Nespole![1]
Negotiating through Spoken Language in E-commerce

NSF-99-102

1 Introduction

Recent technological advances in computing and networking infrastructure have opened up the opportunity of new and exciting applications for the broad general public. Among these, Electronic Commerce (E-commerce) has emerged as the application with the greatest utility for everyday general usage, resulting in enormous growth over the last few years, with expectations of orders of magnitude further growth in the upcoming decade. E-commerce opens the possibility of selling goods and services on a global market supported by the World-Wide Web. Up to now, however, interactions have mainly been menu based, guiding customers through a predesigned choice space allowing a limited set of predefined results. The next generation of E-commerce will need to overcome this limitation. Similar to today's "traditional" commerce, E-commerce should develop into a collaborative activity involving information presentation, negotiation, recommendation and problem solving, in which the customer is allowed to explore alternative solutions, dynamically adapting his/her goals, while the E-service provider reacts by eliciting unexpressed needs and configuring viable alternatives.

In order to enable E-commerce with human negotiation, the underlying technology must be capable of dealing with human-to-human communication. Furthermore, as the reach of electronic markets naturally transcends national boundaries, linguistic boundaries must also be removed. Transnational multilingual human-to-human communication must then be provided to enable flexible and effective E-commerce world-wide. E-commerce service providers can partially solve the problem by using teams of human operators covering a variety of relevant languages. However, such solutions are expensive and may even be infeasible in the context of specific tasks such as product maintenance, trouble shooting and repair, where the relevant operator must possess both technical and language competence. Resorting to a lingua franca, such as English, is only a poor substitute and disenfranchises the providers from their markets and customers.

Nespole! aims at providing a system capable of supporting advanced needs in E-commerce and E-service by fully integrating automatic speech-to-speech translation (henceforth, STST) into the underlying technology. The project builds on years of research experience on STST by the project partners both within and outside of C-STAR [2], as well as on research conducted by various other national and international projects (such as Verbmobil). However, Nespole! will not only address accuracy of translation, but will extend the investigation to the ability of humans to negotiate and jointly solve problems. Enabling effective cross-lingual spoken human-to-human communication in the context of multimodal interaction in a joint workspace is the overarching challenge of this project. It will feature multiple synergistic translation programs (using a multi-engine approach) that range from concept driven analysis and interlingua based translation modules, to simple dictionary and phrase-book lookup to statistical direct translation schemes. Nespole! will also provide for non-verbal communication by way of multimedia presentations, shared collaborative spaces and multimodal interaction and manipulation of the objects in these joint E-commerce spaces.

[1] *Nespole* (`NES-po-lay`) means literally *loquat fruit* in Italian. As an exclamation, it means *Wow!*.

[2] Consortium for Speech Translation Advanced Research `http://www.c-star.org`

The STST and multimedia technologies developed in the course of the project will be demonstrated by means of two showcases. The first, at the end of the first fourteen months of the project, will support multilingual negotiation and discussion between a tourist information/service provider (a so-called *destination*) and a customer. The goal of the interaction is to plan a trip, by exploring a wide range of available options for various sub-tasks, including travel, accommodation, attractions and recreation, cultural events, dining and so on.[3] In a second showcase, the tourist scenario will be greatly expanded. In addition, a second, completely different domain will be addressed, consisting of a video help-desk for technical support, trouble shooting and repair, also capable of acting as a training environment for local retailers and technicians of an international company. Four languages will be considered: English, German, French and Italian.

2 Scientific Challenges and Innovations

NESPOLE! intends to improve over current experimental speech-to-speech translation systems, addressing a number of scientific and technological issues: robustness, scalability, cross-domain portability, and multimodal interaction with multimedia content. The specific innovations that will be contributed by CMU include: two new robust parsing methods; a robust, scalable interchange format for multilingual meaning representation in multiple domains; and the integration of the speech translation modules in a multimodal shared workspace.

Robustness: In order to fully support natural interaction between the partners, the system must be able to cope with corrupted inputs, due to either the peculiarities of the input utterance or to errors of the acoustic recognizer, and with incomplete information. Robustness in NESPOLE! will be achieved with an array of methods including two new, innovative parsers — the Concept Classification parser (CC) and the LCFLEX parser (Section 8) — and multi-engine integration of knowledge-based and corpus-based machine translation methods.

Scalability and Cross-Domain Portability: In present STST systems, robustness is achieved at the cost of breadth. Thus, STST systems are usually confined to narrow semantic domains. Although attempts at addressing scalability and cross-domain portability have already been made in many areas of language technologies, these are rather new concerns in STST. They are crucial, though, to offer STST as a viable mode of communication. The two NESPOLE! showcase systems will demonstrate broadening STST in two ways. The first showcase will demonstrate scalability through the expansion of the travel domain of our previous work [29, 20]. The second showcase will demonstrate cross-domain portability through a new application in a video help-desk domain.

The key to scalability in NESPOLE! is the distinction between *task-oriented* and *descriptive* sentences. (See Section 7.) Most STST systems (possibly with the exception of [8, 12]) are limited to task-oriented sentences. However, even task-oriented dialogues include some descriptive sentences. Not handling these makes for a dry and not very detailed dialogue. We argue that the robust methods that have been effective for task-oriented sentences are not effective for translation of descriptive sentences, and propose methods for translation of descriptive sentences in spoken

[3] Tourism is one of the economical sectors with more growth chances, and among those to take the greatest advantage of the availability of telecommunication supports such as the Web, video-conference, etc. Also, tourism is involved in an important change consisting in a shift from the current broker-supported market (agencies, tour operators) to a situation in which the customer directly contacts and negotiates with a local representative of service providers (the destinations). Such a change is motivated by, and at the same time induces a parallel shift from ways to organize the offer around a fixed number of solutions, towards scenarios in which the keywords are flexibility and negotiation.

dialogue. Innovations in this area include a renovation of our interlingua design to integrate representations of task-oriented and descriptive sentences, and a lexically-driven option in the LCFlex parser for robust parsing of descriptive sentences.

Cross-Domain portability in Nespole! will be addressed by the development of new translation approaches that are inherently more portable than existing technology and by the integration of direct translation engines that, while shallow in coverage, are far more domain-independent than the proposed IF-based approaches. The new concept classification parser that we propose relies on an underlying phrase-level semantic grammar that is far less domain dependent (thus far more portable) than complete semantic grammars. The concept classification itself will be based on data-trainable technology (such as HMMs and neural nets), that are inherently more suitable for portability to new domains. Our other new translation approach will focus on the translation of descriptive sentences, and will be designed to be domain independent as much as possible, thus supporting portability to new domains.

Multimodality: The integration of STST in an environment supporting multimodality and multimedia presentation in large and complex domains goes well beyond the capabilities of current STST systems and state-of-the-art technologies (telephone-based call-centers, or by Web-based e-commerce/service). The main challenge Nespole! will face is the proper treatment of synchronization between language and visual information: already a hard task, now complicated by STST. The fact that the linguistic message goes through a translation stage means that straightforward methods for synchronizing pointing acts and natural language referring expression based on measurements performed on an underlying time-line are not viable. Synchronization is a symbolic process, which can only be captured as obtaining between pointing acts, as performed by humans on images on the screen, and semantic units, as encoded in IF expressions. Therefore, Nespole! will provide: methods for detecting the synchronization between acoustic events and pointing acts, methods for encoding the synchronization at the interlingua (symbolic) level, and methods for re-building the synchronization into the target multimodal scenario.

3 Appropriateness for the MIAM Program

The Nespole! project deals with multilingual and multimodal, human-human and human-machine communication in an electronic commerce application. Thus it is of central importance in "accessing, managing, and exchanging information (about goods and services) in a multilingual and multicultural context." Nespole!'s machine translation capability is an enabling technology "required for their (multilingual information systems) application in a number of social and organizational contexts," specifically, electronic commerce. We believe that a successful Nespole! will have significant societal impact in both the US and Europe, as it will provide systems that lead to more effective, efficient, trade and exchange between the countries and across their languages/cultures.

Nespole! further contributes to the research infrastructure agenda of defining "coding and interchange standards for multilingual spoken and written language data." The Nespole! Interchange Format (IF) is a general, extendible, and portable representation of task-oriented utterances, which has proven to be useful and reliable in a multiligual application. It extends, refines and generalizes a very successful IF already developed by six countries under C-STAR for travel planning and is thus of broad interest to the international language processing community. (See Section 7.) The proposed work will continue development of IF and extend it to improve domain portability or independence.

The NESPOLE! consortium has a long track record of working together under the consortium for speech translation research, C-STAR, and has already practiced a strong and lively US-European collaboration for many years. Moreover, the Interactive Systems Labs have operated both out of University of Karlsruhe (Germany) and at Carnegie Mellon University (US) since 1991 thereby also establishing a long track record and a unique management example for successful US-European scientific cooperation.

4 The Consortium Members and Management Structure

NESPOLE! will be pursued by a consortium of four scientific partners and two user/technology provider partners. The scientific partners are (1) Istituto Trentino di Cultura, Istituto per la Ricerca Scientifica e Tecnologica (ITC-irst) in Trento, Italy, (2) CLIPS-IMAG, in Grenoble, France, (3) The Interactive Systems Lab at the University of Karlsruhe, Germany, and (4) The Interactive Systems Lab and Language Technologies Institute at Carnegie Mellon University, USA. The scientific partners are amongst the major players in the speech and natural language processing community. All these institutions have also been cooperating since 1994 as partners of an international consortium for spoken language translation: C-STAR II. For this reason, they have a long and well-established experience of joint work and of successful cooperation.

The user and technology provider partners are Azienda per la Promozione Turistica di Trento (APT) and AETHRA. APT is a provider of tourism services. AETHRA is a medium-size enterprise involved in the production and commercialization of video-conference equipment, and an e-commerce/service supplier with users in South America, Asia, Europe, and North America.

In addition to these partners, which form the NESPOLE! consortium, a user-group will be set up from the beginning of the project to act consult on strategic choices, exploitation, external validation, and dissemination.

The role of the various partners is as follows: ITC-irst, UKA, CMU-LTI and CLIPS will provide the human language technologies for Italian, German, English and French, and the technologies for multimedia and multimodality. APT will provide requirements of a typical tourist destination, and will help with validation. AETHRA will provide both the video-call center technology and the user expertise of a help-desk managing organization on a global level.

Project management corresponds to Workpackage 1 in the European NESPOLE! proposal. The management structure will consist of (1) two Project Directors, one from Europe and one from the US, (2) a Project Management Committee with one member from each partner, (3) a Project Technical Committee with one member from each partner, (4) a team of Workpackage Leaders reporting to the Technical Committee, and a Project Exploitation Manager for technology transfer.

5 Technical Specifications and Domain Specifications

Devising specifications and requirements for the domains, scenario, translation modules, and their integration into the showcases corresponds to Workpackage 2 in the European NESPOLE! proposal. It will be lead by CLIPS. This workpackage involves:

Domain Specifications: delimiting the extent of coverage and the structure of the two domains - Travel Planning and Help Desk.

Scenarios Requirements: developing specifications for how the clients are expected to interact with the e-service providers including the types of multimedia presentations and multimodal interactions that are required

Translation Modules and Resources Specification: specifications for the IF representations, the input and output specifications for the translation modules in all languages, the required linguistic resources that are to be developed, and specifications for data and corpora collection and annotation.

Architectural Requirements: requirements and specifications for the adopted hardware, for the communication network and for the software needed for integrating the different functionalities

Testing and Evaluation Specification: a detailed plan for the various testing and evaluation activities, including objectives, thresholds and procedures. Evaluation specifications will be designed for both single modules (such as the analysis module for any particular language), as well as comprehensive specifications for the STST system as a whole, the evaluation of the showcases and the evaluation of the effectiveness of multimedia and multimodal interactions.

6 Showcase Development

Showcase development corresponds to Workpackage 3, lead by AETHRA, in the European proposal. There will be two showcases — a tourism showcase in Month 14 and a help desk showcase in Month 30. Both showcases will involve a customer and a service provider negotiating via multimodal communication. The following tasks are involved in showcase development:

Set-up, testing and validation of the hardware and software platform supports: The communication network will be based on Internet or Intranet. All the data related to human-to-human communication, speech, pointing gestures and video conference will be transmitted using standard protocols. Speech and gestures will be encoded using de facto standards that are now emerging, mainly based on the XML family, for example, SMIL or similar future developments.

NESPOLE! will explore the suitability of a general system architecture based on terminals for the users (PCs) connected via IP to a server. The terminal will have the tasks of: inputting speech, images and gestures; encoding and sending them to the server; and outputting the same information when coming from the server. The server will provide all the relevant STST services, the multimedia data related to the application domain, and the multimodality functionalities.

Integration of the STST, multimedia and multimodal functionalities: Our scenarios integrate visual cues and images with multilingual speech. The two parties will interact with the system by acting on a “whiteboard”, though, for reasons explained in Section 9.2, this will not be configured as a real shared space.

7 The Design of the Interchange Format (IF)

Although NESPOLE! will include multiple translation engines, the showcase systems will focus on translation via an interlingua. Analysis modules map source language utterances onto the interlingua representation, and generation modules map interlingua representations onto target language utterances. The interlingua represents the meaning of the source language utterance independently from how it is expressed in the grammar of the source language [23]. Interlingua systems are efficient for multi-lingual situations because one interlingua representation can be input to several target language generators simultaneously, thus avoiding the need for a different set of transfer rules between each source-target pair.

The Interchange Format corresponds to Workpackage 4 (Intermediate Representation Format) in our partners’ European proposal. Carnegie Mellon will be the leader of this workpackage, with participation from IRST, CLIPS, and UKA on interlingua design and participation from APT and AETHRA on domain-specific content.

History of the Interchange Format: The NESPOLE! interlingua is called Interchange Format (IF), and originated in the C-STAR Consortium in 1997. It has been used since then in translation of task-oriented dialogues in the travel domain in six languages (English, French, German, Italian, Japanese, and Korean), thus demonstrating usefulness in different language families. There is a database of 4423 utterances, sampled from five languages, labelled with IF representations. Intercoder agreement for assigning IF representations has not been checked, but translation success among different C-STAR partners shows that the various C-STAR partners use it consistently with each other. IF can not only serve NESPOLE!'s translation needs, but can be a resource to other projects working with multilingual data.

Balancing robustness and expressive power in task-oriented and descriptive sentences: The interlingua is the medium through which analysis and generation modules communicate; therefore it determines the set of all and only the meanings which are to be made available through translation. It should therefore be rich in order to support expressivity. However, if the interlingua is too complex, we risk failure in not being able to produce it reliably. Robustness is often pursued by giving up full expressivity (as it was in our previous work [19]), with the consequence that parts of the information content of the input are kept out of the reach of the interlingua. Building on previous experience, we intend to provide for an interlingua which optimally balances such conflicting requirements, at the same time effectively contributing to the goals of scalability and cross-domain portability.

In balancing the conflicting requirements on the interlingua, it is important to distinguish between task-oriented utterances and descriptive utterances. Our previous work dealt only with task-oriented utterances — those which involve speech acts such as suggestions (*How about a double room*), requests (*Could you make that reservation for me*), acceptances (*A double room will be fine*), etc. Translation of a task-oriented utterances focuses mainly on the speech act and not very much so on literal translation of the words. In fact, task-oriented sentences often cannot be translated literally. Descriptive utterances, on the other hand, convey mainly semantic content and are less distinguished from each other on the basis of speech act. Examples of descriptive utterances in the NESPOLE! domains are *The castle was built in the thirteenth century* or *When I turned it on, nothing happened.*

Description of IF: In the treatment of task-oriented sentences, IF uses the concept of *domain actions* such as requesting information about the availability of a hotel or giving information about the price of a tour. Each domain action consists of a speech act such as `request-information` or `give-information`, optionally some concepts such as `availability` or `price`, and optionally some arguments with values such as `double room` or `one hundred dollars per night`. The IF specification document includes 44 speech acts and 93 concepts, and 117 argument names, although not all combinations of these are used. The IF database of 4423 sentences contains 554 distinct combinations of speech acts and concepts. However, the current English analysis and generation grammars are designed to recognize and generate many more (around 3000) domain actions. The following examples illustrate some typical uses of IF.

no that's not necessary
`negate`
and i was wondering what you have in the way of rooms available during that time
`request-information+availability+room`
and how will you be paying for this
`request-information+payment (method=question)`

however the rate for the double is one thirty four ninety five per night

```
a:give-information+price+room (room-type=double, price=(quantity=134.95, per-unit=night))
```

i+m now arriving june sixth

```
c:give-information+temporal+arrival (who=I, time=(june, md6))
```

Strengths of the IF representation: *Multilinguality:* IF is suitable for multilingual situations because it retains nothing from the source language grammar, but represents only the speaker's domain action. This type of interchange format is desirable for task-oriented sentences whose surface expression varies greatly among languages.

Suitability for many analysis methods: Another advantage of IF is that it is well suited for knowledge-based parsing ([29]) as robust analysis methods which attempt to extract domain actions without a detailed analysis of each sentence [22, 13].

Scalability and Portability: The speech act portion of an IF representation is domain-independent, as are many of the arguments such as `time` and `quantity`. Extending IF for a new domain involves identifying domain-specific concepts and arguments, which can be done efficiently with corpus-based methods.

Plans for extending IF: In order to accommodate descriptive utterances in Nespole!, we will add an interlingual representation of propositional content ([6]) using an ontology of concepts (e.g., [15]). We plan to address the following research issues:

Treatment of translation mismatches: Speech-act-based interlinguas explicitly address the problem of translation mismatches — instances where a literal translation does not work — which arise frequently in task-oriented sentences. However, translation mismatches also arise in descriptive sentences and have been widely studied (e.g., [5]). It is our feeling that many mismatches in descriptive sentences can be attributed to the way that closed class meanings are expressed (e.g., tense, aspect, certainty, ability, obligation, and evidentiality) [18]. Our work will address the representation of closed class meanings in order to assure high quality translation in the case of translation mismatches.

Representation of topic and focus: Another point that must be addressed concerns information-packaging, in particular the representation of *topic* (given information) and a *focus* (new information) (e.g., [4]). Capturing such information in a grammar-independent representation is crucial because languages have different grammatical means to realize topic-focus. Even in the task-oriented representation of domain actions, the topic/focus distinction is necessary to distinguish *The Hilton Hotel is in Verona* where *The Hilton Hotel* is topic from *The hotel in Verona is the Hilton Hotel* where *The Hilton Hotel* is focus.

Representation of multimodal information in IF: Finally, in the Nespole! project IF plays the important role of providing the symbolic level encoding of multimodal links between linguistic referential items and their pictorial/visual referents.

8 Spoken Language Translation Research

The core of the Nespole! project will be the development of new state-of-the-art speech translation methods, which will be integrated into a complete STST system. While we plan to build upon the technology and scientific advances achieved by previous major STST efforts such as C-STAR [26] and Verbmobil, our main research goal will be to significantly advance the technological capabilities of STST systems in terms of robustness, scalability and portability to new domains.

The partners in this consortium have significant experience in building STST systems using a variety of rule-based, corpus-based, and statistical approaches to translation. For Nespole!, our

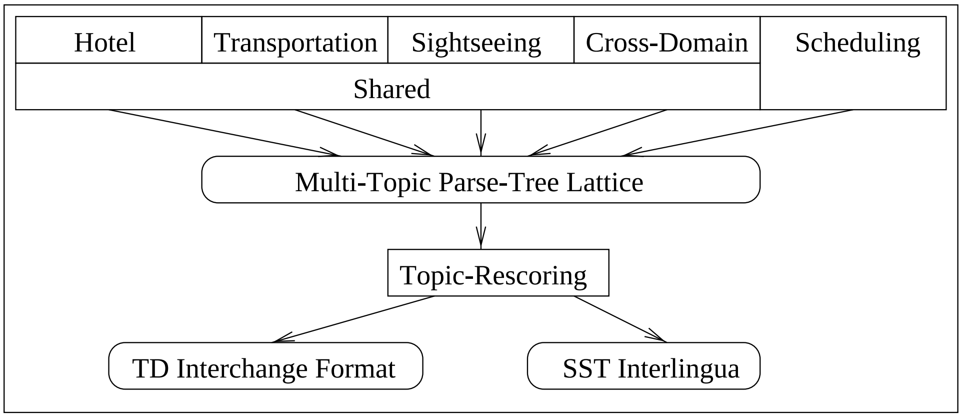

Figure 1: Combining multiple sub-domain grammars with shared and cross domain grammars.

intention is to allow partners to investigate a variety of approaches to STST, some rule-based, some corpus-based, and some a combination of both. The IF will be the main point of interface between partners. Each partner will develop its own analysis module to map sentence in its language onto IF, and its own synthesis module to map IFs onto its language. All-ways translation between all pairs of languages is achieved by combining any analysis module with any synthesis module, mediated by the IF. This mode of collaboration between partners has been effective in C-STAR.

The work described here corresponds to Workpackage 5 in the European proposal. This workpackage will be lead by UKA with help from all the other research partners (CMU, IRST and CLIPS). We focus here primarily on the planned research on development of new spoken language translation components at CMU.

The research work at CMU will focus on the development of a number of different translation approaches that will be combined into a multi-engine system, to make the best of the strengths, and minimize the weaknesses of each single approach. Within this multi-engine context, we will experiment with approaches that use the IF as well as approaches that do not. The approaches that do not use IF will be developed for English and German only, since they have to be specifically trained and developed for each pair of languages separately. Our goal in trying different approaches is to provide comparative evidence concerning the respective merits and weakness of IF- versus non-IF-based approaches with respect to robustness. The individual approaches and our proposed research into multi-engine combination are described in greater detail in the following subsections

8.1 IF-based Analysis Engines

We intend to experiment with three IF-based approaches (symbolic parsing with semantic grammars, concept classification parsing with semantic grammars, and shallow syntactic parsing), comparing and assessing their respective merits and weaknesses.

Semantic Grammar-based Approach: The semantic grammar-based approach [27] has been the focus of our previous STST research within the context of the C-STAR project, and has proven to be effective for large yet limited domains such as travel planning, which can be broken down into several natural sub-domains. For both analysis and generation we have been using semantic grammars. Rather than focusing on the syntactic structure of the input, semantic grammars list ways of expressing semantic concepts. For example, the concept of something being available can be expressed with the phrases *we have* ... or *there are* .... Because they focus on identifying a set

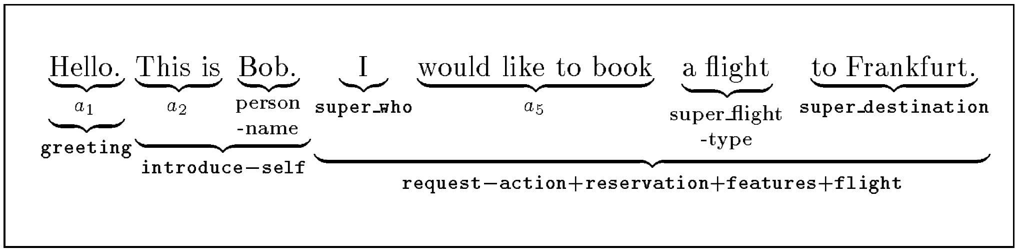

Figure 2: Example: Multi-level Analysis for an Input Utterance

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 our current version of the Janus system we have developed a way to combine modular grammars in order to overcome the problems associated with expanding semantic grammars to new domains. The parser [14] applies multiple sub-grammars in parallel and stores the outputs in 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 [29]. Figure 1 illustrates this modular approach.

For the Nespole! project, we will incorporate our grammar-based approach as one of the multiple analysis engines. We expect to leverage coverage off of the domain-independent portions of our current grammars (labelled as "cross-domain" and "shared" in Figure 1). We will also limit the amount of grammar writing that needs to be done by using our new Concept Classification approach to parsing (described below). With a Concept Classification parser we only need to write mini-grammars for IF arguments; the amount of grammar writing for speech acts and concepts in the IF is significantly reduced.

The Concept Classification Approach: We propose to construct a new Concept Classification parser, CC, for analyzing task-oriented speech utterances. The goal of the parser is to analyze utterances directly into our dialogue-act (DA) based IF representation (Section 7). Complete DA representations consist of a speech-act, a domain concept and a list of arguments. Cc will operate 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, the parser identifies the speech-act and domain concept based on the sequence of detected arguments and the words in the utterance. 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 will be done using the robust SOUP parser [14]. The phrase-level semantic grammars will be developed manually, but we expect to leverage coverage off of portions of our current 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 availabilty of adequate accurate training data.

We have already conducted a pilot study [22] of the above approach. The pilot study was performed on our current C-STAR travel planning domain, with the goal of proving the feasibility of the general approach. In the pilot study, we used a multi-level HMM [24] to model argument-level concepts and speech-acts. The models were trained on a labelled corpus of utterances parsed by our full semantic grammar system. The trained HMM was then used to determine the segmentation and classification of utterances into speech-acts and argument-level segments. The argument-level segments were then parsed (when possible) using the portion of the semantic grammar that corresponds to each of the argument labels. A single neural net was trained to map the sequence of argument labels to a domain concept. In the above example, the sequence of arguments `[super-who a5 super-flight-type super-destination]` should be mapped to the domain concept `reservation+features+flight`. The speech-act, domain concept and arguments were then combined together to form a complete IF dialogue act.

A preliminary evaluation of this pilot system showed promising results. We compared the performance of the pilot system with that of our full semantic grammar analysis system, via an English-to-English "paraphrase" test. Using a test set of 200 English utterances, we analyzed the input into IF representations using both systems. The resulting IFs were then generated back to English (using our grammar-based generation component). The paraphrases were then graded manually by 2 human graders, and categorized as "perfect" (if the paraphrase fluently conveyed the complete content of the input), "OK" (if the content was conveyed, but minor details were missing, or with some disfluency), or "bad" (if the paraphrase was considered unacceptable). "Perfect" and "OK" scores are also summed together as "acceptable". The full semantic grammar component achieved a score of 73.7% acceptable paraphrases, the pilot CC analyzer achieved 57.3% acceptable paraphrases. However, the two analysis systems have significantly different strengths. An "oracle" experiment that picked the better of the two paraphrases results in substantially better performance than each of the individual analyzers.

All aspects of the pilot concept classification analyzer require significant further research and development. The underlying argument grammars will need to be further developed, to support better detection and disambiguation of correct argument-level segments. The second stage detection and classification of speech-acts and domain concepts must also be significantly improved. We will investigate a variety of classification approaches beyond those already tested in the pilot study.

Shallow Syntax-based Analysis: We also plan on developing a shallow syntax-based analysis component for descriptive sentences. Descriptive sentences will require an IF representation that is based on predicate-argument structure rather than on domain actions (Section 7). Parsing will be performed using the **LCFlex** parser [25], a recently developed robust parser for unification-augmented context-free grammars. LCFLEX is a left corner chart parser, that can support a variety of flexibilities: skipping over portions of the input, limited word and category insertions, and limited relaxation of unification constraints. Additionally, LCFLEX supports multi-stage parsing that allows the parser to apply different types and levels of flexibility at different stages. The parser's flexibility parameters can be fine tuned for optimizing performance in specific applications. Pilot experiments have been run on speech translation in the appointment scheduling domain. We plan on making necessary adaptations to the parser based on our experiences from applying it in the NESPOLE! project.

To analyze descriptive sentences into IF representations we plan on using LCFlex to simulate a chunk parser [1, 30, 7] that is lexically driven. Small constituents will be identified by syntactic rules. Then long sentences will be segmented into clauses, and a complete structure for each clause will be assembled according to what arguments are licensed by the lexical head of the clause.

8.2 Generation

In Nespole!, the task of generation from IF will be more complex than in previous STST systems. Because our planned system will explicitly be designed to handle multimodal information, our generation components must be able to produce both sensible output utterances, and synchronization markers between pointings and referring expressions. To this end, the output of generation engines will be a representation in some XML-like language - e.g., SMIL - for synchronized multimedia. Such a language must be capable of encoding both the language part, pointers descriptors, and temporal links between the referring expressions of the output sentence and the pointer. For generation of the linguistic strings, we will draw on research addressing such questions as discourse level phenomena (e.g., the cognitive state of referents) and information-packaging (topic/focus).

8.3 Direct (non-IF-based) Translation Approaches

We will experiment with a number of direct (non-IF) translation approaches and will integrate them, along with the IF-based engines, into a multi-engine system. These will include (1) **Example-Based Translation**, using a generalized EBMT engine capable of producing good results from a relatively small bilingual corpus. It produces direct translation between the languages [3]; and (2) **Glossary/Dictionary-based Direct Translation**, which tends to be of lower quality than other techniques, but is simple enough that it is possible to achieve very large vocabulary coverage, providing a fall-back position when the other engines fail to produce a translation. This has proven useful in the Pangloss [11, 7] and DIPLOMAT [8, 12] projects.

These direct approaches depend on the availability of bilingual data, and we propose to initially limit experimentation therewith to the translation between English and German. The direct translation approaches will most likely not be fully integrated into the showcase systems, given that the absence of the IF would make it more difficult to account for multimodality in a principled way. We believe, nevertheless, that extending the multi-engine approach to such techniques will provide information concerning robustness, scalability and domain portability which will prove valuable for future projects. Our intention is to investigate the relative contribution of the direct translation approaches to each of the above questions, as compared with the IF-based approaches.

8.4 Multi-engine Translation

A significant amount of research effort in Nespole! will be given to combining the individual translation approaches in a multi-engine system. Such a paradigm combines several different translation modules into a single architecture, with the goals of capitalizing on their respective strengths, and minimizing their weaknesses, in terms of robustness, scalability, domain portability and translation quality.

As developed in the Pangloss and DIPLOMAT projects, the Multi-Engine Machine Translation (MEMT) architecture [10] (Figure 3) feeds an input text to several MT engines in parallel, with each engine employing a different MT technology[4]. Each engine attempts to translate the entire input text, segmenting each sentence in whatever manner is most appropriate for its technology, and putting the resulting translated output segments into a shared chart [16, 17, 28] data structure after assigning each segment a score indicating the engine's internal assessment of the

[4] Morphological analysis, part-of-speech tagging, and possibly other text enhancements can be shared by the engines, if available.

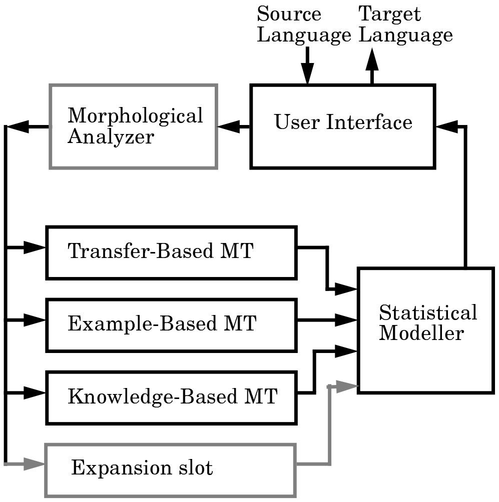

Figure 3: **Structure of MEMT architecture**

quality of the output segment. These output (*target language*) segments are indexed in the chart based on the positions of the corresponding input (*source language*) segments. Thus the chart contains multiple, possibly overlapping, alternative translations. We then use techniques based on statistical language modelling (described in [2]) to choose the best overall set of outputs.

The MEMT architecture makes it possible to exploit the differences between different MT technologies. Differences in translation quality, robustness, scalability, and domain portability can all be exploited simply by merging the best results from different engines. When IF-based techniques can produce high-quality, in-domain translations, their results will tend to be used; when Example-based finds a high-quality match, those results will tend to be used. When none of the other engines produces a high-quality result, the Glossary/Dictionary engine supplies a lower-quality translation, which is still much better than leaving the source language untranslated[5]. The design of the MEMT architecture also allows easy integration of new translation engines (such as direct statistical translation), which fall beyond the scope of the work proposed under the current project.

The main new research challenges in this area are: (1) how best to extend the multi-engine framework to incorporate multiple IF-based and non-IF-based translation approaches; and (2) the development of a significantly improved statistical framework for selecting an optimal combination result from the multiple engines.

8.5 Information Extraction

We plan to exploit IE techniques to enhance scalability and cross-domain portability, taking advantage of the considerable amount of textual data describing places, cultural and sport events, etc., or parts of an equipment, etc., which will be made available by our users. *Named entity Recognition* techniques will provide STST systems with the relevant information about proper names, city and place names, company names, etc. which are needed for both the analysis and

[5] As implemented in Pangloss, unchosen alternative translations can be selected by the user through a special-purpose interface [9], which greatly increases the usefulness of the transfer-based translations.

the synthesis chain modules to work properly. *Template Filling* techniques will provide structured description (e.g., including the agent, objects, date, location, and so on) of relevant events.

8.6 Corpora Acquisition and Annotation

Each Nespole! partner will collect, transcribe, and annotate (with IF representations) data for its home language. This data will be used to develop and test the language knowledge sources of the translation components (grammars, lexica, IF, etc.), to conduct evaluations, and to train the statistical and corpus-based components for speech recognition and translation.

We have many years of experience in data collection, transcription, and annotation for C-STAR and Verbmobil. Data collection scenarios usually involve face-to-face role-playing of traveller and agent. In the later stages of C-STAR we have begun to collect data through user studies. The traveller and agent are not face-to-face, but communicate only through our MT system and interface. Nespole! will begin development with this user study data. However, user studies can only be carried out when the system has reasonable coverage of the desired domain. Therefore, each time we expand or change domains we will have to bootstrap with a small amount of face-to-face data.

9 Speech and Multimodal Components

Multimodality corresponds to Workpackage 6 in the European proposal. This workpackage will be lead by IRST with help from UKA and participation of CMU and CLIPS.

9.1 Speech Recognition Components

The partners of Nespole! already have the infrastructure, the research teams and many years of experience in building large vocabulary speech recognition systems in the languages of the project. Since the proposed effort will focus on enabling STST technologies in practical and commercial situations, we will focus on the language/translation capabilities. For speech recognition components we will therefore leverage as much as possible from on-going speech research efforts at the partners' laboratories and limit the speech activity under Nespole! to integration and adaptation of our state-of-the art systems. This activity will include expanding and adjusting pronunciation dictionaries and language models to the domains at hand, and adaptations of the recognition engines to the microphones used, and the channel requirements of the interconnected systems. We will also have an opportunity to evaluate multi-engine combinations of speech systems. Mono-lingually this technique (also known as "Rover") was shown by NIST to improve results as it combines mutual strengths of different recognizers. At CMU we will be able to evaluate a combination of acoustics produced by the JANUS Speech Recognition Toolkit, ACID (a Neural Net based LV-CSR system), SPHINX, and various derivatives from all of the above. Multilingual integration could also be possible but will be outside of the scope of this proposal.

9.2 Multimodal Components

It is an important innovative goal of the Nespole! project to achieve effective cross-lingual communication in the context of multimedia shared electronic market places. Cross-lingual speech and language technology will therefore be integrated and harmonized with different media, images, sound, text, Web-pages, so as to provide the participants the most effective tools for communication. Rapid progress is expected as all the scientific partners have developed multimodal and multimedia systems, that the consortium can build on.

With respect to STST, multimodality is still a largely unexplored issue, so care will be taken to study the problem both from the point of view of the impact on the users of the system, and from the point of view of the impact on the system itself and its modules. Drawing on the results of these studies, we will design and implement solutions for a selected number of relevant features, which will then be integrated in the final showcase for demonstration. Our purpose is not to solve all problems involved, but to identify viable communication environments and avenues for further research.

We can single out three main problems: the first concerns the impact of multimodality on the system and its users: reaction by the users, conditions of use, types of features and facilities which improve the communication. The second and third concern a major characteristic of multimodality – namely, pointing gestures. Pointing is the act whereby the user exploits images or portion thereof for communicative purposes, using them as meaningful objects. The communicative goal is obtained by associating pointing with appropriate (referential) linguistic expressions. We will need to investigate how pointing is used in our scenario, how it integrates with speech and language, and how domain and language (in)dependent it is. Moreover, we will need to tackle the hard problem of synchronization. To exemplify, suppose the customer utters something like "This/the cable is broken" and accompanies the referring expression with a pointing gesture indicating the relevant cable in a given image on the screen – e.g., by circling the cable, crossing it, by a simple click, etc. On the other party's side, the system must be able to reproduce the pointing gesture at the appropriate point within (the utterance of) the translated sentence. Translation introduces not only a time delay but also linguistic asymmetries, since the linear relationships found in the source sentence may not be reproduced by the target one. We plan on dealing with the synchronizing pointings and their referential expressions at the IF level, where referential expressions will be paired with information relating to their corresponding pointing gesture.

Detection of pointing gestures: Depending on the scenario there are different types of gestures describing technical problems or indicating relevant parts of a document, of a picture, etc. Such pointing gestures can be used to mark regions of interest in a map, to point to devices, or to select a paragraph in a document. Notice that we need not address the problem of 'pointing recognition', namely that of assigning meaning to (possibly different) pointing acts. Meaning is assigned to pointings by the humans involved in the conversation. We need only address the simpler problem of detecting particular events in visual space -namely, pointings. Standard pointing tracking devices will then be used to this end; they will provide for 'pointing descriptors'.

Synchronizing audio and visual information: For simplicity, this task can be broken down into three steps, each producing a pair (a link) consisting of the pointing descriptor and an object, X. At the end of the first step, X is an acoustic event. At the end of Step 2 the second element of the link is the IF object corresponding to the acoustic event. Finally, Step 3 substitutes the IF object with the referring expression in the target language produced by the generation module. The final link is then interpreted by the (target) user interface, in such a way that the correct synchronization between translated sentence and reproduced pointing can be established.

Design and implementation of crosslingual and multimodal workspaces: We will explore methods and design issues to support a rich multimedia environment. One goal is having the system dynamically propose relevant visual and/or textual information, according to the evolution of the conversation, the current topic, salient elements, etc. This will be obtained by enabling the system to keep track of the current topic of interest, of the relevant objects mentioned, etc., and by making it capable of foreseeing informational needs.

10 Testing and Evaluation

Testing and evaluation will focus on robustness, scalability and portability, as well as impact on users for showcases and multimodality. Tests will be conducted on unseen data collected as described in Section 8.6. Testing and Evaluation corresponds to Workpackage 7 in the European proposal. Carnegie Mellon will be the leader of this workpackage.

Task-Based Evaluation: This involves identifying the users' goals in a dialogue and giving the dialogue a score which is a function of whether each goal succeeded or failed and how many turns were spent on repairing wrong translations. (The method for deciding what counts as a goal and the exact nature of the scoring function are topics for research [21].) CMU, with the help of APT and AETHRA, will conduct several task-based evaluations in Years 2 and 3.

Sentence-Based Evaluation of Coverage and Accuracy: For each sentence, the source and target are compared by a bilingual human judge. We evaluate the translation components alone (using hand-transcribed input), and also evaluate end-to-end performance using speech-recognition output as the input to the translation components. CMU will conduct sentence-based evaluations monthly on unseen data starting in Year 1. Results will be plotted over time.

Evaluation of individual components: Analyzers and generators can be evaluated individually using hand-coded perfect output as a gold-standard. Speech recognition can be evaluated using standard measures such as word error rate. Multimodality can be evaluated by comparing the system performances against testbeds obtained from the data collected during the experimental phase described in Section 9.2. Individual components will be evaluated on an as-needed basis (e.g., for error analysis) by CMU, IRST, CLIPS, and UKA.

Evaluating scalability and portability: To evaluate scalability and portability, we will plot coverage and accuracy as a function of the amount of resources. For knowledge- and IF-based STST systems, training data, grammars, lexica, and time spent by human developers will be plotted monthly along with the sentence-based evaluations. For Example-Based STST, glossary-based STST, and statistical methods, experiments plotting the amounts of training data, glossaries, etc., against coverage and accuracy will be conducted in Year 2.

Evaluation of Multi-Engine STST: We will compare the performance (probably sentence-based, not task-based) of the multi-engine STST system to the performance of individual engines. We will also remove individual engines from the multi-engine system to see how performance degrades. This evaluation will take place at CMU in Year 2.

Showcases evaluation: Showcases will be evaluated with respect to standard measures of usability in the area of human-computer interaction. Given the demonstration nature of the applications, subjective measurements will be preferred to objective ones. From the beginning of the project, focus groups will be set up and standard techniques of "discount usability engineering" will be applied to ensure that showcases meet user expectations and needs. IRST, with the help of APT and AETHRA, will be responsible for showcase evaluations.

11 Dissemination

Dissemination corresponds to Workpackage 8 in the European proposal, lead by CLIPS. This includes creation and maintenance of a project web page; publication in national and international conferences (e.g., Eurospeech, ICSLP, IEEE, COLING, ACL); publication in scientific journals (e.g., SpeechCom, Computational Linguistics, Machine Translation, IEEE, ACM); publication in large diffusion journals (e.g., Science et Vie, La Recherche, Scientific American, Science); organization of events such as a final workshop; participation in demonstrations (e.g., ErgoIA, CHI, Interact); and establishment of a users group.