*Proceedings of the Second International Conference On Language Ressources And Evaluation,*
*LREC 2000, Athens, Greece, 31st May-2nd June 2000*

# Shallow discourse genre annotation in CallHome Spanish

**Klaus Ries[*,†], Lori Levin[*], Liza Valle[*], Alon Lavie[*], Alex Waibel[*,†]**

Interactive System Labs
www.is.cs.cmu.edu
{ries|lsl|alavie|ahw}cs.cmu.edu
[*] Carnegie Mellon University, Pittsburgh, PA, USA
[†] University of Karlsruhe, Karlsruhe, Germany

**Abstract**

The classification of speech genre is not yet an established task in language technologies. However we believe that it is a task that will become fairly important as large amounts of audio (and video) data become widely available. The technological cability to easily transmit and store all human interactions in audio and video could have a radical impact on our social structure. The major open question is how this information can be used in practical and beneficial ways. As a first approach to this question we are looking at issues involving information access to databases of human-human interactions. Classification by genre is a first step in the process of retrieving a document out of a large collection. In this paper we introduce a local notion of speech activities that are exist side-by-side in conversations that belong to speech-genre: While the genre of CallHome Spanish is personal telephone calls between family members the actual instances of these calls contain activities such as storytelling, advising, interrogation and so forth. We are presenting experimental work on the detection of those activities using a variety of features. We have also observed that a limited number of distinguised activities can be defined that describes most of the activities in this database in a precise way.

## 1. Introduction

Deep natural languge understanding of unrestricted human speech is a task that, for the most part, still falls beyond current technology. In fact, in many situations, even a casual human overhearer cannot fully comprehend such dialogue, particularly in conversations between close acquaintances. When we inspected human transcripts of our group-meetings the word "tagger" (as in parts-of-speech tagger) was consistently transcribed as "tiger". Instead of aiming for a full understanding of such human conversation, we have been attempting to annotate discourse at different levels with shallow labels. Our goal is to identify information about the discourse that can then be used to index the dialogue for information access, for applications such as browsing through a database of meetings (Waibel et al., 1998). In work conducted under the US Government funded CLARITY Project, we have been focussing on the CallHome Spanish database. CallHome Spanish consists of 120 telephone conversations of Spanish speakers calling from the US back to their home-country. Calls are usually 30 minutes long. We have developed annotation schemes for three levels of discourse structure. In previous publications (Levin et al., 1998; Ries, 1999a), we reported on our annotation of speech acts and dialogue games and the technology used to automatically identify and classify these levels of discourse structure. Work by collaborators in CLARITY included emotion detection and summarization. The focus in this paper is our recent work on the identification of style or "activity" of longer segments of discourse. CallHome Spanish has also been one of the databases used for extensive evaluations in the large vocabulary speech recognition (LVCSR) community. We plan to make our discourse annotations widly available to the research community at large [1]. This paper will first introduce the notions of topic, genre and activity, will then move on to the presentation of a machine learning approach for the detection of activities and conclude. The concrete tagging instructions are included in an appendix.

## 2. Topic, genre and activities

The definition of topic in linguistics is all but consistent. A recent literature review can be found in (Goutsos, 1997), who puts special emphasis on the fact that topic can often be more reliably defined linguistically by not referring to the coherence of the propositional extension of a segment. However in information retrieval and in summarization the definition of topic is purely based on keyword coherence (e.g. (Hearst, 1997)), a narrow application of the coherence of the propositional extension [2] On the other hand the application of the keyword coherence assumption can be done in a fairly robust and general way across registers (Finke et al., 1998) and does not require much fine-tuning. (Goutsos, 1997) himself is more attracted by staging or sequencing behaviors on the discourse level, however his work is related to expository texts.

Earlier work (Longacre, 1996; Gee, 1986; Labov and Waletzky, 1967; Tannen, 1993; Eggins and Slade, 1998;

[1] The database will likely be submitted to the LDC before the conference and is currently in its final processing stage.

[2] In the context e.g. of a discourse between a travel agent and a customer planning and booking a trip usually comes in fixed phases and one may wish to be able to identify those. The keyword based mechanism would completely fail in this situation since coherence might actually predict completely wrong links, e.g. the city of *Heidelberg* as the destination will be mentioned in the phase where the air transportation and the local transportation is planned as well as when lodging and sightseeing is being discussed. However the main topic of each of these phases is very different and a city name might be one of the few repeated items.

Levinson, 1979; Plum, 1988) on conversations and narratives has already followed this route to some extent: Devices that signal a certain quality of discourse were identified and stages particularly of narratives have been identified. If staging behaviour is central to the respective author, the terms *genre* or *frame* are also frequently used. Especially the term genres (Labov and Waletzky, 1967; Eggins and Slade, 1998; Plum, 1988) associates a very fixed sequence of stages.

(Levinson, 1979) uses the term *activities*, but does not elaborate on the staging structure other than pointing to specific introductions: He is just supposing that the choices of the individual discourse participant are restricted by the activity (the *structural* aspect of the activity). He also points out that besides the structural aspect an activity consists of a *stylistic* aspect that he does not elaborate further on in his presentation. Conceptually our approach and detection technology is close to (Levinson, 1979) and it therefore seems natural to use his term "activity" rather than genre or frame. (Linell, 1994; Linell, 1990) takes this one step further by looking at discourse as a joint achievement (Clark, 1996). (Linell, 1994) describes different activity types by the distribution of speech act types.

| Activities (80 dialogues) | | | |
|---|---|---|---|
| Activity | Count | Activity | Count |
| Story-telling | 672 | Undetermined | 57 |
| Planning | 70 | Closing | 22 |
| Advising | 60 | Discussion | 7 |
| CallHome | 59 | Consoling | 6 |
| Interrogation | 57 | | |

| Orthogonal annotations (967 segments in 80 dialogues) | | | |
|---|---|---|---|
| Evaluation | Count | Who or what | Count |
| positive | 50 | speaker A or speaker B only | 49 |
| negative | 56 | speaker A and speaker B only | 271 |
| divergent | 7 | other people (may include A or B) | 401 |
| neutral | 854 | practical topics | 135 |
| | | politics | 14 |
| | | other | 59 |
| | | unknown | 38 |

Table 1: **Activity statistics:** Out of 80 manually annotated dialogues we have measured the distribution of the main activity types (upper table). We also calculated the evaluations and a manual categorization of the "who or what" orthogonal categories (lower table).

Initially we were attempted to apply a coding scheme similar to (Eggins and Slade, 1998) since they have also been working on gossip. However, we discovered that their definitions were hard to apply to personal conversations between family members. Additionally we believed that – even if we were able to annotate this classification by hand – that we would not be able to apply automatic techniques. The two main reasons were:

- gossip is marked as third person oriented (Eggins and Slade, 1998) while we have found numerous examples of discourse that looked like gossip but was first person oriented. Also we have not found a lot of segments that contain explicit evaluations.
- the different types of storytelling are only distinguishable by looking at very fine distinctions at turning points in the conversation. Inspecting our data we have found it difficult to make those fine distinctions by hand and we found the resulting labeling counterintuitive.

We therefore devised a tagging scheme that assigned one major category to each segment (Fig. 1). Additionally we attach a positive/negative/divergent/neutral evaluation annotation to each segment and identify the main person/object of the conversation. Gossip, as defined by (Eggins and Slade, 1998), therefore corresponds approximately to a *story-telling* segment with a *negative* evaluation about a *third person*.

## 3. Detection of functional activities

We have been investigating machine learning techniques for the automatic identification of functional activities from tagged data. So far we have simplified the problem somewhat by assuming that the dialogue is presegmented and only the correct activity label has to be determined and assigned. The problem of segmenting the dialogue into activities is discussed in Sec. 3.2.

### 3.1. Annotating functional activities

We are currently only attempting to assign the major label of the functional activity such as storytelling, we have not attempted to tune our classifiers for the other problems but rather report out-of-the-box performances for those. This task has already turned out to be fairly hard. Out of a number of different machine learning techniques (neural networks, naive Bayes, support vector machines, k-nearest-neighbor, decision trees) that we have successfully applied to other automatic discourse annotation tasks (speech acts and dialogue games), only neural networks were able to deliver results better than picking the most likely category. The input feature space consists of

- interactional features (such as pauses and speaker overlap)
- word level information
- dialogue act information
- dialogue game information
- stylistic information derived from the word level using regular expressions and shallow parsing [3]

Additionally, we use an ngram induction technique that is related to maximum entropy modeling and that allowes us to integrate discriminatory phrases very efficiently (Pietra et al., 1997). Most of the count values

[3] We are thankful for using the shallow grammar developed by Klaus Zechner and to Donna Gates who developed Spanish stylistic features.

are mapped with a logarithmic function, most models reported below are therefore equivalent to a multiplicative model. Models with hidden layers did not improve performance. Using "vanilla" speech acts (traditional DAMSL-speech acts (Core and Allen, 1997; Jurafsky et al., 1997; Stolcke et al., 1998)) we did not see any improvements. However, when we used the enhancements in our tagging scheme such as as future statements, value judgments, certainty and hypotheticals (Levin et al., 1998; Thymé-Gobbel and Levin, 1998) we saw improvements from speech acts. Currently we are also preparing the use of prosodic features such as pitch and power contours into the feature set.

The annotation results reported below have been obtained on just 520 activity segments corresponding to 40 conversations. The basic statistics report is on the recently completed database of 80 conversations and we hope that the detection results on the larger database are improved – the relative frequencies of activities stayed approximately the same from the original set of 40 conversation. This is a relatively small number of tokens given the complexity of the classification task. As it turns, out we can make use of hand annotated activities but using word based regular expressions that mark style (designed similar to (Biber, 1988)) we were able to approach a good initial result without the dialogue act or game classification (Tab. 2). The largest contributor to the error rate is the distinction between storytelling other activities (Tab.3 and4). Finally we tried to detect the orthogonal evaluation attribute but our work in this direction has been preliminary so far (Tab. 5).

| Features | Accuracy in % |
|---|---|
| baseline | 67.7 |
| words per channel | 68.5 |
| + stylistic | 69.3 |
| words per channel | 68.5 |
| + dialogue acts | 68.7 |
| + games | 70.2 |
| + 50 game ngrams | 70.6 |

Table 2: **Activity detection:** Using neural networks with no hidden units we have achieve a reasonable detection accuracy.

### 3.2. The segmentation problem

The topic segmentation algorithm proposed by (Hearst, 1997) is based on the idea that each segment should exhibit a uniform vocabulary profile. Using unigram cache

| manual | automatic | |
|---|---|---|
| | storytelling | other |
| storytelling | 300 | 50 |
| other | 87 | 84 |

Table 3: **Storytelling detection:** Discriminating between stories and non-stories can be done at a 73.7% level while 67% is the baseline result just picking storytelling. This result seems to be the limiting factor for the activity detection results.

| manual | automatic | | |
|---|---|---|---|
| | negative | neutral | positive |
| negative | · | 28 | · |
| neutral | 2 | 290 | 2 |
| positive | · | 24 | 2 |

Table 5: **Evaluation detection:** We have not focussed on this problem at all but the current detection results are just the baseline (84.4%).

models (Kuhn and de Mori, 1990) this may therefore be formulated as finding a segmentation $S$ for the word string $W$ such that

$$max_S p(S|W) = max_S p(W|S)p(S)$$

Assuming that the topics are independently generated we can simplify

$$p(W|S) = \prod_i p(W_i)$$

where $W_i$ is the $i$th segment in $S$ and $p$ is a unigram cache model. In (Hearst, 1997) only the keywords are entering into $W$. This approach can obviously be generalized to include other features that are likely to stay constant across one segment. One could e.g assume that initiative stays constant across a segment and therefore the stream of speaker identies or dialogue act/speaker pairs may be usable with caches as well. We have found that this feature alone gave performance results similar Hearst's approach. Other potential features could be the likelihood of topic word occurence, indicators of syntactic complexity or any of the other features usable for activity detection. We have not integrated these two systems yet.

## 4. Conclusion and Outlook

The information access problem to human-human interactions could be one of the biggest upcoming challenges to language technologies. In this paper we add one new facet to the detection of high-level features of human interactions, the detection of sub-dialogue level activities.

Although the results indicate that automatic activity labelling is very hard, we found that information from the (enhanced) speech act and game level as well as from the word level helps the classification. In other work we have shown that detecting broader genre or register differences is a really easy task, e.g. between different corpora (Ries, 1999b). Currently we are also classifying different TV show types, we are therefore mapping the limit of this technology. It is also not clear at what level this technology is useful for information access. Quantifying the utility of discourse information is part of our current investigation and even if discourse information is not much more helpful then keyword based information it might be easier to obtain from actual audio data.

We have also seen that we can get away by mostly using word level features in combination with stylistic features. The advantage of detecting activity labels from the word level (including stylistic features) is that no additional classification problem needs to be solved. We believe that this

| manual | automatic | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | advising | | closing | | discussion | planning | | |
| | callhome | | consoling | | interrogation | undeterm. | | |
| advising | 12 | 7 | · | · | · | 1 | 4 | 7 |
| callhome | 4 | 25 | 1 | · | · | 2 | 1 | 4 |
| closing | 1 | 2 | 4 | · | · | · | 1 | 1 |
| consoling | · | 1 | · | · | · | · | 1 | · |
| discussion | 1 | · | · | · | · | · | 2 | 1 |
| interrogation | 1 | 2 | · | · | · | 17 | 1 | 5 |
| planning | 6 | 4 | · | · | · | 1 | 11 | 6 |
| undetermined | 2 | 10 | · | · | · | 7 | 3 | 12 |

Table 4: **Activity detection excluding storytelling:** While this detection task is far from being solved (47.4% at a 21.6% baseline) it seems that the activity detection task excluding storytelling is far more tractable.

is mostly due to the fact that there are too few speech acts or games per segment to make effective use of them. The advantage of building models from the speech act or game level would be that they might translate across languages and registers.

Our experience demonstrates the necessity of building a number of compatible resources for one database. We have found that it is necessary to understand the interaction between the coding schemes of different levels of discourse, and that the presence of the variety of discourse annotations allows us to produce interesting comparable results for a variety of methods quickly.

## 5. References


Biber, Douglas, 1988. *Variation across speech and writing*. Cambridge University Press.

Clark, Herbert H., 1996. *Using Language*. Cambridge University Press.

Core, Mark G. and James Allen, 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Eggins, Suzanne and Diana Slade, 1998. *Analysing Casual Conversation*. Cassell.

Finke, Michael, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner, 1998. Clarity: Automatic discourse and dialogue analysis for a speech and natural language processing system. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Gee, J. P., 1986. Units in production of narrative discourse. *Discourse Processes*, 9:391–422.

Goutsos, Dionysis, 1997. *Modeling Discourse Topic: Sequential Relations and Strategies in Expository Text*. Number LIX in Advances in Discourse Processes. Ablex Publishing.

Hearst, Marti A., 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Jurafsky, Dan, Liz Shirberg, and Debra Biasca, 1997. Switchboard swbd-damsl coders manual. Draft 13.

Kuhn, Roland and Renato de Mori, 1990. A cache-base natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and machince Intelligence*, 12(6):570–583.

Labov, W. and J. Waletzky, 1967. Narrative analysis: oral versions of personal experience. In J. Helm (ed.), *Essays of the Verbal and Visual Arts*. American Ethnological Society, proceedings of Spring Meeting 1967, Washington DC: University of Washington Press.

Levin, Lori, Ann Thymé-Gobbel, Klaus Ries, Alon Lavie, and Monika Woszczyna, 1998. A discourse coding scheme for conversational spanish. In *ICSLP*.

Levinson, Stephen C., 1979. Activity types and languagen. *Linguistics*, 17(5):365–399.

Linell, Per, 1990. *The Dynamics of Dialogue*, chapter The power of dialogue dynamics. Harvester Wheatsheaf.

Linell, Per, 1994. *Approaching Dialogue: talk, interaction and contexts in dialogical perspectives.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

Longacre, Robert E., 1996. *The grammar of discourse*. Plenum Press, 2nd edition.

Pietra, S. Della, V. Della Pietra, and J. Lafferty, 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

Plum, G. A., 1988. *Text and contextual conditioning in spoken English: A genre-based approach*. Ph.D. thesis, University of Sidney.

Ries, Klaus, 1999a. HMM and neural network based speech act classification. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1. Phoenix, AZ.

Ries, Klaus, 1999b. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proceedings of the Eurospeech*, volume 3. Budapest, Hungary.

Stolcke, Andreas, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema, 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Tannen, Deborah (ed.), 1993. *Framing in Discourse*. Oxford University Press.

Thymé-Gobbel, Ann and Lori Levin, 1998. *Dialogue Activity Tagging Manual for Spanish Conversational*

*Speech*. can be obtained through our Web-site.

Waibel, Alex, Michael Bett, and Michael Finke, 1998. Meeting browser: Tracking and summarising meetings. In *Proceedings of the DARPA Broadcast News Workshop*.

# A Activity annotation in CallHome Spanish

An *activity* is a unit of the discourse where a fixed patterned activity is present. This activity may come in separate stages and in standardized situations activities can be seen like "rituals". The way, an argument is presented in a courtroom is a prototype of this. Someone telling a story, people gossiping or planning something are other standard examples. All of these activities are *culture and register dependent* although their functions in the cultural or register context might be similar. However the means for communicating these might be very different across registers. E.g., two people gossiping in Spanish might be very different from two people gossiping in Japanese, although they are both exchanging information about acquaintances and exchange value judgments. Or e.g., a mother planning a trip with her son to the mall might be very different to the start of a civilian airliner or the correspondence between individual soldiers planning an attack. What these situations have in common however is that there is critical information that needs to be transported and agreed upon to achieve a successful communication. This manual tries to achieve an initial step towards defining frequent activities in CallHome Spanish. Activities have so far only been described more explicitly for narratives and recently also for some other types (Eggins and Slade, 1998).

Register however describes a more broad variation of discourse style. While a activity segment may be summarized in some standardized form as to what has been achieved in that segment the length of a register is more undetermined. Specifically the register is not really changing in CallHome Spanish as far as we have observed so far: There are always two or more close family members chatting over the phone. A variation that we would deem more relevant for a change in register would e.g. be the difference between a TV game-show and a news broadcast.

Within one activity segment one would usually assume to have one or maybe multiple things that are being talked about. Sometimes there is a separate segment at the end of a activity – especially in story telling – where diverse themes are being touched upon. Another characteristic of a activity segment is that the roles assumed by the speakers usually don't change: In a story-telling segment one speaker is often dominantly telling the story whereas the other one is usually assuming a supportive role: In this sense the activity is governing the general options of the participants. In CallHome Spanish we can also observe a second class of storytelling where the story is being told in a collaborative way.

To ease the tagging and also since we have been observing that these decisions would be hard to make otherwise we have decided to have an orthogonal dimension to activity-type, namely the evaluation. In an evaluation the speakers present their value judgment on the people or matter under scrutiny. Another orthogonal dimension we introduced is the "who or what" function that describes what the primary subjects or objects that are being discussed. According to (Eggins and Slade, 1998) gossip would therefore be a story-telling segment that contains evaluations and is about other people. We decided to factor those issues out for a couple of reasons: In most segments we would have intuitively defined as gossip we have found little or no evaluation. The gossip in CallHome Spanish is usually not explicitly used to construct common believes and values, as (Eggins and Slade, 1998) have found in their data. It may however be serving as a social signal that one can offer gossip about someone else since one is still in touch with those very people. The other issue is that there are sometimes segments that look like gossip – actually often containing explicit evaluations – that concern the speakers proper.

## A1. Main categories

**Storytelling** In (Eggins and Slade, 1998) a variety of ways to present a story about someone or something are presented, e.g. storytelling, recount, narrative and exemplum. We found it extremely difficult and unintuitive to apply those distinctions in our database. Since our goal is to find knowledge that allows us to navigate in a database we figure that only very intuitive notions will serve our purposes. Apart from this economic consideration the individual signals one would have to evaluate are extremely ambiguous and not very well represented on the surface of the conversation. Additionally the regular expression that are classically used to define those differences seem to be getting arbitrarily complex and do not correspond to our intuition anymore. Since one of our goals was to stick to surface cues we have decided to abandon these distinctions.

A strong cue for a storytelling activity are the following subparts in a story, usually in this default sequence, where all elements are optional or repeatable (Labov and Waletzky, 1967; Eggins and Slade, 1998):

1. abstract/introduction
2. orientation (initial part of story)
3. complication
4. evaluation
5. resolution
6. coda (a final wrap-up section containing relating this story to other things, finding the next topic to talk about etc.)

In many situations we also find appraisals in storytelling activities. In most storytelling activities one speaker is dominant and assumes the role of the storyteller. There are two other options: Both are telling the story collaboratively or one speaker basically triggers the other all the time to continue the story and might therefore use a lot of the channel while not being the storyteller.

**Planning** Planning is a activity where people try to figure out the course of some future events they are intending to engage in. Planning also typically entails a mutual commitment to carry out the plan that was agreed upon. In CallHome Spanish planning typically relates to trips/visits, career changes and moving homes. Evaluations are very rare in planning and we have not identified substructure in planning activties.

**Discussion** Discussions are mutual exchanges of information on a certain topic, often coupled with appraisals. The discussion is different from the storytelling activity in that there is not just one central story that is being told and that the exchange is usual mutual. Topics of discussions in CallHome Spanish are usually news, sports and politics, rarely acquaintances.

**Advising** In an advising segment one speaker is giving – solicited or unsolicited – advice to the listener about a specific situation, usually a personal matter. It usually includes instructions (weak or strong, commands and recommendations). The specific function of this activity is to express the speakers opinion about a rather personal issue and try to make the other person follow that advice. The advice is usually offered by the speaker who is more mature or has the higher authority. Evaluations are rare in this category.

**Consoling** Consoling is a activity that described as one speaker giving emotional support to the listener in times of personal misfortune (a divorce, the loss of a family member, an accident). We decided to include also situations were one speaker is praising the other since this is similar on the surface of the conversation and hard to determine. There is little or no evaluation in this category.

**Closing** The function of closing is to end an extended discourse segment because the speaker wants to move on and talk to a third person or just to end the whole conversation. It usually includes all the greetings and farewell expressions. There are no well defined topics and the utterances are usually short, but the activity itself can be long. This is especially common in the CallHome Spanish database since ending conversations in Latin American countries are bound to a set of rules of courtesy. There is no dominant speaker in these discourse segment, but rather an interactive exchange of farewell expressions. Evaluations are rare in closings.

**Interrogation** An interrogations is characterized as obtaining information through the use of questions. There is one dominant speaker who initiates and dominates the conversation, the other speaker would usually not have volunteered the information in another situation. The questions are intended to get specific – usually personal – information from the other participant in the conversation. The responses to the questions are usually short and are limited to answer the questions. The passive speaker does not take the floor of the conversation through his/her answers. There are usually no appraisals and evaluations and the activity is rare in CallHome Spanish.

**CallHome Recording** The CallHome Spanish database contains segments that are directly and obviously generated by the recording environment. Typical topics are: The length of time that the speakers have available for talking, the purpose of the phone call, whether the phone call is free, whether or not the speakers are being recorded and finally how they found out about the free phone calls. In one extreme example the overhearer is explicitly addressed and educated about Spanish phonology.

**Undetermined** Discourse segments that are incomplete due to the segmentation of the transcript, especially when not enough material is available to make an activity decision. Some discourse segments are also labeled undetermined when the participants in the conversation purposely exchange information in codes or use language that can have multiple interpretations and make it, therefore, incomprehensible to the tagger. This activity is quite common at the end of the conversation and at the beginning.

### A2. Orthogonal activity attributes

**Evaluation** Evaluations can be used for people, relationships or behaviors. Evaluations of events, incidents, tangible things or social constructs however will not be marked. The original goal of marking evaluations is to discriminate between neutral stories that are being told and stories that have more gossip character. The evaluations that we are marking have to be *explicit* on the surface of the conversation. Other evaluations are usually very hard to decided as an overhearer so we left them unmarked.

A dialogue can be marked as neutral (no tag), positive, negative evaluation by at least one speaker or divergent evaluations.

**Who or what** The who or what feature is orthogonal to the other activity features. It is meant to capture the main object/subject that is being discussed. In CallHome Spanish this is in for the storytelling activities usually a person, for the planning sections often a trip to some place etc. The "who or what" would have been called the atomic or discrete topic by (Goutsos, 1997): This may seem to contradict our focus on "how" something is presented: However there seem to be canonical ways e.g. in storytelling how to introduce people as a topic for the discourse which may lead to simple algorithms for identifying these topics without a deep analysis.