A  Project Summary

(I will redo this. Right now it is just a copy of the overview.)

Large scale annotated corpora have played a critical role in speech and natural language research. They have enabled large scale integration of statistical knowledge (derived from the corpora) with linguistic knowledge (as represented in the annotations) leading to both scientific as well as technological advances, for example in robust parsing, automatic extraction of relations, coreferences, among others and their applications to information extraction, question answering, summarization, and machine translation, for example. The Penn Treebank (PTB) a resource developed here at Penn almost a decade ago is one spectacular example of such a resource with worldwide impact on natural language processing.

Although the existing annotated corpora such as PTB have been highly successful, it has long been recognized that we need a new generation of annotated resources, which encode on a large scale and in a reliable manner, some key aspects of discourse and dialog structure. Although resources such as PTB deal with corpora, in a sense they deal with them at the sentence level only. These resources have been immensely useful. However, by now we have almost exhausted their use. We need a new large scale and reliable discourse and dialog structure annotated corpora. Although there have been some attempts to construct such resources, they are small scale and not incremental, i.e., not built on top of (or associated with existing large scale resources). Our goal is to build two such large scale resources, one in the discourse domain and one in the dialog domain.

In particular we propose the following two major projects.

1. Discourse Penn Treebank (DPTB): In this project we will develop a large scale and reliably annotated corpus which will encode coherence relations associated with discourse connectives, including their argument structure and anaphoric links, thus exposing a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with the discourse connectives. This annotation will be 'on top of' the PTB annotations as well as the predicate-argument annotations on PTB (called the Proposition Bank or Prop Bank), which will become available soon.

2. MultiFORM– Augmenting the FORM corpus with body movement, speech and intonation: FORM is a corpus of gesture-annotated videos. It was designed to be extensible in order to eventually represent the entire multimodal experience of conversational interaction. We will create this multimodal FORM (MultiFORM) by adding (1) body movement, (2) speech and syntactic structure, and (3) intonation.

In the following sections we will describe these two projects as separate projects. Although there are some interesting connections between these two efforts from the point of the resource development, it is best to treat them as separate activities. To be sure, from time to time, we will be examining the relationships between the tools developed in these two projects, especially from the point of view of the possible use of the tools developed in one project for the other project.

1 Although 'discourse' is a rather general term, we will use the term for corpora such the Wall Street Journal Corpus, the Brown Corpus, among others. We will use the term 'dialog' for the corpora that deal with interactions among two or more participants.
B Project Description

a Overview

Large scale annotated corpora have played a critical role in speech and natural language research. They have enabled large scale integration of statistical knowledge (derived from the corpora) with linguistic knowledge (as represented in the annotations) leading to both scientific as well as technological advances, for example in robust parsing, automatic extraction of relations, coreferences, among others and their applications to information extraction, question answering, summarization, and machine translation, for example. The Penn Treebank (PTB), a resource developed here at Penn almost a decade ago, is one spectacular example of such a resource with worldwide impact on natural language processing.

Although the existing annotated corpora such as PTB have been highly successful, it has long been recognized that we need a new generation of annotated resources, which encode on a large scale and in a reliable manner, some key aspects of discourse and dialog structure. Although resources such as PTB deal with corpora, in a sense they deal with them at the sentence level only. These resources have been immensely useful. However, by now we have almost exhausted their use. We need a new large scale and reliable discourse and dialog structure annotated corpora. Although there have been some attempts to construct such resources, they are small scale and not incremental, i.e., not built on top of (or associated with) existing large scale resources. Our goal is to build two such large scale resources, one in the discourse domain and one in the dialog domain.\(^2\)

In particular we propose the following two major projects.

1. Discourse Penn Treebank (DPTB): In this project we will develop a large scale and reliably annotated corpus which will encode coherence relations associated with discourse connectives, including their argument structure and anaphoric links, thus exposing a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with the discourse connectives. This annotation will be ‘on top of’ the PTB annotations as well as the predicate-argument annotations on PTB (called the Proposition Bank or Prop Bank), which will become available soon.

2. MultiFORM– Augmenting the FORM corpus with body movement, speech and intonation: FORM is a corpus of gesture-annotated videos. It was designed to be extensible in order to eventually represent the entire multimodal experience of conversational interaction. We will create this multimodal FORM (MultiFORM) by adding (1) body movement, (2) speech and syntactic structure, and (3) intonation.

In the following sections we will describe these two projects as separate projects. Although there are some interesting connections between these two efforts from the point of the resource development, it is best to treat them as separate activities. To be sure, from time to time, we will be examining the relationships between the tools developed in these two projects, especially from the point of view of the possible use of the tools developed in one project for the other project.

In the following, for each project we will provide a resource description and then a detailed description of the development of the resource together with the associated research. The prior work associated with these projects will be described in C.

The collaborating faculty at Penn will be (in alphabetical order), Aravind Joshi, Mark Liberman, Mitch Marcus, Martha Palmer, and Fernando Pereira. All have very extensive experience in

\(^2\) Although ‘discourse’ is a rather general term, we will use the term for corpora such the Wall Street Journal Corpus, the Brown Corpus, among others. We will use the term ‘dialog’ for the corpora that deal with interactions among two or more participants.
natural language processing. Liberman as the Director of the Linguistic Data Consortium (LDC) has expertise in the development, maintenance, and distribution of a wide range of language and speech resources. Marcus and his team were the creators of PTB. Palmer is leading the Prop Bank effort (jointly with researchers at BBN and NYU). Pereira is an expert in machine learning and its applications to natural language processing. He has in the past worked on some aspects of computational semantics, which have relevance to inferences in discourse. Joshi has worked on various aspects of computational syntax and semantics. He has also worked on local coherence of discourse and extension of the LTAG model (DLTAG) for discourse\(^3\).

b Discourse Penn Treebank

b.1 Resource Description

We provide below one example of a discourse from PTB and the corresponding DPTB annotation. The details of the annotation as well as the research and development issues are discussed later. We will also describe briefly the annotation implementation plan and the estimated sizes of the annotated corpora to be available during the period of the grant.

A Sample Discourse from PTB–Example 1: Each sentence appears on a new line.

Although IBM last year produced its first strong results in four years and was expected to continue to roll this year, it began faltering as early as January.

First, it had trouble manufacturing a chip for its mainframes, IBM's bread-and-butter business.

Then it had a series of smaller glitches, including problems manufacturing certain personal computers and the delay in the announcement of some important workstations.

Finally, IBM had to delay the introduction of some high-end disk drives, which account for 10% of its $60 billion of annual revenue.

None of the problems is necessarily fatal, and they aren’t all necessarily even related.

There are also other factors at work that are outside IBM’s control, such as currency exchange rates.

The strong dollar, which reduces the value of overseas earnings and revenue when they are translated into dollars, is expected to knock 80 to 85 cents off IBM’s per-share earnings for the full year.

Without that problem, IBM might have matched last year’s earnings of $5.81 billion, or $9.80 a share.

Still, investors will take some convincing before they get back into IBM’s stock in a big way.

The corresponding Discourse Penn Treebank (DPTB) tree is shown in Figure 1.

We label each discourse connective with an ID “DC\textsubscript{i}”, while including the discourse connective as a node in the tree. The purpose of this ID is to record the argument structure of the discourse connective. The current annotation system we use is as follows: For each clause C\textsubscript{j} connected by a discourse connective DC\textsubscript{m} to prior clause C\textsubscript{i}, we tag C\textsubscript{i} with an annotation tag “Arg\textsubscript{1}:DC\textsubscript{m}” and we tag C\textsubscript{j} with an annotation tag “Arg\textsubscript{2}:DC\textsubscript{m}”. We further tag each of these argument clauses with the tag “ANAPH” or “STRUCT”, thereby recording whether this argument is structural or anaphoric.

Details about the various types of connectives considered in this annotation, and the additional information associated with each node will be described in detail in Section b.2. The DPTB annotation is ‘on top of’ the PTB and Prop Bank annotations. It will be a stand alone annotation with links to the PTB and Prop Bank annotations, to be explained later in Section b.2.

Annotation plans and deliverables

1. An initial set of guidelines will be developed by the end of first 6 months. These will be used to annotate a corpus of about 200K words by the end Year 1.

\(^3\)This is a joint effort with Bonnie Webber (University of Edinburgh), who will be a visitor from time to time, Matthew Stone (Rutgers U.) and Alistair Knott (University of Otago, New Zealand, formerly at the University of Edinburgh).
Figure 1: A Discourse Structure Tree for the Sample Discourse
2. A revised version of the guidelines will be available at the end of Year 2. By the end of Year 2 a corpus of about 600K words will be annotated which will include the 200K word corpus from Year 1, updated with respect to the revised guidelines.

3. During Year 3 we will continue annotating more corpus and will also devote substantial amount of time for further developing the semantic tags associated with the connectives. By the end of Year 3 we will tag a corpus of about 1M words (the entire PTB with Prop Bank annotations). About a third of this corpus will also be annotated with the revised semantic tags information available by the middle of Year 3.

4. During the entire process the inter-annotator agreement will be evaluated every 100K words initially and then for every 200K words (see Section C for the results of some preliminary experiments).

5. In summary, by the end of the project there will be a very substantial resource, integrating the PTB and Prop Bank annotations with the Discourse Structure Bank annotations. For further details about the annotations and the related research issues see Section b.2.

b.2 Development of Resource and Associated Research

Our goal is to develop a large scale and reliable corpus with annotations that will encode coherence relations associated with discourse connectives, including their argument structure and anaphoric links, thus specifying a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with discourse connectives. Figure 1 shows the discourse structure tree associated with the discourse in Example 1. This annotation is on top of the PTB annotation and the Prop Bank annotation.

The PTB annotation for the first sentence in Example 1 is shown below.

Still, investors will take some convincing before they get back into IBM’s stock in a big way.

( (S
  (ADVP (RB Still) )
  (. .)
  (NP-SBJ (NNS investors) )
  (VP (MD will)
    (VP (VB take)
      (NP (DT some) (NN convincing) )
      (SBAR-TMP (IN before)
        (S
          (NP-SBJ (PRP they) )
          (VP (VBP get)
            (ADVP-CLR (RB back) )
            (PP-DIR (IN into)
              (NP
                (NP (NNP IBM) (POS ’s) )
                (NN stock) )
              (PP-MNR (IN in)
                (NP (DT a) (JJ big) (NN way) ))))))))
  (. . )))
)

Prop Bank annotations provide for each predicate (each sense of a verb, for example) a semantic

---

4 There is an effort by Marcu [9] for discourse annotation based on Rhetorical Structure Theory, RST. We will discuss the relationship between this resource and the one we are proposing here at the end of Section b.2.
frame specifying the arguments and adjuncts associated with the predicate. The semantic frames for the predicates for the last sentence in Example 1 are as follows.

Semantic Frame for “take”
Roleset “take, acquire, come to have”:
Roles:
Arg0: Taker
Arg1: thing taken
Arg2: entity taken from

Semantic Frame for “get back”
Roleset “recover”:
Roles:
Arg0: recoverer
Arg1: thing regained
Arg2: who had it before

Therefore the Prop Bank annotation for the last sentence in Example 1 will be as follows.

argM-DIS: Still
arg0: investors
argM-MOD: will
rel: take
arg1: some convincing
argM-TMP: before they...in a big way

arg1: they
rel: get back
arg2-into: IBM’s stock
argM-MNR: in a big way

The predicates, arguments and adjuncts in the Prop Bank annotation also include the links for these back into the PTB annotation. The DPTB annotation in Figure 1 is on top of these two annotations. Its clausal arguments and connective predicates will be linked to the PTB and Prop Bank annotations.

We have carried out some preliminary experiments (see Section C) by first parsing a small PTB corpus with a parser for a Lexicalized Tree-Adjoining Grammar (LTAG), extracted from PTB, and then extracting clauses from the parse. In the next phase the very same parser treats the clauses as units (corresponding to the lexical anchors as the units in LTAG) and parses the corpus with a Discourse LTAG (DLTAG) whose elementary trees are elementary discourse trees anchored by connectives and whose arguments are clauses extracted from the LTAG parse of individual sentences. This process is based on the architecture of DLTAG, which allows a seamless transition from sentences to discourse (see Section C). These experiments have provided a preliminary tool which can be used during the annotation process to obtain a discourse parse. Of course, this parse will most likely have errors. However, it will be easier for the annotators to edit this parse rather than construct the entire discourse parse by hand. This is analogous to the use of ‘shallow parsing’ in the construction of PTB. One of our major research objectives is to make substantial improvements to this discourse parser. This, in turn, requires a corpus based investigation of a large number of connectives. Such a study is already in progress and some preliminary results are presented in Section b.2. As our annotation work progresses our discourse parser will improve also, thereby increasing the annotation speed and reliability.
In our annotations we will consider a variety of connectives whose arguments will be annotated. The three major types are listed below. At the discourse level, discourse predicates can take their arguments either structurally or anaphorically. Also discourse connectives are realized either overtly or covertly.

1. **Subordinate conjunctions**: Subordinate conjunctions form a uniform class of predicates. Each connective has two arguments, not necessarily in a fixed order, and both arguments are structural. For example, the two arguments of *because* can appear in different orders as in
   (1) John failed the exam *because* he was lazy. (2) *Because* he was lazy, John failed the exam.

2. **Coordinating conjunctions**: Coordinating conjunctions, such as *and* take their arguments structurally. Some adverbial connectives are included here also, for example, *xx*.

3. **Adverbial connectives**: Adverbial connectives also take two arguments. One argument is identified structurally, normally the clause that contains the adverbial. The other argument is identified anaphorically. For example, in the discourse D1 below, the discourse predicate is the adverbial *then* which takes one structural argument, *he found out that he was broke*, and one anaphoric argument, *he ordered two cases of the '97*. Adverbial connectives may appear sentence initially, medially or finally. The position of the adverbial connective in the sentence affects the scope of the connective and is often associated with the information structure of the sentence.

   D1:
   (a) On the one hand, John loves Barolo.
   (b) *So*, he ordered two cases of the '97.
   (c) On the other hand, he had to cancel the order
   (d) because he then found out that he was broke.

4. **Empty connectives**: Empty connectives are predicates which ‘realize’ a coherence relation but do not appear lexically in the discourse. In the discourse D2 below, 2(a) and 2(b) are connected via a causal relation in the absence of a lexical connective. The DLTAG parser is unique in providing structural descriptions for these covert discourse relations. The two clauses are connected at the discourse level by a tree with a null anchor. This structural description will prove very useful to the annotators who will be able to annotate the arguments of the null predicate on the DLTAG parse output.

   D2:
   (a) You should not lend John any books.
   (b) He never returns them.

Initially we will start with a set of tags corresponding to (1) the types of connectives—structural, anaphoric and null and their positions—initial, medial, and final, (2) positions of arguments, for example, in an embedded clause, in the preceding sentence, or immediately preceding discourse. Of course, the annotation will also show the links for arguments into the appropriate segments using the PTB and Prop Bank annotations. Further, we will annotate certain kinds of semantic information with the connectives, besides the arguments, including the anaphoric links. One particular aspect of information that we will focus on is a set of inferences that are systematically associated with a connective. For example, the interpretation of the connective *although* involves the denial of a presupposed defeasible rule. Thus in ‘Although C1, C2’, where C1 and C2 are clauses, this
defeasible rule is roughly of the form ‘Normally, if C1 is true then C2 is not true’. By asserting both C1 and C2 ‘Although C1, C2’ contradicts this rule. Such pieces of information which are systematically associated with the discourse predicates will be annotated with appropriate tags. This annotation together with the linking of C1 and C2 (in the although example above) into the PTB and Prop Bank annotations provide enough information to draw the necessary inferences. The annotation itself does not directly represent this information but it provides a semantic frame and links into the PTB and Prop Bank annotations for deriving the inferences by a user of DPTB. We have carried out a preliminary study of such frames and a corpus based study of such frames will be an important part of our research associated with the DPTB project. Specification of these frames, as well as the frames associated with the argument structures of connectives, including anaphoric links, will also help the annotators during the annotation process by letting them judge quickly and accurately the relevant roles played by the surrounding context for each connective in a discourse corpus, enabling them to distinguish the arguments for a connective from the surrounding clauses. This is very similar to providing argument-adjunct frames for a verb to the annotators of Prop Bank.

Returning to the PTB discourse, in Example 1 and the associated discourse structure shown in Figure 1, we have labeled each clause connected by a discourse connective in the tree with an ID ‘Ci’. The clauses corresponding to each ID are as follows:

ID=C1: IBM last year produced its first strong results in four years and was expected to continue to roll this year,
ID=C2 : it began faltering as early as January.
ID=C3: it had trouble manufacturing a chip for its mainframes, IBM’s bread-and-butter business.
ID=C4: it had a series of smaller glitches, including problems manufacturing certain personal computers and the delay in the announcement of some important workstations.
ID=C5: IBM had to delay the introduction of some high-end disk drives, which account for 10% of its $60 billion of annual revenue.
ID=C6: None of the problems is necessarily fatal
ID=C7: they aren’t all necessarily even related.
ID=C8: There are also other factors at work that are outside IBM’s control, such as currency exchange rates.
ID=C9: The strong dollar is expected to knock 80 to 85 cents off IBM’s per-share earnings for the full year.
ID=C9i: which reduces the value of overseas earnings and revenue
ID=C9ii: they are translated into dollars,
ID=C10: Without that problem, IBM might have matched last year’s earnings of $5.81 billion, or $9.80 a share.
ID=C11: investors will take some convincing before they get back into IBM’s stock in a big way.

The tree in Figure 1 is the complete discourse parse in the DPTB for this discourse, the discourse connectives and their arguments are determined, and further, their structural and anaphoric connections are also determined. In addition, certain semantic information systematically associated with the discourse connectives are also annotated, as illustrated before for the connective although, which connects C1 and C2 in the example above.

Other more complex examples in this discourse are the connectives first, then, and finally in sentences 2, 3, and 4 respectively. There are variety of ways to build the semantic frame for the list relation imparted by this series of connectives. One possibility is to give all three connectives the same structure consisting of three arguments: ”what’s before me”, ”what’s me”, and ”what’s after me”. Then what would vary is the way those arguments are filled. “First” would require all three arguments to be filled, and its first argument would be a set, which could be inferred from
or already available in the prior discourse. “Then” (or “second”) would allow “what’s after me” to be empty, and “Finally” would require “what’s after me” to be empty.

To be sure, there will be some cases of uncertainty that are likely to remain. For example, still in the last sentence of Example 1 provides an example of this remaining ambiguity. In the DPTB tree (Figure 1) we have shown its anaphoric argument to be the immediately prior clause, C10. A viable alternative exists, however, namely, C6. If we want to always produce a single gold standard, guidelines will be needed, such as selecting the closest antecedent. On the other hand, some ambiguity is unavoidable, perhaps even desirable for the purpose of achieving multiple communicative goals at the discourse level.

Related work:

Efforts to annotate discourse structure started as a way of providing empirical justification for different theories of discourse structure (e.g., [5, 12], with community interest organized under the Discourse Resource Initiative (DRI). Unfortunately, although many committee individuals devoted a great deal of time and energy to the work (e.g., [3]), the results – being tied to particular theories of what gives structure to discourse – have not been widely used, though they remain a resource for the future.

There is an effort that is closest to what we are proposing here. This is a resource developed by Marcu [9] based on the Rhetorical Structure Theory (RST)\(^5\). The RST principles [7] include: (1) that adjacent units of discourse are related by a single rhetorical relation that accounts for the semantic or pragmatic (intentional) sense associated with their adjacency; (2) that units so related form larger units that participate in rhetorical relations with units that they themselves are adjacent to; and (3) that in many, but not all, such juxtapositions, one of the units (the satellite) provides support for the other (the nucleus), which then appears to be the basis for rhetorical relations that the larger unit participates in.

Given these principles, the two tasks carried out in RST annotation involve identifying the elementary units that participate in relations and labeling the relations that they participate in. The two tasks are not independent. For example, a relation (attribution) postulated between the specification of a speech act (e.g., “Riordan said”) and its content specified as direct or indirect speech (e.g., “We must expand the vision of our party”) means that a subject-verb fragment must be marked as elementary discourse units if the object of the verb is direct or indirect speech.

Marcu’s RST-annotated corpus differs from the proposed annotation effort in three main aspects:

1. The basis for annotating a rhetorical relation between two elementary or derived units is absent in the RST-annotated corpus. Even though there is a strictly ordered protocol to follow [8] in assigning rhetorical relations, the corpus contains no record of either the particular basis on which a rule from the protocol such as “If the relation is one of EXPLANATION, assign relation EXPLANATION”, is taken to hold, or why the conditions for earlier rules in the protocol were taken to fail to hold.

In DPTB we are proposing to annotate all and only the arguments of discourse connectives – adverbials, prepositional phrases and conjunctions – so that the basis for a coherence relation being taken to hold is clear: that is, the discourse units involved are seen as arguments of the higher-order predicate associated with the connective. (The precise semantic nature of that relation may be ambiguous – e.g., whether the relation conveyed by “then” is one of temporal ordering of events or logical consequence of propositions. But existing lexico-syntactic annotations and annotations of clausal predicate-argument relations currently in progress (Prop Bank) will provide a solid basis for disambiguation efforts.

\(^5\) [http://www.ldc.upenn.edu/Catalog/LDC2002T07.html], about 170K words from WSJ
Overt connectives are often missing in natural language discourse. In these cases, the discourse relation holding between two units is inferred on the basis of semantic and pragmatic information. The RST annotation contains no information that would be useful for computing the intended inference. In DPTB the annotation schema that we propose has two major advantages in this respect: a) it is built on the DLTAG parse (seamlessly connected to the sentence parses) which provides structural descriptions for the empty connectives and b) the discourse annotations link up to existing syntactic and semantic annotation for the sentence. Identifying the empty connectives and accessing syntactic and semantic information from the sentence level are crucial steps towards an automated computation of discourse relations in the absence of lexically realized connectives.

2. Because it is meant to assign a structure that covers an entire document, RST annotation requires two separate (but as we noted, not independent) annotation efforts. The actual result of such a stringent demand has been much of text being annotated with semantically empty relations such as span and joint (50% of the rhetorical relation annotations reported in [9]), which say little about what is going on in a text. The current annotation effort, while not attempting to cover as much as RST annotation, can focus annotators’ efforts on those parts of the discourse where the presence of a connective demonstrates that further analysis is needed.

3. In the RST annotated corpus, RST annotation of elementary discourse units, derived discourse units and rhetorical relations is the only text annotation, and thus bears the entire burden of supporting Language Technology algorithms derived from the text. The PTB annotation effort will be an additional layer on top of text already annotated with syntactic structure (PTB) and being annotated with predicate-argument relations (Prop Bank). These layers will be linked, and both their presence and their linkage provide a richer substrate for the development and evaluation of practical algorithms.

4. (Shall we say something about the final size of PTB as compared to the RST corpus?)

We do not downplay the importance of having a large annotated corpus of coherence relations associated with adjacent discourse units. In fact, we believe that the task of producing such a corpus can be made easier by having already identified the coherence relations (i.e., higher-order predicate-argument relations) associated with discourse connectives. They can then be factored into the calculation or removed from the calculation, as appropriate [18].

In summary, developing a large scale reliable Discourse Penn Treebank (DPTB) is the primary goal of this project. The resource will be based on a clearly identified level of discourse structure that will be integrated with existing large scale resources such as the PTB and the Prop Bank, both in terms of providing hooks to these other annotations but also a seamless transition from the sentence structure to the discourse structure. The latter will be achieved by the use of the DLTAG model, the parser for which also serves as a powerful tool for increasing the annotation speed as well as its reliability. This strategy is parallel to the one adopted by the PTB group, leading to its wide acceptance in the natural language processing community. We hope to achieve a similar success at the discourse level.

Applications

PTB has been widely used in natural language processing with applications ranging from parsing, information extraction, question-answering, summarization, machine translation, generation systems, as well as in corpus based studies in linguistics and psycholinguistics. The Prop Bank which is under construction and which is integrated with PTB is expected to have at least the same range of applications with the quality and coverage of the results going well beyond what has been
possible with the use of PTB. This is due to the added ‘knowledge’ incorporated in the Prop Bank. Since the Discourse Penn Treebank (DPTB) will provide a substantial level of discourse structure information, DPTB (together with PTB and Prop Bank) will raise the bar very substantially with respect to the quality and coverage achieved in the applications mentioned above. We will give one simple example to make our point. At present one technique in question answering is as follows. For a given query the system returns the sentences (say, a set R) from a document (even from a set of documents) that have the same relation(s) (predicate(s)) as in the query. Further, it may also return all other sentences which mention entities that are coreferential with the entities mentioned in the sentences in R. This collection of sentences with a minimal smoothing forms a very useful response to the initial query. With DPTB, which provides argument information for connectives as well as anaphoric links, there is a possibility of retrieving a set of sentences which are arguments of connectives in the set R including the anaphoric links associated with connectives. It might even be possible to use the same architecture as the one developed for entity coreference in a question-answer system [11].

c MultiFORM: Augmenting the FORM corpus with body movement, speech and intonation

c.1 Resource Description

A sample MultiFORM annotation

The FORM annotation scheme[10] is a useful tool which allows for multi-tier gesture annotation of videos of speakers or signers. As an example, below, we provide a video sequence and its corresponding FORM annotation. We then show the MultiFORM annotation for the same video sequence.

Figure 1’s four stills are from a video sequence of Brian MacWhinney teaching a research methods course at Carnegie Mellon University. The FORM annotation of the video, from timestamp 1:13.34 (1 minute 13.34 seconds) to timestamp 1:14.01 is shown in Figure 2. It shows the view of the data that a particular tool, Anvil[6], presents to the annotator. However, FORM uses annotation graphs[2] as its logical representation of the data. So regardless of which annotation tool is used, FORM’s internal view is the annotation graph given in Figure 3.

Figure 2: Snapshots of Brian MacWhinney on January 24, 2000

In the “Results from Prior Support” section, both FORM and Annotation Graphs (AGs) are presented in greater detail. However, for now note, first, that an annotation graph is a directed acyclic graph (DAG) such that the nodes represent timestamps of some given signal and the arcs represent some linguistic event that spans the time between the timestamps. Second, note that FORM uses vectors of attribute:value pairs to capture the gestural information of each section of
the arms and hands. In Figure 1, then, the arc labeled $\text{Hand and Wrist. Movement}$ from 1:13.34 to 1:13.57 encodes the kinematics of Brian’s moving his right hand or wrist during this time period, and the arc from 1:13.24 to 1:13.67 encodes a change his right hand’s shape.\footnote{For the example given in Figure 1, Brian is only moving his right hand. Accordingly, the Right. which normally would have been prepended to the arc-labels has been left off.}

The particular advantage to using AGs to encode the kinematics of gesture, or any linguistic signal, is the ease with which the annotation can be extended to include other data. The only constraint is that all the data share the same timeline. As such, we can easily extend the FORM corpus to include, body movement, speech and syntactic information, and intonation (Figure 4).

We propose to build such an extensible corpus of annotated videos to facilitate general research on the relationship among the many different aspects of conversational interaction. Additionally, further tools and algorithms to add these annotations and evaluate inter-annotator agreement will be developed. The end result of this work will be a corpus of annotated conversational interaction, which can be:
Figure 5: MultiFORM representation of example gesture. New information is represented by the bold arcs.

<table>
<thead>
<tr>
<th>Year</th>
<th>Prior Corpus</th>
<th>Current Year</th>
<th>Total</th>
<th>Type(s) of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3.5</td>
<td>3.5</td>
<td>G, B</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>2.5</td>
<td>6</td>
<td>G, B, S</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>1.5</td>
<td>7.5</td>
<td>G, B, S, I</td>
</tr>
</tbody>
</table>

Table 1: Summary of projected corpus development in number of hours of data. (G=Gesture, B=Body Movement, S=Speech, I=Intonation)

- extended to include new types of information concerning the same conversations; as new tags and coding schemes are developed—discourse-structure or facial-expression, for example—new annotations could easily be added;
- used to test scientific hypotheses concerning the relationship of the paralinguistic (gesture, facial expression, intonation, etc) aspects of communication to meaning;
- used to develop statistical algorithms to automatically analyze and generate these paralinguistic aspects of communication (e.g., for Human-Computer Interface research).

Annotation Plans and Deliverables

An initial set of guidelines concerning the addition of body movement to the FORM corpus will be completed by the first 6 months of year 1. These will be used to annotate roughly 3.5 hours of videos of conversational interaction. By the end of year 2, additional guidelines will be developed for speech transcription and syntactic information. These will be used to add additional information to the year 1 data, as well as to create new data using the updated guidelines. By the end of year 2, we estimate that we will have approximately 6 hours total of dialog information annotated. Similarly, by the end of year 3, we will have added intonation information to the year 1 and 2 data and create new data according to the year 3 guidelines. By the end of year 3, we expect to have a total 7.5 hours of annotated videos of conversational interaction in varied settings. Table 1 summarizes the above information, which is given in greater detail in the next section.

Interannotator agreement data will be collected and evaluated throughout the whole three years. To do this, we will randomly assign 10% of the video data to more than one annotator. Evaluation of inter-annotator agreement will take place 1, 3, and 6 months after the completion of each set of updated guidelines.
a.2 Development of Resource and Associated Research

Year 1—Gesture and Body Movement: FORM already contains tracks for head and torso movement, but further study is needed to develop the best set of parameters to capture the salient information. The first 6 months will be spent on further experimentation with different versions of head and torso annotation-schemes, as well as developing standards by which to judge correctness and agreement of FORM with this new information added. We believe that, by the end of this time we can begin in earnest annotating new videos with FORM plus head and torso information. Further as the FORM team is already quite experienced with FORM annotation techniques, by the end of the first year, we should have a corpus containing approximately 3.5 hours of annotated videos of conversations. These conversations will be of various types and in various settings.

Three and a half hours might not seem like much, however we discovered that the amount of information contained in conversational gesturing is substantial—on the order of 3500 distinct attribute:value arcs per minute. In the worst case, a trained-annotator can take a week to encode the gestures of a speaker in a 90 second video clip. However, this should not be seen as daunting. Instead, we believe that this only underscores the need for the proposed corpus, viz. there is so much more information in 90 seconds of communicative interaction than we are currently capturing by only transcribing speech.

Year 2—Gesture, Body Movement and Speech: The point of building a corpus of multi-modal annotation of conversational interaction is to be able to better understand what parts the paralinguistic aspects of these interactions play in the conversation. They seem, at least, to play a part in conveying the intended meaning of the speaker. In order to discover possible correlations between meaning and the paralinguistic, then, it is necessary to be able to encode the intended meaning of the speaker. Although this is obviously impossible, we do have access to a proxy for the meaning of the speaker, viz. what the speaker said. Accordingly, for the second year of this project, we will further augment the corpus with transcriptions of the words spoken, as well as syntactic and/or part-of-speech information. The exact form of these annotations will require some experimentation.

By the end of Year 2, we expect to have a more experienced team in place. Additionally, for the speech transcription and syntactic information, we will be calling upon the wealth of expertise at the University of Pennsylvania:

- **The Linguistic Data Consortium (LDC)** has produced many corpora containing transcribed speech. We will call upon these resource when adding transcription information. Again, since the underlying data-structure is an annotation graph, adding this information should present no problems.

- **The Department of Computer and Information Sciences and The Institute for Research in Cognitive Science (IRCS)** both have rich resources for automatically extracting syntactic/part-of-speech information from these transcriptions. As such, adding the syntactic/part-of-speech information should also present minimal difficulties.

Given the above, we expect to spend the first 3 months of Year 2 developing coding standards for the new information. The second 3 months will be used to add this new information to the

---

7 An experienced annotator can create approximately 3 secs of annotation per hour. He/she can annotate at most for 6 hours per day, generating 18 secs/day. The FORM team will consist of the equivalent of 6 full-time annotators each working 5 days per week. The weekly output will, then, be 540 secs. Assuming only 4 weeks per month to be conservative, the monthly output will be 2160 secs. So, for 6 months, we can hope to see 12,960 secs, which is 3.6 hours.
old data. The last 6 months will be used to generate new data using the new Year 2 specification. Given that this specification will be more complex than the Year 1 specification, we conservatively estimate that an additional 2.5 hours of data will be created in Year 2. This will give us a total of 6 hours of data containing FORM, head and torso, and transcription and syntactic information by the end of Year 2.

**Year 3—Gesture, Body Movement, Speech and Intonation:** Given different discourse contexts and intonational patterns, the same sentence, $S$, could be interpreted in different ways. We might, for example, assume the speaker is asserting the content of $S$, or we might understand that the speaker is being sarcastic, and is, therefore, asserting not-$S$. Intonation, then, plays a significant role in conveying meaning, and we will add this information to the corpus in Year 3. Again, we will call upon the significant expertise at the University of Pennsylvania and the LDC when experimenting with the best representation of intonation to use for annotation.

As in Year 2, we will spend the first 3 months of Year 3 developing coding standards for the new information. However, since the developed corpus will be quite substantial by this point, we estimate 5 months will be necessary to add the Year 3 additions to the Year 2 data. The last 4 months of Year 3 will be used to annotate new data with the Year 3 specification. With the complexity of the Year 3 specification and the size of the pre-existing corpus, we expect to only produce approximately 1.5 hours of data. So, by the end of Year 3, we expect to have approximately 7.5 hours of annotated video in the corpus—containing FORM, head/torso, transcription/syntactic, intonation/pitch contour information.

**Associated Research**

In order to build a useful, multi-modal corpus for human communication research, tools capable of not only annotation, but also of searching and verifying, must be built. The FORM project has the beginnings of a tool, FORMTool, which allows for easy input of gestural information while viewing video. The problem is that there is no easy way for the annotator to assess the ”correctness” of his/her annotation. Inter-annotator agreement studies help assure that the annotators are all creating similar data, but this does not guarantee that these data accurately represent the phenomena. With a strong annotation-manager, and a very specific coding manual, a team might achieve high inter-annotator agreement scores, but still not have sufficiently captured the phenomena in question.

To deal with these issues, we propose to concomitantly do research concerning the following.

- **Visualization and Animation Tools** which will “play back” an annotation. This will allow the annotator to better judge how well he/she has captured the linguistic phenomenon in question.

- **New Metrics for Inner-Annnotator Agreement.** As mentioned in the “Results from Prior Support” section, our current numbers are based on the bag-of-arcs technique. However, as the scores there indicate, often annotators agree to a large degree on structure, but differ only on exact beginning or ending timestamp, or on the value of an attribute. Unfortunately, small differences in timestamp and value are judged incorrect to the same degree as large differences. Visual feedback, as just described, will allow us to discover whether small differences in coding actually have little difference visually. If this proves to be the case, then we will need to experiment with more geometrically-based measures of similarity, e.g., distance in n-dimensional space.

- **Augmented Search Algorithms for Annotation Graphs.** The annotation-graph community has already begun research into the most efficient ways to search AG data[1] But, as
we add richer information, we need to extend the search capabilities to allow researchers fast access to this complex data. An example would be the need to search for all gestures similar to the one given in our example. Further, the researcher might want to then search those results for gestures which accompany certain syntactic or intonational structures.

This research will be carried out by the current FORM research team, as well as by one or two new graduate students. The animation and search-algorithm research, at least, is of dissertation difficulty. And, as the research progresses, we expect that more research questions suitable for dissertation work will emerge. Additionally, up to five part-time undergraduate engineering students will assist in this research as part of their senior-project work.

**Conclusion: Applications to HLT and HCI**

The augmentation of FORM to include richer paralinguistic information (Head/Torso Movement, Transcription/Syntactic Information, and Intonation/Pitch Information) will create a corpus that allows for research that heretofore we have been unable to do. It will facilitate experiments that we predict will be useful for speech recognition, as well as other Human-Language Technologies (HLT). As an example of similar research, consider the work of Francis Quek et al. [13]. They have been able to demonstrate that gestural information is useful in helping with automatic detection of discourse transition. However, their results are limited by the amount of kinematic information they can gather with their video-capture system. Further, an augmented-FORM corpus will contain much more specific data and will allow for more fine-grained analyses than is currently feasible.

Additionally, knowing the relationships among the different facets of human conversation will allow for more informed research in Human-Computer Interaction (HCI). If one of the goals of HCI is to have better immersive-training, then it will be imperative that we understand the subtle connections among the paralinguistic aspects of interaction. A virtual human, for example, would be much better if it were able to understand, and act in accordance with, all of our communicative quirks.

Having an extensible corpus such as we describe in this proposal is a first-step that will allow many researchers, across many disciplines, to explore these and other useful ideas.

**C  Results from Prior NSF Support**

**a  Results Relevant to the Discourse Penn Treebank Project**

**Discourse LTAG (DLTAG) and the DLTAG Parser:**

A discourse parsing system for a lexicalized Tree-Adjoining Grammar for discourse, DLTAG, has been described in [4]. DLTAG describes integration of sentence and discourse level processing. Our system is based on the assumption that the compositional aspects of semantics at the discourse-level parallel those at the sentence-level. This coupling is achieved by factoring away inferential semantics and anaphoric features of discourse connectives. Computationally, this parallelism is achieved because both the sentence and discourse grammar are LTAG based and the same parser works at both levels. The approach to a LTAG for discourse has been described in [14, 15, 16, 17, ?].

A DLTAG posits two kinds of elementary trees: *initial* trees, which encode predicate-argument dependencies, and *auxiliary* trees, which are recursive and modify and/or elaborate elementary trees. All structural composition is achieved with two operations, *substitution* and *adjunction*.

---

Clauses connected by a subordinating conjunction form an initial tree whose compositional semantics is determined by the semantic requirements of the subordinate conjunction (the predicate) on its arguments (the clauses). Auxiliary trees are used for providing further information through conjunction. They can be anchored by adverbials, by conjunctions like and, or may have no lexical realization. Furthermore, a discourse predicate may take all its arguments structurally, as in the case of subordinating conjunctions, or anaphorically, by making use of events or situations available from the previous discourse, as in the case of then.\(^9\) This division between the compositional part of discourse meaning (projected by the tree structures) and the non-compositional contributions due to general inferencing and anaphora is a key insight of the approach to an LTAG for discourse. It simplifies the structure of discourse and extends compositional semantic representations from the sentence level to the discourse.

An initial implementation of the DLTAG system is described in [4]. Discourse structure is derived in two passes of parsing. In the first pass, the sentences in the discourse are parsed, whereas discourse parsing is done in the second pass. This two pass implementation achieves considerable simplification over a single pass parsing, especially in terms of the parsing time and space requirements that would result from using both the sentence-level and the discourse-level grammar at the same time.

The input discourse is submitted to the XTAG parser, which parses each sentence in the discourse with reference to the sentence grammar. The output derivations (one derivation each for each of the sentences) are then submitted to a tree extractor, which extracts the basic discourse constituent units from each sentence derivation. The basic discourse units constitute the elementary trees lexicalized by discourse connectives in the sentence-level grammar, and the derivation (and derived) structures associated with the clausal units. In the next step, the sentence-level elementary trees anchored by the connectives are mapped by a tree mapper to their corresponding elementary trees in the discourse grammar. The output of the tree mapper, together with the clausal units and the input discourse, is then used to construct a discourse input representation that consists of a sequence of lexicalized trees with the extracted connectives and clausal units as the lexicalizing elements.

Some preliminary work on annotating connectives:
In this annotation work our concern was trying to see if we could identify what the arguments were, and whether or not these arguments could be retrieved via discourse structure. Each annotator annotated 3 distinct connectives. We chose three sets of connectives each providing a different type of discourse relation: concessive: nevertheless, yet, however, resultative: so, as a result, therefore, additive: in addition, moreover, also. 75 tokens of each connective were tagged as follows: First, the size and location of the clearly structural argument was marked, i.e. the syntactic material it modified. Any preceding punctuation (when it began a new sentence or clause) or connectives (e.g. and therefore or but nevertheless) was annotated. In addition, the position of the connective within the clause—initial, final or medial—was indicated. The size and location of the left argument was also marked. The syntactic type of this argument was indicated, i.e. whether it was a main clause, multiple main clauses, a subordinate, relative, or sentential complement clause, or a syntactic phrase smaller than the clausal level (marked XP). Later, 100 connectives were tagged for one from each: in addition, as a result, nevertheless.

So far within our system it seems best to treat these connectives as anaphoric rather than structural. Although the current results are limited in scope, already distinct patterns can be detected in the behavior of these connectives. These patterns include whether the connectives can combine with other connectives; the possible size and syntactic type of the anaphoric argument

\(^9\)Our use of the term anaphora does not include anaphoric relations such as those established by pronouns and definite descriptions.
(multiple clauses, single clause, or less than a full clause); how likely it is to find the argument intrasententially vs. intersententially. With a much larger corpus, the interaction between these patterns seems likely to be statistically significant in such a way that it could be used as input to train an anaphora resolution algorithm.

b Results Relevant to the MultiFORM Project

Results from Prior NSF Support

Annotation Graphs:
As described in [2], annotation graphs are a formal framework for “representing linguistic annotations of time series data.” AGs do this by extracting away from the physical-storage layer, as well as from application-specific formatting, to provide a “logical layer for annotation systems.” An annotation graph is a collection of arcs and nodes which share a common timeline, that of a video tape, for example. Each node represents a timestamp and each arc represents some linguistic event spanning the time between the nodes. The arcs are labeled with both attributes and values, so that the arc given by the 4-tuple (1,5,Wrist Movement,Side-to-side) represents that there was side-to-side wrist movement between timestamp 1 and timestamp 5. The advantage of using annotation graphs as the logical representation is that it is easy to combine heterogeneous data—as long as they share a common time line. So, if we have a dataset consisting of gesture-arcs, as above, we can easily extend this dataset by adding more arcs representing discourse structure, for example, simply by adding other arcs which have discourse-structure attributes and values. Again, this allows different researchers to use the same linguistic data for many different purposes, while, at the same time, allowing others to explore the correlations between the different phenomena being studied.

FORM:
The FORM annotation scheme was developed in order to capture the kinematic information of gesture from videos of speakers. This is done by annotating the video with geometric descriptions of the positions and movements of the upper and lower arms, and the hands and wrists[10]. FORM uses AGs as its logical representation.

Before FORM, recording kinematic information required special laboratory equipment in order to motion-capture the data. Usually this requires special sensors to be attached to the speaker in 10-15 key points on the body. These sensors have cables which physically impede motion, as well as create a very unnatural feel to the conversation itself. Subjects are self-conscious of this equipment, and of being in a laboratory. In this way, it is difficult to get natural communication. However, by combining a geometrically-based tag set with annotation techniques which have been very successful in other fields (e.g., NLP), FORM has succeeded in being able to record the kinematics of gesture without the expense and unnaturalness associated with conventional methods. Additionally, with motion-capture methods, one is only able to capture conversations staged for a specific lab session. The FORM method allows a researcher to capture kinematic information from any video-recorded conversation. This allows us to take advantage of vast amounts of already-recorded conversational data. This aspect is invaluable if we want to study the gestures of other cultures or linguistic groups in natural, unimpeded settings. And, even when doing controlled experiments, the setting can be almost natural. The only addition to the natural setting would be video cameras, which are now quite small and can be easily hidden.

Preliminary Inter-Annotator Agreement Results:
Preliminary results from FORM show that with sufficient training, agreement among the annotators can be very high. Table 2 shows preliminary interannotator agreement results from a
<table>
<thead>
<tr>
<th>Gesture Excursion</th>
<th>Precision</th>
<th>Exact%</th>
<th>No-Value%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3.41</td>
<td>4.35</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>10.07</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>29.44</td>
<td>41.38</td>
</tr>
<tr>
<td>SAME</td>
<td></td>
<td>56.92</td>
<td>86.15</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>37.5</td>
<td>52.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>60</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>75.56</td>
<td>94.81</td>
</tr>
<tr>
<td>SAME</td>
<td></td>
<td>73.24</td>
<td>95.77</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>19.25</td>
<td>27.81</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>62.5</td>
<td>86.11</td>
</tr>
<tr>
<td>SAME</td>
<td></td>
<td>67.61</td>
<td>95.77</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>10.2</td>
<td>12.06</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>25.68</td>
<td>31.72</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>57.77</td>
<td>77.67</td>
</tr>
<tr>
<td>SAME</td>
<td></td>
<td>68.29</td>
<td>95.12</td>
</tr>
</tbody>
</table>

Table 2: Inter-Annotator Agreement on Jan24-09.mov

FORM pilot study.\textsuperscript{10} The results are for two trained annotators for approximately 1.5 minutes of Jan24-09.mov, the video from Figure 1. For this clip, the two annotators agreed that there were at least these 4 gesture excursions. One annotator found 2 additional excursions. Precision refers to the decimal precision of the time stamps given for the beginning and end of gestural components. The \textit{SAME} value means that all time-stamps were given the same value. This was done in order to judge agreement with having to judge the exact beginning and end of an excursion factored out. \textit{Exact} vs. \textit{No-Value} percentage refers to whether both the attributes and values matched exactly or whether just the attributes matched exactly. This distinction is included because a gesture excursion is defined as all movement between two rest positions of the arms and hands. For an excursion, the annotators have to judge both which parts of the arms and hands are salient to the movement (e.g., upper-arm lift and rotation, as well as forearm change in orientation and hand/wrist position) as well as what values to assign (e.g., the upper-arm lifted 15-degrees and rotated 45-degrees). So, the \textit{No-Value\%} column captures the degree to which the annotators agree just on the structure of the movement, while \textit{Exact\%} measures agreement on both structure and values.

The degree to which inter-annotator agreement varies among these gestures might suggest difficulty in reaching consensus. However, the results on \textit{intra}-annotator agreement studies demonstrate that a single annotator shows similar variance when doing the same video-clip at different times. Table 3 gives the \textit{intra}-annotator results for one annotator annotating the first 2 gesture excursions of Jan24-09.mov.

For both sets of data, the pattern is the same:

- the less precise the time-stamps, the better the results;
- \textit{No-Value\%} is significantly higher than \textit{Exact\%}.

\textsuperscript{10}Essentially, all the arcs for each annotator are thrown into a bag. Then all the bags are combined and the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented.
<table>
<thead>
<tr>
<th>Gesture Excursion</th>
<th>Precision</th>
<th>Exact%</th>
<th>No-Value%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>5.98</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>20.52</td>
<td>25.21</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>58.03</td>
<td>74.64</td>
</tr>
<tr>
<td></td>
<td>SAME</td>
<td>85.52</td>
<td>96.55</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>25.81</td>
<td>28.39</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>89.06</td>
<td>95.31</td>
</tr>
<tr>
<td></td>
<td>SAME</td>
<td>90.91</td>
<td>93.94</td>
</tr>
</tbody>
</table>

Table 3: Intra-annotator Agreement on Jan24-09.mov

It is also important to note that Gesture Excursion 1 is far more complex than Gesture Excursion 2. And, in both simple and complex gestures, inter-annotator agreement is approaching intra-annotator agreement. Notice, also, that for Excursion 2, inner-annotator agreement is actually better than intra-annotator agreement for the first two rows. This is a result of the difficulty for even the same person over time to precisely pin down the beginning and end of a gesture excursion. Although the preliminary results are very encouraging, all of the above suggests that further research concerning training and how to judge similarity of gestures is necessary. Visual information may need very different similarity criteria.

References


