up until the purpose of the game has been either fulfilled (e.g. the requested information has been transferred) or abandoned." A game will therefore consist of all turns up to the point where the tagger finds that the game has been completed or, if incomplete, includes only the initiation of a game. Games are much like modified adjacency pairs, consisting of Moves that are required, expected, or optional. Each game consists of a required Initiative Move by Speaker A, a Response Move by Speaker B that is required or optional depending on the type of game, a Feedback Move by Speaker A that is always optional, and a possible second Feedback Move by Speaker B which is also always optional. Our system contains eight main types of games plus eight modifiers. The game types are seeking information, giving information, giving directive, action commit, giving opinion, expressive, seeking confirmation, and communication filler. Taggers label turns within a game as Initiative, Response, and Feedback. Games may overlap, either as nested games or as interleaved games.

```
#Topic Boundary
#Activ:Convince 0   #Activ:Inform 4
#Activ:Planning 5   #Activ:Negotiate 2
#Activ:Inquire 1    #Activ:Gossip 0
#Activ:Argue 0      #Activ:Conv-mgmt 0
#Activ:Greet-cls 0  #Activ:Instruct 1
#Game:Questˆ Aband
<I> qw B: pero como,
          but how
#Game:Quest
<I> qy B: pero pagan impuestos,
          but are they taxed
<I> sˆcert-
    B: pero se supone que el menaje no paga
    but household items are not supposed
    to be taxed
<R> ny A: si'
          yes
#Game:Infoˆ Elab
<I> sˆe A: no si' paga impuestos,
           no yes it is taxed
<I> sˆcert+
    A: paga el quince por ciento, si' sen˜or
       it's taxed fifteen per cent, yes sir
<R> b  B: ah si'
          oh yes
```

Figure 1: A Fragement of a Tagged Dialogue

4. MANUAL TAGGING OF TOPIC SEGMENTS AND ACTIVITIES

The highest or most general level of tagging identifies a discourse segment and an activity focusing on the purpose and goal of the speakers within the segment. Since we are not working with task oriented dialogues, segments cannot be defined in terms of subtasks. Instead they are defined by topic boundaries. We are currently labeling ten activities: inform, inquire, plan, convince, negotiate, gossip, argue, conversation management, greet-close, and instruct. The presence of each activity is judged on a continuum. The tagger decides to what degree a certain activity is present and assigns a numerical value on a scale from 0 to 5, 0 being "not present in the segment" and 5 being "strongly present in the segment."

Figure 4 shows a fragment of a dialogue tagged with activities, games, and speech acts. The speech act tags are `s` (statement), `qw` (wh-question), `qy` (yes-no question), `ny` (no answer), and `b` (backchannel). `ˆcert+` and `ˆcert--` indicate the semantic features certainty and uncertainty. `ˆe` tags an elaborated reply to a yes-no question. `ˆm` is the label for a mimic of the other speaker's utterance. `<I>` and `<R>` are initiative and response moves at the game level of tagging.

5. RELIABILITY OF THE CODING SCHEME

We conducted a variety of evaluations on the reliablilty of the preliminary speech act coding scheme, including intercoder agreement, intracoder agreement, confusion matrixes, and the effects of listening or not listening while tagging. The corpus was not presegmented into speech act units before manual tagging. Segmentation was therefore performed as part of the tagging process. Using one tagger as a reference, the other had a segment boundary precision of 86.1% and recall of 88.9%. Because the speech act boundaries do not necessarily agree, intercoder agreement is measured in terms of the percentage of *words* that are tagged with the same speech act by both taggers. Intercoder agreement on ten dialogues was 78.7% for two coders using the preliminary coding scheme. The most confusable speech act tags `sd` and `sv` (descriptive and opinion statements). `sd/sv` disagreements encompassed 9.2% of the word tokens. On a test of three dialogues, intracoder agreement (with a time lag between codings) was about 85%. Tagging first without listening and then later with listening resulted in revision of about 3.5% of the tags.

6. AUTOMATIC CLASSIFICATION AND SEGMENTATION

The purpose of the manually tagged data is to train automatic classifiers for identifying the three levels of discourse structure described above. We have trained a classifier for the speech act level and a segmenter for breaking dialogues into topical segments.

The classifier technology we are using is an HMM of speech acts, were each speech act is modeled as an ngram of words. Since the segmentation and labeling of text is known in training, these models can be trained as Markov Models using standard language modeling tools. Using an A^* search algorithm the best segmentation and labeling is simultaneously found (see [6, 15] for a more detailed discussion of our implementation and a comparison with other approaches).

We ran a preliminary speech act classification experiment using the preliminary speech act tags, training on a set of 80 dialogues from CHS and testing on 40 CHS additional dialogues that were not used for training. Performance results are shown in Table 2 and will be compared with results from the updated coding scheme in our ICSLP-98 presentation. The results are presented in terms of how many words have correct speech act tags, not how many sentences have the correct speech act tags. This is because the classifier performs segmentation of utterances into speech act level units simultaneously with the classification, and the automatic segmentation may have different speech act boundaries from the manually tagged data that it is checked against.

To automatically determine topical segment boundaries in the CallHome dialogues, we chose to use [8]'s TextTiling algorithm. First we compute text similarity scores for all potential boundaries using blocks of text to the left and to the right of the boundary. We then move this 2-block-window over the entire dialogue and compute the similarity scores for each potential boundary. The resulting

Method	Result
baseline (most likely tag)	55%
classifier	69%
intercoder agreement	79%
interacoder agreement	85%

Table 2: Results of Automatic Speech Act Classification in the preliminary coding scheme

Language	English	Spanish
nr. of dialogues	9	15
avg. nr. of turns	345	252
uniform baseline (F1)	0.48	0.47
crossvalidation (F1)	0.58	0.53

Table 3: Topic Segmentation Results for English and Spanish CallHome

similarity graph is then smoothed, and segment boundaries are hypothesized at the minima of the similarity graph.

We ran experiments on English and Spanish CallHome data. The results of a k-fold crossvalidation are presented in Table 3. [1] [2]

ACKNOWLEDGEMENTS

We would like to thank our taggers Liza Valle, Santiago Cortes, and Susana Fuentes Arce. This work was conducted under funding from the U.S. Department of Defense.

7. REFERENCES

1. Allen, James and Mark Core, Draft of DAMSL: Dialog Act Markup in Several Layers, http://www.cs.rochester.edu/research/trains/ annotation/RevisedManual/RevisedManual.html, 1997.
2. Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J., & Anderson, A., "The Reliability of a Dialogue Structure Coding Scheme, *Computational Linguistics 23(1): 13-31*, 1997.
3. Carlson, L., *Dialogue games: An approach to discourse analysis*, Dordrecht, Holland: D. Reidel Publishing, 1983.
4. Chafe, W. L., "Prosodic and Functional Units of Language," In Edwards and Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale NJ: Lawrence Erlbaum Associates, pages 33-43, 1993.
5. Chu-Carroll, Jennifer and Nancy Green (eds.), *Applying Machine Learning to Discourse Processing, Papers from the 1998 AAAI Symposium*, Stanford University, Technical Report SS-98-01.
6. Finke, M., Lapata, M., Lavie, A., Levin, L., Mayfield-Tomokiyo, L., Polzin, T., Ries, K., Waibel, A., Zechner, K., "CLARITY: Inferring Discourse Structure from Speech." In *Applying Machine Learning to Discourse Processing, AAAI'98 Spring Symposium Series*, March 23-25, 1998, Stanford University.
7. Fritsch, Jürgen and Michael Finke: 1997, 'Improving Performance on Switchboard by combining Hybrid HME/HMM and Mixture of Gaussians Acoustic Models', *Proceedings of Eurospeech 97.*
8. Hearst, Marti A., "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages," *Computational Linguistics, Vol. 23.1*, pages 33-64, 1997.
9. Jurafsky, D., Shriberg, L., & Biasca, D., Coder's Manual Draft 11, Switchboard Shallow-Discourse-Type Labeling Project Switchboard DAMSL (SWBD-DAMSL) Labeling System, http://stripe.colorado.edu/ jurafsky/manual.august1.html, 1997.
10. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Andreas Stolcke, A., Taylor, P., and Van Ess-Dykema, C., *SWBD Discourse Language Modeling Project, Final Report*, Johns Hopkins LVCSR Workshop-97, 1997
11. Lampert, M. D. & Ervin-Tripp, S. M., "Structured Coding for the Study of Language and Social Interaction," In Edwards and Lampert (eds.) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale NJ: Lawrence Erlbaum Associates, pages 169-206, 1993.
12. Ma, K. W., Zavaliagkos, G., and Marie Meteer, M., "Sub-Sentence Discourse Models for Conversational Speech Recognition," In *ICASSP 1998 Proceedings*, 1998.
13. Searle, J. R., "The classification of illocutionary acts," *Language in Society 5, 1-24*, 1976.
14. Thymé-Gobbel, Ann and Levin, L., Speech Act, Dialogue Game, and Dialogue Activity Tagging Manual for Spanish Conversational Speech, http://www.cnbc.cmu.edu/ gobbel/clarity/manualintro.html, 1998.
15. Van Ess-Dykema, C., and Ries, K., "Linguistically Engineered Tools for Speech Recognition Error Analysis", in *ICSLP 1998 Proceedings*, 1998.

[1] $F1 = \frac{2PR}{P+R}$, where R=recall and P=precision.

[2] The "uniform baseline" inserts evenly spaced boundaries in all dialogues.