SPECTRUM: Speech Communication and Translation Under Mobile Environments Proposal for NSF ITR Program

1 Scientific Goals and Objectives

The focus of the proposed SPECTRUM project is communication robustness in spoken language translation (SLT). In addition to conducting scientific research to advance the state-of-the-art in SLT, we will build a prototype of a mobile platform for travel related services. The platform will include spoken language translation for human-human communication as well as spoken human-machine dialogue for other information and services. Both spoken human-human communication and human-machine communication will be enhanced with fall-back modalities to smooth over communication failures. Evaluation will be based on a suite of metrics related to usability.

The context of the research is the continuing progress of the computing revolution toward mobile, wearable, and ubiquitous devices. Wireless devices and increased bandwidth provide an enormous spectrum of possibilities for communication, provided that content and services can be brought to the user in a contextually appropriate and natural manner. SPECTRUM aims to exploit the emerging power of mobility to better serve cross-cultural and cross-linguistic communication and computing needs.

SPECTRUM builds on our previous research in the NESPOLE! consortium [1]. However, it goes beyond NESPOLE! in addressing a new round of scientific challenges and in focusing on a broader notion of communication robustness that goes beyond the noise tolerance of parsing and speech recognition.

Most current efforts in spoken language translation address a rather narrow notion of robustness, targeting the capability of modules and of the system as a whole to provide sensible answers even in the presence of corrupted input, incomplete information, etc. Such a concern is natural as long as SLT systems play a passive role, and their objective is limited to translation of isolated messages from one language to another. For actual use, though, it is crucial to start addressing the question of how SLT systems can be designed to play a more active role in securing achievement of communicative goals. Communication robustness encompasses, in our view, all the ways in which communication can fail: communicative goals misalignments, misunderstandings among the parties, problems due to the limitations of the system itself (e.g., HLT modules), or of the underlying networking, etc.

We identify the following facets of communication robustness: tolerance to noise; adaptability and learning; detection of communication failure; seamless integration of fall-back modalities for repair; seamless integration of services using different modalities (e.g., human-human translation and machine-human navigation assistance); and absorbing information from more modalities (e.g., prosody and information about the emotion of the speaker). Also, we are moving from evaluations of translation accuracy to evaluations of system usability. Each of the new scientific challenges of SPECTRUM addresses one of these desiderata. Following is a list of the research areas of SPECTRUM with indications of how they contribute to communication robustness.

- In speech recognition: tolerance to environmental noise, speaking rate, stress and emotion (*addressing noise tolerance*); new work on the optimal integration of multilingual recognizers (*addressing adaptability and learning*); and detection of emotions (*addressing absorbing information from more modalities*).

[1] *Nespole* (`NES-po-lay`) means literally *loquat fruit* in Italian. As an exclamation, it means *Wow!*. See the project web-site at `http://nespole.itc.it`

- For machine translation, we will maintain our interlingua approach because of its usefulness in multilingual situations. However we propose two new automated learning techniques to enhance the human-engineered language analyzers (*addressing adaptability and learning*), and a new model of multi-engine integration (*addressing seamless fall-back*).

- A new dialogue monitor will flag communication failures (*addressing detection of failure*). The core of the dialogue monitor will be new confidence measures for the language analysis components and analysis of speech prosody (*addressing absorbing information from more modalities*).

- A new modality manager will facilitate the communication process. The modality manager will facilitate between several human agents via speech translation as well as with computer agents that supply complementary information or requested data. Taking cues from the dialogue monitor, it will suggest fall-back modalities when communication failures are detected (*addressing seamless fall-back and integration of services using different modalities*).

- A flexible design of the telecommunication architecture will be undertaken, with careful attention given to the potential and limitations of foreseeable networking solutions. We will propose architectures that can be actually deployed in the medium term, but expanded as more advanced communication systems become available.

- Our methodology for sentence-level accuracy-based evaluations will be augmented by new techniques for usability, including achievement of communicative goals, platform adequacy (footprint and bandwidth), and optimality of selecting and integrating different services (such as translation and navigation).

SPECTRUM is designed to be a joint collaborative project between our group at the Language Technologies Institute at Carnegie Mellon and three European research groups at ITC-irst in Italy, Université Joseph Fourier in France and the University of Karlsruhe in Germany. The project consortium will also include Aethra - an Italian tele-communications commercial company. It follows on the footsteps of the currently in progress NESPOLE! project involving the same consortium members. NESPOLE! is funded jointly by the European Commission Fifth Framework (funding the European groups) and the NSF MLIAM Program (funding our group at CMU). A parallel proposal for the SPECTRUM project has been submitted to the EC (IST-2001-34738), and is appended as an addendum to this document. This proposal submitted to the NSF-ITR Program requests funding for Carnegie Mellon's participation in the project.

In this proposal we focus specifically on the core scientific research issues for which our group at Carnegie Mellon will have primary responsibility. Additional research issues for which the European partners have primary responsibility are described in greater detail in the proposal submitted to the EC. It is our belief that the investigation of these core research questions identified above can be carried out to a large extent independently, in case European funding is not secured. We believe the research described clearly fits under the ITR focus area of "Augmenting Individuals and Transforming Society", particularly as it aims to design and develop a multi-lingual system "to serve the needs of multi-national industry, collaborating science teams, or virtual cultural exchanges".

2 Existing Research Framework

The proposed SPECTRUM project builds on a long-term existing collaboration between our group at Carnegie Mellon and the three European research groups at ITC-irst in Italy, UJF in France and the University of Karlsruhe in Germany. We have been working closely together for the past seven years within the C-STAR consortium[2] and in the past 18 months have been jointly carrying our the NESPOLE! project. The SPECTRUM project aims to further this collaboration with a new focus on communication robustness in multi-lingual speech-to-speech communication in wireless environments. This section briefly describes the NESPOLE! framework that we will be building on.

[2]See the consortium web-site at `http://www.c-star.org`

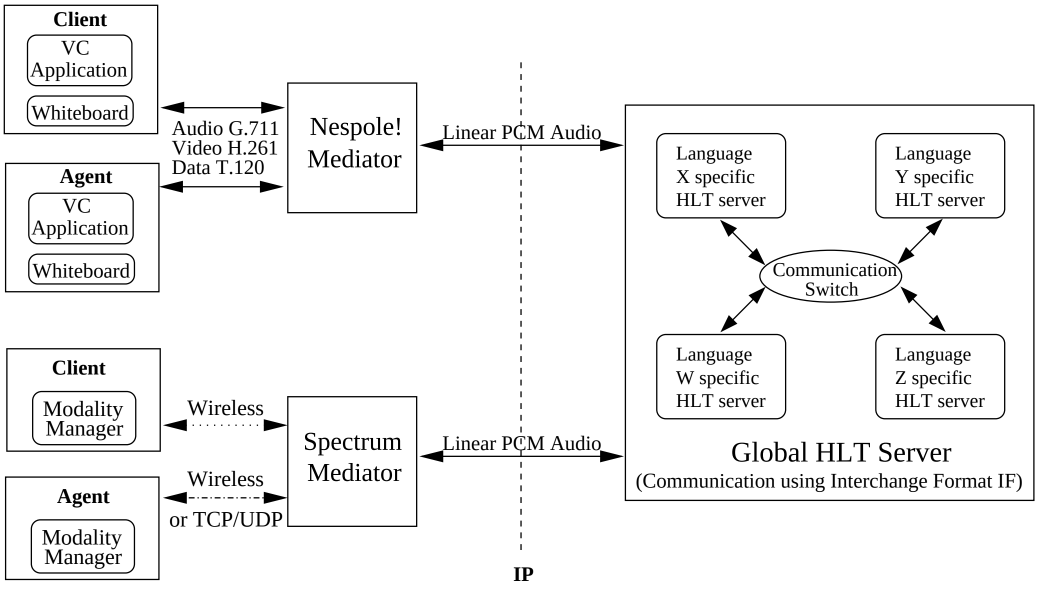


Figure 1: General Architecture Infrastructure for NESPOLE! and SPECTRUM

The main goal of the NESPOLE! project is to advance the state-of-the-art of speech-to-speech translation in a real-world setting of common users involved in e-commerce applications. The speech-to-speech translation approach builds on previous work of the C-STAR consortium. The prototype system developed in NESPOLE! is intended to provide effective multi-lingual speech-to-speech communication between Italian travel agents and German, French and English clients within broad, but yet restricted domains. The first showcase, currently in final stages of development, is in the domain of tourism and travel information.

2.1 The NESPOLE! Architecture

The SPECTRUM Project will build upon the existing architectural infrastructure developed in the course of the NESPOLE! Project. A main feature of this architecture is the clean separation of the speech-translation components of the system from the actual communication system between the two parties participating in the dialogue. The speech translation sub-system is organized into language-specific Human Language Technology (HLT) servers that can be physically distributed and that communicate with each other over the internet. Communication between the client and agent on the other hand is facilitated by a dedicated module — the *Mediator*. In NESPOLE! , the mediator is designed to connect the parties using standard video-conferencing applications. In SPECTRUM , a new mediator module will be developed to support the wireless communication between the mobile device on the client side and the stationary apparatus on the agent side. We believe that the architecture of the HLT servers will not require any significant modifications. Figure 1 shows the general architecture of the two systems.

Each language-specific HLT server consists of an *analysis chain* and a *generation chain*. The analysis chain receives an audio stream corresponding to a single utterance and performs speech recognition followed by parsing and analysis of the input utterance into the interlingua representation (IF). The interlingua is then transmitted to a central HLT communication switch (the CS) that forwards it to the HLT servers for the other languages as appropriate. IF messages received from the central communication switch are processed by the generation chain. A generation module first generates text in the target language from the IF. The text utterance is then sent to a speech synthesis module that produces an audio stream for the utterance. The audio is then communicated externally to the mediator in order to be integrated back into the communication stream between the two parties. The system also supports the display of paraphrases generated from the interlingua back into the language of the original speaker, which is very useful for verification of accurate translation.

3 Robust Speech Recognition

In the context of an autonomous system in a wireless environment, an automatic speech recognition (ASR) engine is confronted with many challenging conditions. These include varying environmental noise levels, bandwidth limits, varying speaking styles, disfluencies, and speech under stress. To obtain acceptable performance, current speech recognition technology must be adapted to effectively cope with such conditions. This adaptation requires large amounts of training data that match the working conditions in question.

Until recently most ASR systems have been trained and used under controlled conditions. However, when embedded in an immersive environment such as an autonomous tourist assistant, these limitations are unacceptable since the system must be easy to set-up and socially acceptable, and should require little maintenance or attention on the part of the user. In particular, (1) human users are reluctant to gear up with lapel microphones, let alone headsets and headgear, or to engage in enrollment phases; (2) users wish to talk freely without regard to gain settings or start/stop buttons of the behind-the-scenes recognition technology; (3) environmental noises need to be filtered out to enhance recognition, but must also be explicitly identified and tracked in order to model the environment.

We are planning to develop techniques and algorithms that allow for robust speech recognition in the face of mismatched training and testing conditions and that build on the robust spoken language technology of the JANUS project available at our lab. We propose the following approaches to achieve the goals of unrestricted open space recognition for multiple languages:

1. Robust signal enhancement for remote and variable microphone positioning
2. Models of extreme speech, for example, under stress and with hyper-articulation
3. Using articulatory features in preparation for language-independent speech recognition

3.1 Robust Signal Enhancement for Remote and Variable Microphone Positioning

The key issue in developing accurate recognition of noisy and reverberant speech is to overcome the mismatch between training and testing conditions. This problem is usually solved by a separate estimation of additive and convolutional noise in the spectrum domain, and then a process of feature compensation or model adaptation through approximation. To this end, two classes of algorithms have been developed. First is speech signal enhancement, which can be considered as a form of unsupervised adaptation in the signal feature domain. This type of algorithm is simple and efficient but generally makes assumptions that limit their scope of application. Second is acoustic model adaptation, which is usually accomplished in two ways for HMM-based modeling. One way is direct adaptation of the HMM parameters. The other is indirect adaptation of a set of transformation parameters followed by adaptation of the HMM parameters using these transformations. Both methods are effective, but are generally slow to adapt or to compute.

Recently, promising new research has been conducted on model adaptation [31]. However this approach is computationally costly and therefore only applicable for small, low complexity tasks and is not scalable to LVCSR systems. In order to overcome these limitations we propose a combination of signal enhancement and model adaptation algorithms. This goal can be achieved by extending our Model-combination-based Acoustic Mapping (MAM) approach. MAM was proven to be very effective for additive noise [37], but it only applies a simple cepstral mean normalization for convolutional noise. However, both kinds of noises interact in a very complex way on speech features [Pan00] and both are very likely in the SPECTRUM scenario (e.g., environmental noises as well as remote microphone conditions, head movements while speaking, different distances to the microphone etc.). Therefore, MAM will be extended by using a small set of secondary acoustic models, which will be used to predict and map the input signal distortion. Since only the small secondary model is modified to fit the current signal conditions, this signal compensation method is a rapid on-line unsupervised model adaptation method. The large set of core acoustic models no longer needs to be re-estimated, while the resulting system should be independent from microphone positioning. The core acoustic models will be very robust since they will be trained on various speech databases, like Broadcast News (various channel conditions), English Spontaneous Scheduling Task (conversational speech), and Switchboard (telephone speech).

3.2 Models of Extreme Speech

To date, considerable progress has been achieved in speech recognition through techniques such as vocal tract length normalization (VTLN), maximum likelihood linear regression (MLLR), and speaker adaptive training. However, for highly accurate dictation applications, it is still necessary to correct errors. It is therefore crucial to investigate and understand how users react to recognition errors. In previous studies [22],[28], it has been shown that humans apply similar recovery strategies in speech interactions with machines to those that are applied in speech interactions with other humans. Speaking style for error correction often becomes more accentuated in an attempt to more clearly articulate the original mistake. Unfortunately, contrary to the user's intention, the word error rate increases significantly under such hyper-articulations. This well known phenomena is mostly due to the unmatched speaking style conditions between training and testing, and to limitations of the current phone based acoustic model approaches.

We propose to build acoustic models to handle hyper-articulated speech more robustly and therefore enable our speech components to respond more naturally to correction situations. We plan to achieve this goal in four steps. In the first step we will analyze the effects of hyper-articulated speech, i.e., which acoustic features (such as speaking rate, pitch contour, formant frequencies) are effected by hyper-articulation. In the second step we will develop a classifier for detecting hyper-articulated speech. To determine appropriate features for the statistically trained classifier we will take advantage of the results of the investigation of articulatory features (see below). In the third step we will train two sets of acoustic models, one for normal speech, one for hyper-articulated speech (emotion-adaptive training). In the forth step we plan to do an early integration of the hyper-articulated effects into the context decision tree by introducing a "hyper-articulation mode". This would allow for better parameter sharing, and thus a very efficient usage of available training data. As a result of this approach, those acoustic models which are not effected by hyper-articulation could share all available speech data, and only those acoustic models which are effected by hyper-articulation would be treated separately. This approach had already been applied successfully to handle dialectal variations [6].

The described techniques will be evaluated on various tasks to examine the robustness against speaking style variations due to stress, emotion, hyper-articulation as well as the Lombard effect. To carry out the experiments we will use public available databases like SUSAS (speech under simulated and actual stress), as well as an in-house collected database which provides us with parallel data on normal and and hyper-articulated speech.

3.3 Recognition Based on Articulatory Features and Processes

Most current speech recognition systems represent speech as a linear sequence of sound units or phones, with transitions between these units at well-defined points in time. During the training phase, the acoustic characteristics of these phones are learned by aligning the transcription with the speech waveform, assigning exactly one phone to each part of the utterance ("beads-on-a-string model" [21]). In reality, these clear-cut transitions do not exist, as human speech production is a continuous process due to the nature of our articulatory organs. To compensate for this, a typical state-of-the-art LVCSR system represents a phone in context of its neighbors and uses several thousands of these context-dependent units to model speech.

To improve the system behavior under adverse conditions these models are adapted to the new conditions. Because these adaptation techniques optimize the criterion used during the training phase of the recognizer, they usually improve the performance of the speech engine to a certain extent. However, these methods do not really lead to robustness against changes in speaking style (for example, where a realized pronunciation exhibits only certain articulatory features, which do not correspond to a canonical phoneme).

We therefore propose to model phones no longer as a linear sequence of models, but as an articulatory feature bundle. Each feature in itself is modeled by a state sequence, but the transitions do not have to happen time synchronously. For each point in time the separate probability streams coming from feature detectors are combined by multiplying their probabilities. This combination-of-streams approach has already been successfully applied to noise-robust recognition, where different acoustic models have been trained. In contrast to systems presented in [11], our approach applies this principle to feature detectors and additionally allows the different streams to segment the data differently.

As a first step, we propose to limit the asynchronicity to one state and determine the transitions for each stream separately simply by the optimization of the training criterion (Maximum Likelihood in our case). The goal is to train feature detectors, to determine suitable features streams as well as their respective stream weights, and to compare the computational effort. The completion of this step will show the validity of the concept of articulatory feature based ASR. We have already conducted initial experiments, in which we combined six feature streams with a stream that simply consists of our current 40k vocabulary English single-pass recognizer. The feature detectors were only trained on a subset of the original training data and they increased the size of the recognizer from approximately 133k to slightly less than 140k Gaussian models. The conventional HMM approach has a 14.1% error rate and is significantly outperformed by adding the six feature streams, which has an error rate of 11.8%. These results were obtained on a 20 minute in-house test-set comparable to the F0 condition of the Broadcast News data. Other groups have reported gains on small systems using feature-based approaches on noisy environments [5].

In a second step, we propose to use the resulting stream weights for speaker and channel adaptation. We will investigate the effectiveness of existing adaptation methods with the goal of matching or improving the speech recognition performance of the proposed articulatory feature stream method on a broad range of tasks. Training, adapting and testing the systems on different in-house and external tasks such as Broadcast News and Switchboard will show whether the proposed method is indeed robust. The proposed system would enable a much finer modeling, by allowing slow articulators to stay in place, while requiring fast articulators to make explicit transitions. Therefore, the confusability with other words should be reduced, improving the overall performance.

4 Learning Approaches to Spoken Language Analysis

Our primary approach to speech translation within SPECTRUM will continue to be anchored on our task-based interlingua representation. The main approach we have used in the past for analysis of spoken language into the interlingua has been rule-based using human-engineered semantic grammars [36] and a robust parser [18] designed to effectively parse with such grammars. This approach has proven to be effective for large yet limited domains such as travel planning, which can be broken down into several natural sub-domains. Rather than focusing on the syntactic structure of the input, semantic grammars list ways of expressing semantic concepts. For example, the concept of a service being available can be expressed with the phrases "*we have ...*" or "*there are ...*". Because semantic grammars focus on identifying a set of predefined semantic concepts, they are relatively well suited to handle the types of meaningful but ungrammatical disfluencies that are typical of spoken language, and are also less sensitive to speech recognition errors. Semantic grammars are also relatively fast to develop for limited domains, where the set of concepts being described is relatively small. However, they are usually hard to expand to cover new domains. New rules are required for each new semantic concept, since syntactic generalities cannot usually be fully utilized.

In the previous version of our speech-to-speech translation system we developed a way to combine modular grammars in order to overcome some of the problems associated with expanding semantic grammars to new domains. We used sub-domain grammars whose outputs were combined into a parse tree lattice. A number of heuristics are used to rank the paths through the lattice, including the likelihood of a string of words belonging to a particular sub-domain module [38],[16].

Within SPECTRUM , our goal is to fully address the issues of domain portability and the intensive manual labor inherent in the development of semantic grammars. We propose to focus on two different approaches that are based on machine-learning to significantly reduce the amount of effort required in expanding the coverage of semantic grammars to broader or new domains.

4.1 The IF Interlingua

NESPOLE! uses an interlingua-based approach with a relatively shallow task-oriented interlingua representation [15], that was initially designed for the C-STAR consortium and has been significantly extended for the NESPOLE! project. The interlingua approach is well known to be beneficial when there are more than two language pairs [20]. The NESPOLE! interlingua, called Interchange Format (IF), consists of four

$$\underbrace{\underbrace{\text{Hello.}}_{a_1}}_{\text{greeting}}\ \underbrace{\underbrace{\text{This is}}_{a_2}\ \underbrace{\text{Bob.}}_{\text{person-name}}}_{\text{introduce}-\text{self}}\ \underbrace{\underbrace{\text{I}}_{\text{super\_who}}\ \underbrace{\text{would like to book}}_{a_5}\ \underbrace{\text{a flight}}_{\text{super\_flight-type}}\ \underbrace{\text{to Frankfurt.}}_{\text{super\_destination}}}_{\text{request}-\text{action+reservation+features+flight}}$$

Figure 2: Example: Multi-level Analysis for an Input Utterance

representational components: (1) the speaker tag ("*a:*" stands for agent, "*c:*" for client); (2) the speech act (e.g. *thank*, *give-information*); (3) a possibly empty sequence of concepts, describing the focus (e.g. *+hotel*, *+room*); (4) a possibly empty list of arguments-value pairs (e.g. *room-type=double*). The following are three examples of utterances tagged with their corresponding IF label:

1. `Thank you very much`
   c:thank
2. `And we'll see you on February twelfth`
   a:closing (time=(february, md12))
3. `There is an hotel in the town`
   a:give-information+existence+accommodation (accommodation-spec=hotel, location=town)

The same interlingua formalism will be used in the SPECTRUM project, with appropriate extensions for coverage of the new domains and sub-domains.

4.2 The DA Classification Approach

Within NESPOLE! , we have begun developing a new Domain Action Classification Parser, DACP, for analyzing task-oriented speech utterances. The goal of the parser is to analyze utterances directly into our domain-action (DA) based interlingua representation. A DA is a combination of speech acts and concepts such as `give-information+existence+accommodation` in the examples above. Complete IF representations consist of a speech-act, a collection of domain concepts and a list of arguments. The DACP parser operates in two stages. In the first stage, the parser uses a phrase-level semantic grammar for analyzing the input into a sequence of arguments. In the second stage, using the sequence of detected arguments and the words in the utterance, the parser uses statistical classification methods in order to identify the speech-act and the sequence of domain concepts that form the Domain Action. An example utterance and its levels of analysis is shown in Figure 2. The utterance `I would like to book a flight to Frankfurt` is analyzed first as a sequence of phrase arguments (such as *to Frankfurt* analyzed as a `destination`). The sequence of arguments then gets mapped to a speech-act, in this case `request-action`, and a domain-concept — `reservation+features+flight`.

The first stage, parsing of argument level phrases, is done using the robust SOUP parser [18]. The phrase-level semantic grammars are still developed manually, but we expect to leverage coverage off of portions of our existing semantic grammars. We also expect the phrase-level semantic grammars to be far less domain dependent (thus far more portable) than complete semantic grammars. The second stage – speech-act and concept classification – will be based on data trainable classification technology. We will experiment with hidden markov models (HMMs), neural nets, and decision trees. These techniques are by design more robust and portable than complete semantic grammars, but their accuracy depends on the availability of adequate accurate training data.

We have so far put together a baseline prototype of the new DACP parser and have tested it within the NESPOLE! project. Semantic grammars were developed for analyzing the argument-level concepts within the NESPOLE! travel-domain scenarios. These grammars are significantly more compact and were faster to develop than full semantic grammars. The speech-act and concept sequence classification task is currently done using TiMBL — a classifier based on a "k-nearest neighbor" approach [4]. The training data is a corpus of dialogues that were hand-tagged with DAs.

Two important questions are whether the statistical methods used for classifying strings of arguments into domain actions are accurate enough, and what amounts of tagged data are required to obtain reasonable levels of performance. To assess these two questions, we recently tested the performance of the current speech-act and concept classifiers for the expanded travel-domain when trained with increasing amounts of training data. The results of a ten-fold cross-validation experiment indicate that the classifier currently detects DAs with about 50% accuracy (when the complete training corpus is used), and that the performance reaches a relative plateau at around 4000-5000 training utterances. We see these results as indicative that this approach should indeed prove to be significantly easier to port to new domains. Creating a tagged database of this order of magnitude can be done in a few weeks, rather than the months required for complete manual grammar development time. Furthermore, even with 50% correct classification, when the Dacp parser was incorporated into our end-to-end translation system, translation accuracy on an unseen test set was around 65%. We believe this indicates that when fully developed, the Dacp approach will be able to produce accuracy levels comparable with manually well developed semantic grammars.

All aspects of the Dacp analyzer require significant further research and development. The main research issues are the following:

- **Segmentation:** Developing methods for reliably segmenting utterances that consist of multiple domain actions. We will investigate using statistical information that can model the likelihood of a boundary between identified argument concepts as well as using knowledge from the interlingua specification that defines well-formedness constraints on the sequences of arguments that can belong to a single DA. Other approaches will also be considered.

- **Argument Parsing:** Investigating methods for dealing with ambiguity at the level of argument parses. The underlying Soup parser for arguments currently uses a set of heuristics for selecting a single sequence of argument parses for a given input utterance. These were originally developed for disambiguation when parsing with full semantic grammars and need to be revisited in the context of argument parsing. We also intend to consider passing multiple sequences of argument parses to the classification stage of the parser, either in the form of an n-best list or in the form of a lattice.

- **Classification Approaches:** Performing a detailed investigation of several different classification approaches and analyzing their effectiveness and appropriateness to the task of domain action classification. These will include the current k-NN approach (TiMBL), Decision Trees (i.e C5.0), Neural Networks, Statistical Language Models and possibly others. Both accuracy and classification speed need to be taken into account.

- **Feature Set for Classification:** Investigating the set of features that should be provided to the classifier for optimal classification performance. The space of features includes the words in the utterance, n-grams of words, information about whether words were parsable or not, the semantic labels of the argument parses, the detected speech-act (for concept classification), and contextual information (DA of previous utterance of the same speaker and other speaker). Additional features will also be considered.

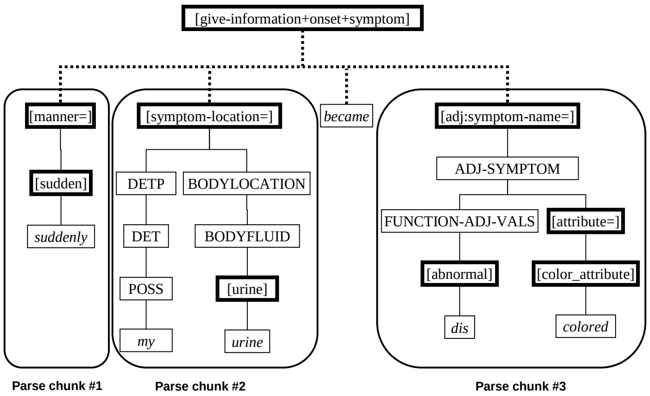
- **Concept-sequence Classification:** Investigating whether concept sequences that form the DA should be classified jointly or as individual concepts that are then combined together.

- **Use of Interlingua Specification Constraints:** Investigating how to use the the well-formedness constraints of the interlingua specification in order to augment the statistical approach used by the classifier. This includes using information about which arguments are licensed by each DA, information about which concept sequences are semantically allowed, and devising fall-back strategies for cases where the best classified DA is not legal according to the interlingua specification.

- **Portability Across Domains and Across Languages:** Evaluating the extent to which the new Dacp approach is in fact more portable to new domains and to new languages.

4.3 Automatic and Interactive Semantic Grammar Induction

To further facilitate the development of semantic grammar coverage for a domain, and in support of faster portability to new semantic domains, we plan to develop a new approach to semantic grammar induction. The premise here is that once the interlingua representation for a new domain or sub-domain has been




```
Original interlingua:
give-information+onset+symptom
(symptom-name=(abnormal,attribute=color_attribute),symptom-location=urine,
manner=sudden)

Learned Grammar Rule:
s[give-information+onset+symptom]
        ( [manner=] [symptom-location=] *+became [adj:symptom-name=] )
```

Figure 3: A reconstructed parse tree from the Interlingua

defined, it is fairly fast and straightforward to develop a small core semantic grammar for the new domain, based on a small set of example utterances. The most labor intensive and time consuming task is to then expand the coverage of the grammar to correctly parse the wide variety of ways in which concepts of the new domain can be expressed in spoken language. Grammar induction tools can therefore significantly speed up this process. Our past experience has shown that our grammar and interlingua developers can annotate utterances with their corresponding interlingua representations relatively fast and with high reliability. It is far more time consuming to expand the coverage of the semantic grammars to correctly produce the interlingua representations for new previously unseen utterances. Our key goal is therefore to develop tools for expanding the coverage of an initial core grammar by learning new grammar rules semi-automatically using partial parse information from the existing grammar and the "target representation" information provided by the interlingua. We plan to first investigate the extent to which grammar rules can be learned completely automatically, with no interaction with an actual human developer. At a second stage, we plan on adding interactivity, to fill in knowledge gaps and resolve ambiguity in a more reliable fashion. Previous work within our project [8] has already partly investigated the learning of grammar rules with user interaction.

We have already begun implementing an initial prototype of the completely automatic grammar induction component. The automatic induction is based on performing tree matching between a skeletal tree representation obtained from the interlingua, and a collection of parse fragments that is derived from parsing the new sentence with the core grammar. Extensions to the existing rules are hypothesized in a way that would produce the correct interlingua representation for the input utterance. Figure 3 shows a tree corresponding to an automatically learned rule. The input to the learning algorithm is the interlingua (bold boxes in the figure) and three parse chunks (circled in the figure). The dashed edges are augmented by the learning algorithm. Preliminary tests of the current prototype have shown promising results.

The main stages of processing can be seen in Figure 4. All of the above stages require significant research and development. The main issues to be addressed are the following:

- **Tree Matching:** The skeletal tree is constructed by a reverse mapping from the interlingua. Although the use of a semantic grammar makes such a mapping easier, in certain cases the mapping produces more than one possible skeleton. The tree matching algorithm thus needs to match parse fragments against a collection of skeletons, and give an indication of the quality of the matches.

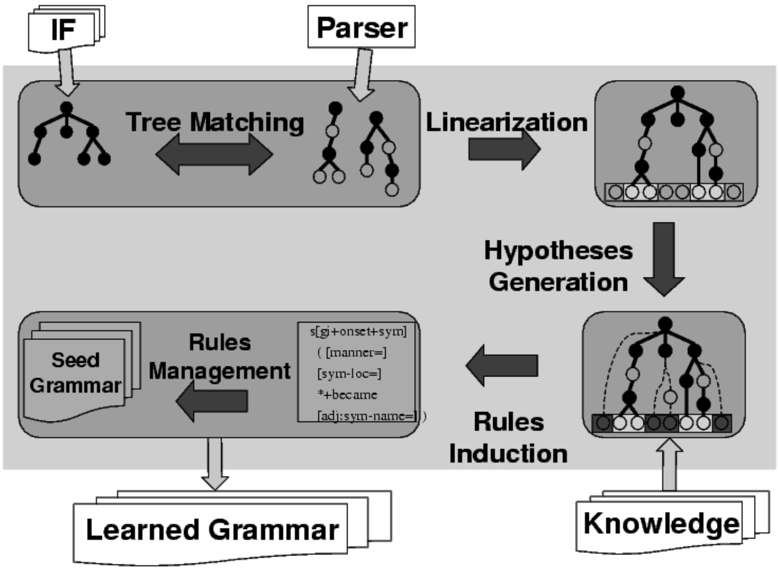


Figure 4: Stages of Processing in Automatic Grammar Induction

The algorithm could be made more robust by adopting approximate or error-correcting matching techniques.

- **Linearization:** Since the ordering of the meaning components within an interlingua is not significant, the matches between a skeletal tree and the parse fragments need to be linearized (reordered) with respect to the actual word ordering in the utterance. The process also needs to detect any cross-branch violations, where one meaning component is mapped to multiple non-contiguous words. The violations can then be eliminated by certain meaning-preserving tree transformations. Finally if more than one linearization is possible, the algorithm needs to give indication of the quality of each linearization.

- **Hypothesis Generation:** At this stage only a partial parse tree is reconstructed by attaching the matched parse fragments with the skeletal tree. For any remaining unparsed words and the meaning components which are not mapped with any words, the correspondences between them needs to be hypothesized. The process could benefit from using additional knowledge sources, such as statistical language models or ontological information, to reliably map words to their most relevant meaning components. Heuristics on which types of words/meaning components need to be mapped first could also be helpful to achieve higher accuracy.

- **Rule Induction:** At this stage a complete parse tree is reconstructed, and a set of rules which can be used to produce the tree are read off from the tree. The rules can be made more robust by taking optionality and repeatability of part of the right-hand side of the rules into consideration.

- **Rule Management:** The rules induced are inserted back to the core grammar at this stage. In updating the core grammar the algorithm should minimize the impact on the existing rules so that the performance of the combined grammar does not degenerate. It is also desirable to merge or rewrite rules to make the updated grammar compact. Theoretical results have shown that detecting rule subsumption is an undecidable problem, thus further constraints that are appropriate for a specific application domain need to be introduced to achieve these goals.

Once these issues are well addressed, we will turn our attention to how to boost the accuracy of the approach by adding user interaction for better ambiguity resolution and validation of the induced grammar rules. User interaction will play a role in the following ways:

- **Learning Confirmation:** Theoretical results have shown that grammar learning using only positive examples is not feasible [9]. To avoid over-generalization, user interactions can be introduced at the stage of hypothesis generation and rule induction to confirm the learning results. However the interactions should be designed in such a way that they do not require linguistic sophistication from the users. For example, in hypothesizing correspondences between unparsed words and unmapped meaning components, the system could generate alternative utterances by realizing the meaning components with the other words known to map to the components, and ask the user if the utterances convey legitimate meaning. The correspondences can then be established if a confirmation is obtained. Other alternative ways of taking advantages of user interactions include asking ontological questions about unparsed words, and hypothesizing the correspondences in light of the obtained ontological relations between the words and the meaning components.

- **Proactive Learning:** With user interaction the system can not only ask confirmation questions when learning from problematic utterances, but is also able to resolve grammar ambiguities by proactively planning experiments, i.e., by generating utterances from potentially ambiguous rules and asking the user for confirmation. This will enable the system to generalize the grammar even when the relevant data is unknown at the point of learning.

- **Acquiring New Concepts:** In addition to the incomplete grammar coverage, an utterance could be deemed ungrammatical because of the lack of the appropriate concepts in the interlingua. Via user interactions the system will be able to detect such errors and initiate an interactive process to acquire the missing concepts, thus making them available for later use. The interactive process itself requires another set of grammars, for which the grammar learning approach could also be useful.

The grammar induction techniques we plan to develop are complimentary to the work on the Domain-action Classification parser described earlier. The grammar induction method can be used to learn either full semantic grammars or semantic grammars limited to the argument-level representations, from which we can then rely on the DACP parser to complete the analysis process. We plan to investigate both, with the goal of devising the most effective and accurate way to leverage from these two approaches.

5 Dialogue Monitoring and Modality Management

In this section we describe our plans for a communication manager that detects communication problems and deploys recovery strategies. The communication manager has two main parts, dialogue monitoring and fall-back management of the modalities (e.g., speech, text messages, menus) available to the system. The goal of dialogue monitoring is to reliably detect failures in the communication and their causes. The goal of modality management is to actively assist the user in overcoming communication difficulties by using alternative modalities.

5.1 Dialogue Monitoring

The Enthusiast, Artwork, and Verbmobil projects [23], [24], [26], [25] have studied the use of discourse information for resolution of ambiguity and improvement of quality of SLT in task oriented dialogues. For example, if a plan inference tree of dialogue goals were available, could it be used to disambiguate the speech act of an incoming utterance and thereby increase translation accuracy? These projects experimented with a variety of techniques including plan inference, finite state models of speech act sequences, and n-gram models of speech act sequences. However, the results were discouraging. Even with reasonably high accuracy – for example, around 70% correct detection of speech acts – the improvement in translation quality was negligible. Similarly discouraging results were obtained at the 1997 Johns Hopkins Summer Workshop [10] regarding the impact of speech act detection on word accuracy for speech recognition in non-task-oriented Switchboard dialogues. [27] showed mildly positive results for using discourse knowledge in interactive repair.

While the above results appear to discourage the general utility of attempting to track the progression of the dialogue for improving translation quality, we believe that contextual information (such as speech-acts and domain-actions of previous utterances) can still play an important role in identifying concrete points within the dialogue, particularly communication failures. In SPECTRUM we plan to focus our attention specifically on monitoring the dialogue for communication failures. Discourse knowledge will play only one part in a suite of confidence measures that will be designed to detect communication trouble spots. The job of repair will then be handed over to the Modality Manager. We plan to develop and integrate a suite of strategies for detection of translation failure and communication breakdown. These will include:

- **Effective Translation Confidence Measures:** We will develop new measures for evaluating the confidence in the quality of translation outputs of our specific translation engines. The focus will be on our interlingua-based engine. Features that will be taken into account include the confidence score from the speech recognizer, parser information about words that were not incorporated into the analysis, the "distance" of the produced paraphrase from the original input, and a new measure for rating the fitness of produced interlingua representations. The improved confidence measures associated with the different translation engines will also be used to improve the multi-engine integration.

- **Explicit Detection of Communication Failure:** We will monitor the progression of the dialogue with the goal of explicitly detecting points of communication failure. Our goal will be to find reliable characteristics of points of communication failure based on a variety of indicators. These will include: (1) speech recognition and translation confidence measures for the current and previous turns; (2) specific domain-actions, and sequences of domain-actions, that may indicate communication failure; (3) prosody and intonation features. We will also attempt to characterize the cause of the failure: poor speech recognition, poor translation, deviation from the domain, etc. This information will then be passed along to the modality manager.

5.2 Modality Manager

The Modality Manager will have the task of actively assisting the user in overcoming communication failures once a failure and the cause of the failure are identified. While constrained by the limitations imposed by our mobile platform, we envision the system will have at least the following modalities at its disposal: (1) speech translation activated by voice; (2) menus of common "phrase-book" entries for translation; (3) ability to enter short text messages for translation; and (4) transmission of simple annotated hand drawings.

The modality manager will take an active role in suggesting to the user when to switch from the primary modality of speech translation activated by voice to any of the other modalities available. For example, if we detect that the cause of the failure is poor speech recognition, text message translation will be proposed as an alternative. If on the other hand speech input appears to be reliable, but the main translation approach is failing, the system can present a menu of phrase-book entries based on the speech input and allow the user to select an appropriate entry for translation. User studies will play a large role in the investigation of how the modality manager can effectively assist the user without being intrusive.

Previous work within the framework of human-machine dialogue systems has shown positive results using such active dialogue management. In a spoken dialogue system for accessing online train schedules, an adaptive version of the dialogue manager was developed that monitored ASR confidence scores to incrementally predict whether a user was having problems as the dialogue progressed. The system adapted to a more conservative set of dialogue strategies whenever the predictions classified the dialogue as problematic. An experimental evaluation demonstrated that changing the behavior of the dialogue manager based on monitoring acoustic confidence scores significantly increased the task success rate [17]. In the context of the JANUS project we have also conducted experiments on multi-modal repair of speech recognition [29].

Based on our recent experience, we believe that active management of repair modalities is also important for machine translation. In April 2001, at the conclusion of the Tongues project [1], we carried out a field-test of a SLT system running on a laptop, with regular Army officers and naive Croatian civilians as users. While the detailed analysis of the test results is still underway, our informal observation was that user interaction was clearly the weakest link. While ASR and TTS were certainly not perfect, the majority of the failures to complete tasks successfully were clearly the result of difficulties the monolingual English speakers had in assessing whether the MT component had produced correct translations in the target language. Assessing the quality of the translation involved gestures and menus. A team member who was bilingual observed that often the English-to-Croatian translation was reasonably correct, but the English-speaker would abort the translation because the available feedback suggested that the translation was not adequate.

6 Evaluation

In spite of the growing number of commercial and research-oriented MT systems, evaluation of machine translation is still a topic of much disagreement. The confusion stems in part from the diversity of machine translation applications and the multi-faceted nature of the research. Some evaluation methods are relevant for the user, whereas others are relevant for the system developer. Some are relevant when high quality is required, whereas others are relevant when gisting is satisfactory. Evaluation by humans is costly, whereas automatic evaluation may not accurately reflect human judgement.

In the course of our research we have developed a suite of evaluation methods that we find informative including end-to-end evaluations of translation accuracy on a sentence-by-sentence basis; plots of accuracy as a function of development time; plots of accuracy as a function of the amount of development data; plots of coverage as a function of training data (for corpus-based methods); and evaluation of individual system components against hand-coded gold-standard output.

Our evaluation methods have been largely accuracy-based [7], focusing on whether the meaning of a source language segment is totally and accurately conveyed in the target language. This type of evaluation is useful for tracking our improvement over time. The measure we use is percent of sentences that are accurate (we call these `acceptable`) and the percent that are both accurate and fluent (we call these `perfect`).

Accuracy Based Evaluations (ABE) continue to be useful. However, the ability of a user to complete a task (for example, getting a plane reservation) is higher than would be expected based on an ABE. For example, the ABE might be around 70% acceptable, but the users could almost always complete the task. Recently we have begun to design a Task Based Evaluation (TBE) for speech-to-speech translation [30], [14] that measures goal completion. Goals can be identified at the dialogue level (e.g., did the user get a hotel room by the end of the dialogue?) or at the utterance level (e.g., Was the utterance *How much is a double room* translated accurately enough for the interlocutor to understand that it was a question about price and that it concerned a double room?).

However, the correct notion of goal completion is difficult to pin down. First we must decide what counts as a goal. For example, suppose the question *How much does a double room cost?* is translated as *How much?*. Has a goal of requesting information about some quantity been met? Or suppose it is translated as *A double room costs*. Has a goal of talking about the price of a double room been met? Or has the communicative goal failed completely? Furthermore, utterances can be rephrased and repeated in the face of communicative failure. The repetitions and rephrased sentences should not count as new communicative goals, but as multiple attempts to achieve the same communicative goal. In [14] we formulated a scoring function for communicative goal completion that took into account whether each communicative goal ultimately succeeded or failed and how many times it was attempted before succeeding or failing. Of course, intercoder agreement on new versus repeated goals then becomes an issue.

In addition to translation accuracy and goal completion, there are likely many other factors that are useful for measuring performance, such as dialogue length. In human-machine spoken dialogues, the PARADISE evaluation framework has been successfully used to understand the relative contribution of a wide variety of such factors (representing performance dimensions such as task completion, dialogue quality, dialogue efficiency, and usability) to overall performance, across many different spoken dialogue systems [34], [33], [35], [32]. PARADISE models performance as a weighted function of a task-based success measure and dialogue-based cost measures, where weights are computed from experimental data by correlating a meaningful external criterion of usability with performance.

To date, most approaches that go beyond accuracy-based evaluations have been conducted in the context of human-machine spoken dialogue systems. For machine translation, we need an evaluation paradigm that is suitable for two humans each expressing communicative goals, but mediated by a machine. Additionally, we have to allow for a large and unpredictable number of communicative goals in each dialogue. For example, following the coding scheme we developed for [14], the dialogues we are evaluating each contain over one hundred communicative goals. Our new work on performance evaluation will extend previous work on task-based and usability-based approaches for human-machine dialogue, to the context of a computer-mediated human-human spoken translation system.

7 Showcases and Scenarios

The general scenario we are targeting is a wireless help-desk, allowing a user/customer to interact over a wireless mobile device with an agent/service provider, where the two parties speak different languages. The agent is equipped with an ordinary PC, while the customer exploits a mobile device. A typical situation would be one in which the user is travelling in the country of the service provider, and is in need of assistance with any number of travel related situations. We envision both remote and face-to-face conversations. Concrete examples include accommodation reservations, doctor-patient conversations and car mechanical road assistance.

In collaboration with the European project partners, we will design and implement two showcase prototype systems in the course of the project. The first showcase will be completed at the mid-point of the project (month 18). It will focus on the domain of medical assistance (emergency help and doctor-patient interactions) and will demonstrate the results obtained with respect to the new HLT modules (both interlingua-based and direct approaches), communication robustness and multimodality. The second showcase will be completed at the end of the project. In this showcase, we will expand our coverage of the medical assistance domain and add on the domain of mechanical car assistance. It will be devoted to demonstrating our results with respect to portability to new domains, multi-engine translation, the translation of emotions, and further advancements achieved in communication robustness.

Taken together, the two showcases aim at demonstrating the advantages and the feasibility of the proposed solutions for multilingual and multi-modal communication in a wireless scenario. Both showcases will target the wireless help-desk domain. The choice of the infrastructural and architectural solutions to be adopted will be made taking into account technical feasibility for experimental and demonstrative purposes, and portability of the developed solutions to emerging wireless platforms.

8 Results from Prior Relevant NSF Support

Spectrum will build upon extensive infrastructure we are currently developing within the scope of the Nespole! project (PI Waibel), funded jointly by the European Commission and NSF (under the NSF MLIAM program). Some current accomplishments of Nespole! were described in section 2. Our research work under Nespole! is funding in part PIs Lavie, Levin, Schultz and Waibel, three Ph.D students, two M.S. students and several undergraduate assistants. It has resulted so far in five refereed conference publications [13] [12] [2] [19] [3]. The most recent accuracy based evaluation shows around 50% end-to-end translation accuracy. Now that the HLT servers, interlingua, and mediator are in place, and preliminary user studies have been conducted, we expect to see rapid improvement in translation quality.

We are also involved in a collaborative project with a group of researchers at the University of Cairo, under an NSF International Programs grant (PI Levin). This is a small seed grant from the U.S.-Egypt Joint Technology Board. The goal of the project is to add a prototype speech recognizer, analyzer, and generator for Egyptian Arabic within the Nespole! architecture. So far, researchers from Egypt have been trained in data collection, interlingua design, and writing analysis and generation grammars.

PIs Lavie, Levin, Schultz and Waibel are also involved in the recently awarded NSF-ITR grant for the Avenue Project. The main goal of Avenue is to develop new Speech and Machine Translation technology for minority languages and languages with scarce online resources.

9 Work-Plan Summary

The following work-plan was developed jointly with our European partners and describes the work-plan for the project as a whole, including activities that were not described in detail in this proposal, but which are detailed in the European version of the proposal. We identify four major sets of activities spanning the whole temporal extent of the Spectrum project: (a) the study, development and evaluation of HLT modules (speech recognition/synthesis, interlingua-based and direct translation methods, multi-engine approaches); (b) activities related to the multimedia/multimodality issues; (c) activities targeting communication robustness; and (d) the translation of emotions. The first three lines of investigation and development will converge with the realization of Showcase-1, due by month 18, which will demonstrate (communication) robust multilingual and multimedia-based conversation in the wireless help-desk for the traveller. The showcase will be evaluated on real data in a real mobile setting, in order to stress and evaluate robustness, and quality of interaction. The results will feed the second phase, whereby improvements on all the relevant research lines will be pursued, and converge within Showcase-2, due by month 36. The second showcase will also integrate the results concerning the translation of emotions.

The following table summarizes the joint work-plan for the Spectrum project. The list of deliverables mentioned in the last column of the table can be found in the European proposal for Spectrum , which is included as an appendix.

Workpackage list							
Work-package No.	**Workpackage title**	**Lead contractor**	**Person-months**	**Start month**	**End month**	**Phase**	**Deliverable No.**
W1	Project Management	ITC-irst	76	m0	m36		D1
W2	Requirements	UJF	66	m0	m23		D2, D12
W3	Showcases development	AETHRA	80	m4	m36		D10, D17
W4	Translation of Emotions	ITC-irst	73	m4	m34		D7, D14
W5	Communication Robustness	CMU	100	m4	m34		D8, D15
W6	HLT development	UKA	230	m4	m34		D6, D13, D14,
W7	Multimodality	ITC-irst	76	m4	m34		D9, D16
W8	Assessment and Evaluation	CMU	47	m16	m36		D11, D18
W9	Dissemination and Implementation	ITC-irst	24	m0	m36		D3, D4, D5, D19, D20
	TOTAL		772				

Figure 5: Spectrum Project Work-plan

Project Management Plan

The management plan presented here is identical to the plan contained in the proposal submitted on the European side to the EC. The main goals of the management of this project are: (1) to organize the project as a whole, initiate its different activities, perform all the necessary administrative tasks; maintain the contacts and report to the Commission, the NSF and the partners; (2) to supervise the technical progress of the project, and give technical advice; (3) to insure that the project adheres to its scientific goals and satisfies user needs, in order to achieve a high exploitation potential of the project's results. An important consideration is the "transatlantic" nature of the SPECTRUM project. Thus the consortium will have to report to two funding agencies, the EC and the NSF, which differ in many important respects: cost monitoring practices, overall control on projects achievements, etc. Our experience carrying out the NESPOLE! project has been extremely positive, so we propose to maintain the same structure and methodology adopted therein, which is described in the following.

Management of the project will be assured by the partners through the following structures:

- Administrative Director
- Technical Directors (PD)
- Project Management and Technical Committee (PMTC)
- Project Exploitation Manager (PEM)
- Project Managers
- Work-package leaders

In addition to the classical structure defining managerial and technical committees, we add the following concerns: Knowledge protection; Exploitation preparation; Dissemination of results; Quality Assurance; Establishing and interacting with a User Group.

Project Administrative Director: The Project Administrative Director will be responsible for: (1) chairing the PMC; preparing and managing the management reports; (2) handling all communications with the EC and the NSF; monitoring project costs; (3) creating and maintaining the conditions necessary for successful and effective collaboration; (4) proposing and implementing the quality assurance procedures; (5) creating and coordinating the user groups; (6) planning and implementing the project's contribution to IST/HLT project clusters; (7) representing the project (or delegating the project representation to the appropriate project staff) in the occasion of cluster events and meetings.

Project Technical Directors (Project Coordinators): Given the important role of the US partner, the Consortium will have two Technical Project Directors, one for the European side and the other for the US one; they will be nominated by the (European) Prime Contractor, and by the US partner, respectively. They will be responsible for: (1) co-chairing the PMTC meetings; (2) managing the progress reports; (3) monitoring the time schedule and the timing of the related activities; (4) recommending appropriate actions to rectify delays; (5) ensuring that all project deliverables are available on time; (6) ensuring that all the resources consumed in the performance of the work are actually relevant to the specific work involved; (7) representing the project (or delegating the project representation to the appropriate project staff) at various scientific events.

The coordinating partner will operate a Project Secretariat office for the duration of the project. The office will support the project by: maintaining a central archive of all documents produced within the project; distribution of information inside and outside the project; maintaining the Project Plan and producing consolidated reports on efforts, results, schedule, and resource consumption.

Project Management and Technical Committee (PMTC): This committee will be formed by one key person of each full contracting partner (Project Manager), with the role of Administrative and Technical Manager involved in the project and by the Project Exploitation Manager. The role of the PMTC is to: (1) assist the Project Directors when carrying out their duties; (2) make sure that the activities and results thereof conform to the proposed quality standards; approve all official deliverables; (3) approve all significant changes in the project work-plan; (4) approve the Exploitation Plan; (5) establish Knowledge Protection policies; (6) assign specific responsibility to the most suitable partner representative, when new events require it; (7) monitor the technical direction of the project; (8) approve all major technical decisions: reviewing and/or amending the work-plan, the cost or time schedule under the EC Contract, the termination of the EC Contract, lay down procedures for publications and press releases with regard to the project.

Conflict Resolution: The decision-making procedure is organized as follows: each full contracting partner has a vote. Decisions will normally be taken by seeking consensus. In cases where consensus cannot be reached, decisions will be made based on majority vote.

Project Exploitation Planning Manager (PEM): . The PEM will be responsible for coordinating the overall project exploitation planning strategies and actions. He will also coordinate the preparation of a detailed Exploitation Plan. This Plan will be defined throughout the project in order to be able to support, effectively, the project operation and exploitation phase.

User Groups (UG): Two User Groups will be created and actively involved in the following project activities: (i) user requirements; (ii) system validation; (iii) system demonstration, and (iv) system exploitation. The Project Administrative Director will be directly responsible for creating and coordinating these two user groups.

Administrative and Technical Project Managers: Each contracting organization will appoint an Administrative Project Manager (APM). All official communications will be addressed to him. He will attend the PMTC meetings and also liase with it to ensure the alignment between the organization's objectives and the direction of the project. He will also be responsible for ensuring that the organization provides resources to the level specified in the project. In addition, he will provide to the project Director all the needed information regarding his organization for the preparation of the management reports. Moreover the same person will be responsible for ensuring that the organization respects the planned schedule, both with respect to activities and their results. He/she will provide to the project Director all the needed information regarding his organization for the preparation of the advancement reports.

Work-package Leaders: The work-package leader is responsible for the coordination of the activities carried out by his work-package. He/she reports to the PMTC.

Information Flow: The preparation process of a deliverable is the following: a project deliverable is prepared under the responsibility of the person appointed by the responsible organization for a specific task. The deliverable is sent to the Project Director, who submits it to the Commission.

Periodic Reports: The coordinator will supply a full report on a quarterly, semi-annual and/or annual basis, detailing the progress of the work, any problems encountered, actual expenditures (of money and manpower) versus plan, and plans for the coming year.

List of Personnel Associated with the Proposal

- **Laurent Besacier**, Université Joseph Fourier, Grenoble, France. (Senior Personnel)
- **Hervé Blanchon**, Université Joseph Fourier, Grenoble, France. (Technical Manager)
- **Christian Boitet**, Université Joseph Fourier, Grenoble, France. (Senior Personnel)
- **Jean Caelen**, Université Joseph Fourier, Grenoble, France. (Senior Personnel)
- **Roldano Cattoni**, ITC-irst, Trento, Italy. (Senior Personnel)
- **Robert Frederking**, Language Technologies Institute, Carnegie Mellon University. (Senior Personnel)
- **Roberto Giamagli**, AETHRA, Ancona, Italy. (Industry Partner)
- **Damien Genthial**, Université Joseph Fourier, Grenoble, France. (Senior Personnel)
- **Jean-Philippe Guilbaud**, Université Joseph Fourier, Grenoble, France. (Senior Personnel)
- **Alon Lavie**, Language Technologies Institute, Carnegie Mellon University. (PI)
- **Gianni Lazzari**, ITC-irst, Trento, Italy. (Project Director)
- **Lori Levin**, Language Technologies Institute, Carnegie Mellon University. (co-PI)
- **Diane Litman**, Computer Science Department, University of Pittsburgh. (Senior Personnel)
- **John McDonough**, Univerity of Karlsruhe, Germany. (Technical Manager)
- **Florian Metze**, Univerity of Karlsruhe, Germany. (Senior Personnel)
- **Fabio Pianesi**, ITC-irst, Trento, Italy. (Project Technical Manager)
- **Emanuele Pianta**, ITC-irst, Trento, Italy. (Senior Personnel)
- **Ivica Rogina**, Univerity of Karlsruhe, Germany. (Senior Personnel)
- **Tanja Schultz**, Language Technologies Institute, Carnegie Mellon University. (co-PI)
- **Hagen Soltau**, Univerity of Karlsruhe, Germany. (Senior Personnel)
- **Loredana Taddei**, AETHRA, Ancona, Italy. (Industry Partner)
- **Alex Waibel**, Carnegie Mellon University and University of Karlsruhe. (co-PI)
- **Hannes Werthner**, ITC-irst, Trento, Italy. (Senior Personnel)

References

[1] ACT II Tongues: Audio Voice Translation Guide System. http://www.avt-actii.lmowego.com/.

[2] S. Burger, L. Besacier, P. Coletti, F. Metze, and C. Morel. The nespole! voip dialogue database. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 2001. ISCA.

[3] R. Cattoni, M. Federico, and A. Lavie. Robust analysis of spoken input combining statistical and knowledge-based information sources. In *Proceedings of ASRU-01*, December 2001. to appear.

[4] Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory Based Learner, version 3.0 Reference Guide. Technical Report Technical Report 00-01, ILK, 2000. Avaliable at http://ilk.kub.nl/ ilk/papers/ilk0001.ps.gz.

[5] E. Eide. Distinctive features for use in an automatic speech recognition system. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 2001. ISCA.

[6] C. Fügen and I. Rogina. Integrating dynamic speech modalities into context decision trees. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000. IEEE.

[7] Donna Gates, Alon Lavie, Lori Levin, Marsal Gavaldà, Monika Woszczyna, and Puming Zhan. End-to-end evaluation in janus: a speech-to-speech translation system. In M. Mast E. Maier and S. Luperfoy, editors, *Dialogue Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence (1236), pages 195–206. Springer Verlag, 1997.

[8] M. Gavaldà. Epiphenomenal Grammar Acquisition with GSG. In *Proceedings of the Workshop on Conversational Systems of the 6th Conference on Applied Natural Language Processing and the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000)*, Seattle, U.S.A, May 2000.

[9] E. M. Gold. Language Identification in the Limit. *Information and Control*, 10(5), 1967.

[10] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolke, Paul Taylor, and Carol Van Ess-Dykema. Switchboard discourse language modelling project final report. Technical Report Research Note No. 30, Johns Hopkins University, Center for Speech and Language Processing, 1997 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports, 1998.

[11] K. Kirchhoff. Integrating articulatory features into acoustic models for speech recognition. In *Proc. Phonus 5*, Saarbrücken, Germany, 2001. Universität des Saarlandes.

[12] A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei, and F. Balducci. Architecture and design considerations in nespole!: a speech translation system for e-commerce applications. In *Proc. Human Language Technology Conference*, pages 31–34, San Diego, CA, USA, March 2001. DARPA.

[13] A. Lavie, L. Levin, T. Schultz, C. Langley, B. Han, A. Tribble, D. Gates, D. Wallace, and K. Peterson. Domain portability in speech-to-speech translation. In *Proc. Human Language Technology Conference*, pages 82–86, San Diego, CA, USA, March 2001. DARPA.

[14] Lori Levin, Boris Bartlog, Ariadna Font-Llitjos, Donna Gates, Alon Lavie, Dorcas Wallace, Taro Watanabe, and Monika Woszczyna. Lessons learned from aa task-based evaluation of speech-to-speech machine translation. In *Proceedings of LREC*, Athens, Greece, 2000.

[15] Lori Levin, Donna Gates, Alon Lavie, and Alex Waibel. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, pages Vol. 4, 1155–1158, Sydney, Australia, 1998.

[16] Lori Levin, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Gavaldà, Detlef Koll, and Alex Waibel. The Janus-III Translation System. *Machine Translation*, 15(1-2):3–25, 2000.

[17] D. Litman and S. Pan. Predicting and adapting to poor speech recognition in a spoken dialogue system. In *Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI)*, Austin, TX, 2000.

[18] Marsal Gavalda. SOUP: a Parser for Real-world Spontaneous Speech. In *Sixth International Workshop on Parsing Technologies*, pages 101–110, Trento, Italy, February 2000.

[19] F. Metze, J. McDonough, and H. Soltau. Speech recognition over netmeeting connections. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, pages 2389–2392, Aalborg, Denmark, September 2001. ISCA.

[20] Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman. *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann, San Mateo, California, 1992.

[21] M. Ostendorf. Moving beyond the beads-on-a-string model of speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, USA, 1999. IEEE.

[22] S. Oviatt. The cham model of hyperarticulated adaptation during human-computer error resolution. In *Proc. International Conference on Speech and Language Processing*, Sydney, Australia, November 1998. IEEE.

[23] Yan Qu, Barbara Di Eugenio, Alon Lavie, Lori Levin, and Carolyn Rosé. *Minnimizing Cumulative Error in Discourse Context*. Springer Verlag, Berlin, Heidelberg, New York, 1997.

[24] Yan Qu, Carolyn Rosé, and Barbara Di Eugenio. Using discourse predictions for ambiguity resolution. In *Proceedings of COLING*, 1996.

[25] Norbert Reithinger and Elizabeth Maier. Using statistical dialogue act processing in verbmobil. In *Proceedings of ACL*, 1995.

[26] C. Rosé, B. Di Eugenio, L. Levin, and C. Van Ess-Dykema. Discourse processing of dialogues with multiple threads. In *Proceedings of ACL*, 1995.

[27] Carolyn Rosé and Lori Levin. An Interactive Domain Independent Approach to Robust Dialogue Interpretation. In *Proceedings of COLING/ACL*, 1998.

[28] H. Soltau and A. Waibel. On the influence of hyperarticulated speech on recognition performance. In *Proc. International Conference on Speech and Language Processing*, Sydney, Australia, November 1998. IEEE.

[29] B. Suhm, B. Myers, and A. Waibel. Model-based and empirical evaluation of multimodal interactive error correction. In *Proc. CHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA, USA, May 1999. ACM SIG.

[30] Kavita Thomas. Designing a Task-Based Evaluation Methodology for a Spoken Machine Translation System. In *Proceedings of ACL-99 (Student Session)*, College Park, MD, 1999.

[31] T.Takiguchi, S. Nakamura, and K.Shikano. HMM-Separation-based Speech Recognition for a Distant Moving Speaker. *IEEE Transactions on Speech and Audio Processing*, 9(2):127–140, 2001.

[32] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. DARPA communicator dialog travel planning systems: The june 2000 data collection. In *Proceedings of Eurospeech*, 2001.

[33] M. Walker, D. Litman, C. Kamm, and A. Abella. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 1998.

[34] Marilyn Walker, D. Litman, C. Kamm, and A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL'97)*, 1997.

[35] Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[36] Wayne Ward. The CMU Air Travel Information Service: Understanding Spontaneous Speech. In *Proceedings of the DARPA Speech and Language Workshop*, 1990.

[37] M. Westphal and A. Waibel. Model-combination-based acoustic mapping. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001. IEEE.

[38] Monika Woszczyna, Matthew Broadhead, Donna Gates, Marsal Gavaldà, Alon Lavie, Lori Levin, and Alex Waibel. A Modular Approach to Spoken Language Translation for Large Domains. In *Proceedings of Conference of the Association for Machine Translation in the Americas (AMTA'98)*, Langhorn, PA, October 1998.