# Speech-to-Speech Translation Services for the Olympic Games 2008

Sebastian Stüker[1], Chengqing Zong[4], Jürgen Reichert[1], Wenjie Cao[4], Muntsin Kolss[1], Guodong Xie[4], Kay Peterson[2], Peng Ding[4], Victoria Arranz[3], Jian Yu[4], and Alex Waibel[1,2]

[1] interACT, Universität Karlsruhe (TH), D-76131 Karlsruhe, Germany
stueker@ira.uka.de, juergen@ira.uka.de, kolss@ira.uka.de, waibel@ira.uka.de
[2] interACT, Carnegie Mellon University, Pittsburgh, PA 15213, USA
kay.peterson@cs.cmu.edu, waibel@cs.cmu.edu
[3] ELRA/ELDA, 75013 Paris, France
arranz@elda.org
[4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
cqzong@nlpr.ia.ac.cn, wjcao@nlpr.ia.ac.cn, gdxie@nlpr.ia.ac.cn, pding@nlpr.ia.ac.cn

**Abstract.** In 2008 the Olympics Games will be held in Beijing. For this purpose the city government of Beijing has launched the *Special Programme for Construction of Digital Olympics*. One of the objectives of the program is the use of artificial intelligence technology to overcome language barriers during the games. In order to demonstrate the contribution that speech-to-speech translation technology (SST) can make to solving this problem and in order to prove the feasibility of deploying such technology in the environment of the Olympic Games 2008 in Beijing, we have developed the *Digital Olympics Speech-to-Speech Translation System* that addresses a general touristic domain with a special focus on pre-arrival hotel reservation. The system allows for rapid development of SST prototypes, the study of different user-interfaces and the on-the-fly comparison of alternative approaches to the individual problems involved in this task.

## 1 Introduction

In today's world traveling the globe has become increasingly possible for a growing number of people in the world. With the advent of affordable and fast intercontinental transportation, mostly by means of air travel, and through increasingly more transparent national borders, the number of international tourists rises steadily.

As a tourist in a foreign country one has to satisfy certain basic needs, such as shelter, food, and transportation. But when being on vacation one does not only want to fulfill these basic requirements. As a tourist one wants to interact

with the people of the visited country and experience their culture. Key to this experience is the ability to communicate with the natives of the country that one visits. However, learning the language of every country one wants to travel to is clearly infeasible. In some places of the world, e.g. in Europe, English has been established as lingua franca. But even here, as in many places of the world, English is not always spoken, especially among people with little international contact. Therefore, English as a mean of communication with the native population of an arbitrary country is often not an option. Also, language is a key component of culture. Thus unifying the languages of the world into one common language comes with a loss in cultural diversity which we want to avoid. Here modern speech and language processing technologies can be the savior for keeping the language diversity of a globalized world.

While speech-to-speech translation for arbitrary unconstrained domains is only starting to become the topic of research activities, translation systems for limited, pre-defined domains have been developed that have reached a grade of maturity that makes them ready for field deployment in the near future.

In 2008 the Olympic Games will be held in Beijing, the capital of the People's Republic of China. It is expected that many visitors from all over the world will take this opportunity to visit China. Only very few of them will be able to speak Chinese but will want to seek deeper understanding of the Chinese culture and to come in contact with the local population. In order to make the Olympic games an attractive and enjoyable event, the city government of Beijing has launched the *Special Programme for Construction of Digital Olympics*. In the spirit of the Olympic idea one of the objectives of the program is to remove language barriers with the aid of artificial intelligence technology in order to promote friendship and mutual understanding. To achieve this goal, speech-to-speech translation technology can make a substantial contribution. In order to prove the feasibility of the deployment of speech-to-speech translation technology in the environment of the Olympic Games in 2008, we have produced the *Digital Olympics Speech-to-speech Translation System* prototype for a tourist application. The development of the prototype was a joint, international effort between *CapInfo*, the *National Laboratory of Pattern Recognition* (NLPR) at the Chinese Academy of Sciences and the *International Center for Advanced Communication Technologies* (interACT) at both Universität Karlsruhe (TH) and Carnegie Mellon University. The resulting system was successfully demonstrated at the *Beijing International High-Tech Expo* (ChiTec) 2004 in Beijing and during the *Language Technology Days* at FORUM 2004 in Barcelona.

By providing a flexible, modularized platform it is able to demonstrate the different aspects of the technologies involved in speech-to-speech translation: automatic speech recognition, machine translation, and speech synthesis. By allowing to run different modules in parallel it is possible to compare the strengths and weaknesses of different approaches to the individual problems on the fly under real-world conditions.

## 2 System Overview

The *Digital Olympics Speech-to-Speech Translation System* in its current form is able to translate spontaneously spoken speech between arbitrary pairs taken from the languages Chinese, English, and Spanish. It covers a tourist expressions domain, with a strong focus on pre-arrival hotel reservation. The system integrates automatic speech recognition components, machine translation engines and speech synthesis modules and can either work as a stand-alone solution with a user interface running on a laptop or can be extended by two PDA clients that are connected via wireless network to the laptop. The two PDAs then work as the interface to the system while the translation and recognition engines are running on the laptop in a client-server setup. The system can exchange any of the six components necessary for the translation service (two speech recognizers, two translation components, and two speech synthesis modules) on the fly. In that way it is possible, for example, to compare the translations for the same sentence as given by two different components on the fly.

### 2.1 Domain

The Digital Olympics speech-to-speech translation system demonstrates the usability of speech translation technology for the Olympic Games 2008. In order to do so, it addresses three different domains: *pre-arrival hotel reservation, basic travel expressions, basic medical needs.*

The main focus of the system is the capability to translate in the domain of pre-arrival hotel reservation. Hotel reservation is a domain in which our labs already had some experience before the beginning of the development of this system. Therefore, it was possible to develop a prototype for this scenario in a very short time. At the same time the hotel reservation scenario is close enough to the actual scenario of a foreign tourist making inquiries at a hotel desk or at a shop, that a proof of feasibility for the pre-arrival hotel reservation scenario implies the feasibility for the other scenarios.

The domain of basic travel expressions demonstrates the capabilities of speech translation technology to act as a helper in predictable, re-occurring scenarios in which tourists need to communicate in situations typical to travelers. Currently in those situations tourists often make use of phrase books. However PDA based translation systems can solve this issue more elegantly by providing greater flexibility in finding the correct phrases and by providing a more convenient and more natural interface. They do so by taking natural, spontaneous speech as input, giving speech output of the translation and by showing more sophisticated functionality than a phrase book. In this way communication flows smoother than it would when utilizing an old fashioned phrasebook. This is especially true in situations where the tourist will not be able to speak the phrase book entries, as it is often the case for Chinese. As foundation for this domain we took the *Basic Traveler's Expression Corpus* (BTEC) [1].

The domain of basic medical needs can be seen as a specialized case of the travel expressions domain. Its application provides significant leverage to the

speech-to-speech translation technology due to the importance of this domain. Knowing beforehand to be able to overcome language difficulties in medical emergency situations will encourage people to take on a visit to a foreign country.


**Fig. 1.** PDA user interface

For integrating the different, often very heterogeneous, components we developed the *Active Speech* framework.

The *Active Speech* framework is an environment for building and testing multi-modal interfaces. The framework allows for the easy creation of demos, prototypes, and analysis of interface issues. The basic idea of the framework is the data flow paradigm. Each component transforms received data and sends the transformed output to one, none or several receivers. The receivers themselves then again transform and resend the data. For example, speech recognizers transform audio to text, translator components transform text from one language to another, synthesis components transform text to audio etc. Special components can reconfigure the links between the components depending on what data they receive. In order to allow for distributed solutions or client-server setups, components can be located on different computers communicating over the network. Besides Windows desktops, portable devices using Windows CE are also supported.

In the *Active Speech* framework there are two kinds of components: *service components* which provide a service to other components and *visual components* which interact with the user. Visual components can be placed on the screen by drag and drop and can be connected to other components by simple mouse clicks. A system setup running in the framework can be configured and changed

during run-time, in order to allow for short development cycles and interactively observe the behavior of different components in the translation task.

### 2.2 User Interface

The system itself has two different user interfaces, one for devices with large displays, such as laptops, and one for devices with smaller, lower resolution displays, such as PDAs. Both displays provide basically the same functionality to the user. For recording purposes it shows a waveform representation of the recorded signal. This is actually a good and intuitive display for the user to indicate the quality of the audio recording. The interface further shows the recognized sentence in the original language and the translation into the target language. In between is the phrasebook that automatically picks sentences that are close to the recognized one. In addition to the PDA user interface, the interface of the laptop allows for control of the loaded components, the selection of the components to be used in the translation process, and the manual repetition of translations and synthesis, e.g. after changing a component or correcting an error, e.g. via key-board. Figure 2.1 show the user interface for the PDA.

## 3 Speech Recognition Component

For the speech recognition components we developed large vocabulary spontaneous speech recognition systems in the three languages Chinese, English, and Spanish. The Chinese speech recognition component was developed by NLPR, the Spanish and English systems by Universität Karlsruhe (TH). Different linguistic resources for fitting the recognizers to the new domain were available, among them BTEC and data we collected at our labs, such as prototype dialogs for hotel reservation, in domain word lists, and protocols of previous demos.

### 3.1 Chinese Recognition System

The NLPR Chinese recognition system extracts features from overlapping 24ms long frames of audio data at a rate of 100 frames per second. From these frames 12 Mel-warped cepstral coefficients are extracted. Together with the normalized energy plus 1 dimension normalized pitch, and their first and second derivatives, they compose the final 42-dimensional feature vector.

Decision tree based gender dependent class triphones based on 62 phones (including silence as a separated phone) are trained on 800 hours of speech. Each state is modeled by 16 Gaussian components. To incorporate tone information, the question set is specially designed to take the Mandarin tonal information into consideration. The obtained acoustic model typically includes tonal information for both, the right and left context phones, and the base phone itself, and can provide detailed information for a tonal language, such as Mandarin.

The main language model used by the system is a word based N-Gram model using Katz backoff, and is estimated for a 50K word set vocabulary using hundreds of million of words of training texts. We have also investigated the use of

class-based language models using automatically derived word classes to smooth the word model probabilities in the system. The final LM used in the system is an interpolation of a 4-Gram and a class based 3-Gram model. Our LVCSR system uses a time-synchronous decoder that can either operate in a single pass or can be used to generate or rescore word lattices. The decoder can operate with cross-word triphone models and direct incorporation of trigram language model. Afterwards the output lattice can be rescored with more advanced language models. To fit the proposed tonal AM, the search engine was updated to consider the tone information during path propagation, merge and pruning which make our decoder quinphone like.

### 3.2 English Recognition System

The English and Spanish recognition systems were developed with the help of the Janus Recognition Toolkit JRTk featuring the IBIS one-pass decoder[2]. The English recognition system is derived from the *ISL Meeting Transcription System*[3]. For the translation system the Meeting System was reclustered into a fully-continuous system with 6000 models, each being a Gaussian Mixture models with 32 Gaussians. The system uses an MFCC based front-end with per utterance Cepstral Mean Subtraction and Cepstral Variance Normalization. The vocabulary of the resulting system contains 2500 entries. As language model we use a 3-Gram language model with Kneser-Ney backoff trained on 2.1M word corpus with the help of the *SRI Language Model Toolkit*. In the language model we make use of manually defined semantic classes, such as hotel names, person names, local points of interest, etc. In that way it is possible to port the recognizer to new scenarios, e.g. a new city, without the need of retraining the language model.

During decoding we employ incremental vocal tract length normalization and feature space constrained MLLR on a sliding window in order to be able to adapt to changing speakers. We evaluated the system on a collection of 13 dialogs in which hotel reservation scenarios were reenacted in a spontaneous manner. The scenarios were collected with the help of 17 native speakers of American English at Carnegie Mellon University. The database contains approximately 23 minutes of speech in 319 turns. On this set the recognizer achieved a word error rate of 18.1%. On a laptop with a Pentium M 1.7GHz the system runs in realtime.

### 3.3 Spanish Recognition System

The Spanish system was also trained with the help of the JRTk and derived from a Global Phone speech recognition system for South American Spanish [4]. The phoneme set was modified to better fit Castilian Spanish and then trained on ca. 14h of Castilian speech. The acoustic model consists of 2000 triphones with roughly 51k Gaussians. The recognition system has a vocabulary of 24k. A 3gram language model with the same semantic classes as for the English system was trained on roughly 255k words. The training corpus includes the Spanish portion of the BTEC and several in-house, in domain text collections.

For decoding the IBIS decoder was also used, and the same incremental adaptation scheme as for the English recognizer was applied. The sytem was tested on hotel reservation dialogs collected at Universität Karlsruhe (TH) with the help of 18 native speakers. The database contains 2973 turns, giving about 170 minutes of speech. The system yielded a word error rate of 16.3% on these dialogs. On a laptop with a Pentium M 1.7GHz the system runs in about realtime.

## 4 Speech Translation

For the *Digital Olympics SST System* we incorporated two different approaches to machine translation. One rule-based approach based on an interlingua described in section 5 and one statistical machine translation approach described in detail in section 6. Also, in addition to the machine translation approaches we have incorporated a phrase book constructed out of the BTEC.

### 4.1 Phrasebook

The phrase book contains 162320 phrases taken out of the BTEC, which were categorized into 13 classes by NLPR, and a dictionary with about 100000 translation pairs. In order to restrict the result set to a query, the user can select categories to filter the output phrases. The phrase book can be used in multiple modes. It can be used like a normal electronic dictionary or the user can search for phrases with the help of key words. The similarity mode returns related phrases to a given phrase, with a maximum word distance. This often allows to correct errors from the speech recognition, if a similar or the same phrase to the spoken one is in the database.

### 4.2 Translation modes

Our speech-to-speech translation system can run in two different modes. In the first mode one can pre-select one of the provided translation components as the default for translating the speech input, e.g. the IF based translation component. The second mode is a cascade of different stages that are triggered one after the other as necessary. In the first stage a look-up in the internal phrasebook is made, to see whether a translation of the input sentence is stored there. If so, the stored translation is taken. If not the input sentence is forwarded to the IF based translation component. If the IF translation component however fails to parse the input sentence, e.g. because it is out-of-domain, the input is forwarded to the last fall-back stage, the statistical machine translation component which is guaranteed to produce a translation.

## 5 Interlingua Based Translation

The interlingua-based engine of the MT system utilizes the Interchange Format (IF) that was originally developed for CSTAR II and later on expanded and

modified for the LingWear [5] and NESPOLE! [6] systems. This interlingua was designed to cover spontaneous task-oriented dialogs in specific, limited domains [7]. It captures speaker intention rather than literal meaning, abstracting away from syntax and language-specific idiosyncrasies. Utterances to be represented in IF are segmented into semantic dialog units (SDUs).

The assumption underlying such a domain-action based interlingua for MT purposes is that utterances relevant to a particular domain can be classified into a limited number of domain actions [8]. For each language, analysis components from input language to interlingua and generation components from interlingua to output are written to cover those domain actions reliably. Consequently, an MT engine based on this interlingua can be expected to perform best on translating utterances highly relevant to the domain it was designed for, and exhibit more limited or no translation capabilities for out-of-domain utterances. The original domain covered for CSTAR II was pre-arrival hotel reservation; for NESPOLE!, this was expanded to cover more general tourism-related inquiries and later on medical patient-doctor conversations.

### 5.1 Interlingua-based analysis and generation

Interlingua-based MT allows for the use of different analysis and generation components for different languages that can all be integrated into the same MT engine. For the system at hand, the English and Spanish analysis and generation use the same parsing and generation mechanisms, while Chinese uses different ones and is therefore described in a separate section.

For both the analysis and generation side, the English grammars for the Digital Olympics system were expanded and adapted from pre-existing grammars. The expansion and adaptation was also done at Carnegie Mellon. Some of the structure of the English grammars was taken as a seed for the Spanish grammars, which were otherwise written from scratch specifically for the FAME [9] and Digital Olympics projects. The Spanish grammars were developed at Universitat Politècnica de Catalunya. The data used for grammar development were taken from the IF-annotated C-STAR II database, which was translated into Spanish.

### 5.2 English and Spanish interlingua-based analysis

The SOUP parser [10], a stochastic, top-down, chart-based parser specifically developed for parsing spontaneous speech in real time, is used for English and Spanish analysis into IF. In a pre-processing step, the speech recognition output is standardized to optimize parsing performance; for example, filler words and hesitations may be removed before parsing. The SOUP parser calls context-free grammar rules from the grammar for the respective language that contain top-level domain action and lower level rules. SOUP segments utterances into SDUs at parse time. The SOUP output is then mapped into standard IF format in a subsequent step.

### 5.3 English and Spanish interlingua-based generation

The GenKit generator [11], a powerful pseudo-unification-based generation tool, generates English and Spanish output from IF. GenKit uses language-specific hybrid syntactic/semantic grammars in combination with generation lexica to generate natural language text output from an IF-based feature structure. Highly domain-relevant, frequent domain actions may be generated via rules specific to them to guarantee reliable and highly fluent output. On the other hand, more general rules serve as fall-back rules to enable generation from as wide a variety of domain actions as possible. GenKit grammars use syntactic, lexical, and morphological knowledge. Generation of the correct morphological form is achieved via inflectional grammar rules that draw on additional information stored in the lexical entries. In the more complex case of verb morphology in Spanish, the correct form is retrieved from an additional morphological form look-up table. Also specific to Spanish, several linguistic phenomena were tackled that had not been implemented in GenKit-based grammars before. At the end, the GenKit output is sent through a post-processing stage to ensure that clean text is produced.

### 5.4 Evaluation of English-Spanish IF based translation

In [12] the English-Spanish IF based translation was evaluated by human judgment of the fidelity and naturalness of the translations on ten hotel reservation dialogs with an English speaking client and a Spanish speaking agent. When dealing with perfect transcriptions, instead of error prone automatic transcriptions obtained from the described speech recognizers, 91.7% of the translations from Spanish to English were good in content, only 1.2% were considered bad, the rest O.K. From English to Spanish 82.4% were good in content, only 1.6% bad, the rest O.K. When using speech recognition output as input to the translation, for Spanish to English still 96.4% of the translations were at least O.K.; but only 62.4% for English to Spanish, showing the higher complexity of that direction for speech recognition as well as for translation.

### 5.5 Parsing and Generation of Chinese Sentences

The parsing of the Chinese sentences into IF uses a robust hybrid parsing scheme [13] that combines Hidden Markov Model (HMM) based approaches with rule based ones for extracting semantic information and mapping parsed result into IF. The parsing takes place in three distinct steps: *chunk identification, chunk and HMM based analysis, chunk interpretation.*

A chunk is defined under the restrictions of semantic level and syntactic level, which is a head-word cored semantic unit that has relatively independent phrase structure and relatively complete semantic composition. In our paraphrase system, we classify all words in the domain of travel information into 324 semantic types. Similar to the syntactic parsing with context free grammar (CFG) and Chart parsing algorithm, we have developed a grammar for semantic chunk

recognition, in which all rules are described by semantic type markers. From the resulting chunk sequence the HMM based analysis picks up the skeleton of the IF. Hereby The chunk sequence is interpreted as the observation of a HMM while the IF is corresponds to the internal states of the HMM. The parameters of the HMM were trained on a large set of tagged spoken Chinese sentences. The IF skeleton for an input chunk sequence is then found by calculating the most likely HMM state sequence with the help of the Viterbi algorithm. Finally our chunk interpreter fills the slots of the IF skeleton by using the internal structure and semantic information of the chunks acquired during the chunk identification according to the corresponding parsing rule in the rule base.

For the Chinese generation part, we implemented a feature-based generation approach in our spoken Chinese paraphraser for generating the Chinese translation from the generated IF [14].

The spoken Chinese generator consists of two functional modules, the micro-planner and the surface generator. The micro-planner generates the functional structure of the resulting sentence.

The surface generation is then the final stage of the sentence generation. The final sentence is generated based on the micro-planning results and a system functional grammar for the generation language. Our surface generator employs a top-down and depth-first unification algorithm. After that, in the sentence linearization, the order of components in the generating sentence is arranged according to the unification order of the sentence and phrase. In a post-processing step modifier words expressing the tense and voice are being added.

## 6 Statistical Machine Translation

Statistical machine translation is based on the noisy channel approach. The ISL system uses as primary building blocks phrase-to-phrase translations extracted from bilingual data by optimizing a constrained word-to-word alignment for an entire sentence pair [15]. Phrase translation candidates are scored by a modified IBM1 alignment model. For words inside the source phrase the summation of the lexicon probabilities is restricted to the probabilities for words inside the target phrase candidate, and for words outside of the source phrase it is restricted to the probabilities for the words outside. The alignment probabilities from both alignment directions are interpolated. Single source words are treated as phrases of length 1. Most phrase pairs are seen only a few times, even in very large corpora. Therefore, probabilities based on occurrence counts have little discriminative power. Phrase translation probabilities are calculated based on a statistical lexicon, i.e. on the word translation probabilities. The language model used in the decoder is a standard 3-gram language model. We use the SRI language model toolkit to build language models of different sizes, using the target side of the bilingual data only or using additional monolingual data. The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations are used to generate a translation lattice. Second, a first-best search

is performed on this lattice, using the language model probabilities in addition to the translation model probabilities to find the overall best translation.

For training data we compiled a trilingual corpus of spoken utterances covering the target domains of general tourism, hotel reservation, medical assistance, and specific tourist assistance for the Olympic Games in Beijing. The main components of this corpus were the BTEC, as well as in-house collected dialogs for medical assistance, and hotel reservation dialogs collected at Carnegie Mellon University and translated from English into Chinese and Spanish. In addition, a manual list of about 4600 named entities specific for the city of Beijing was added, consisting of the names of bus and metro stations, tourist attractions and sites, hotel names, and person names. Overall, the corpus currently has a size of about 190K utterance and named entity tuples. As a preprocessing step, the Chinese part of the corpus was segmented into words using a segmenter derived from the LDC segmenter. The final system also uses a small number of translation rules for number and date expressions.

## 7 Speech Synthesis

Concatenative speech synthesis technology has been employed in the Chinese TTS system. The system contains three main parts: text analysis model, prosody generation model, and unit selection model. The text analysis applies some preprocessing such as text normalization, parsing, and text-to-pinyin conversion. In the prosody generation model, special attention has been paid to tones during the prosody information prediction process. Depending on the prosody information and segmental information, the synthesizer selects units from a real waveform corpus, and smoothens the pitch contour before outputting the final speech.

The English and Spanish speech synthesis were provided by Cepstral LLC, Pittsburgh, PA, USA. Cepstral provides state-of-the-art unit selection text-to-speech synthesis and voices that are small and fast enough to run on handheld devices or distribute over the network. Cepstral voices support SSML, VoiceXML tags, and Microsoft(R) SAPI.

## 8 Conclusion

In this paper we have introduced our *Digital Olympics Speech-to-Speech Translation System*. It incorporates modules from our different labs for the technologies necessary for SST: Speech recognition, machine translation, and speech synthesis. Its successful implementation and demonstration on various occasions gives proof to the feasibility of deploying speech-to-speech translation technology in the environment of the Olympics Games 2008 in Beijing. Automatic speech-to-speech translation for tourist domains, such as general tourist needs, hotel inquiries or medical needs can provide significant leverage for the distribution of this technology due to the practical significance of the domains in real-life. The introduced Active Speech framework allows for fast development of prototypes,

studies of user interfaces, and on-the-fly comparison of different approaches to the technologies necessary for translation systems of this kind.

## 9 Acknowledgements

The authors would like to thank Raquel Tato and Marta Tolos for their help in the development of the Spanish recognition system, Dorcas Alexander for her contribution to the development of the IF components, and Victoria MacLaren for her help in the data collection. Special thanks go also to Elisabet Comelles for her help in the development of the IF components and the carrying out of the corresponding evaluation.

## References

1. Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S.: Creating corpora for speech-to-speech translation. In: EUROSPEECH. (2003)
2. Soltau, H., Metze, F., Fügen, C., Waibel, A.: A one pass-decoder based on polymorphic linguistic context assignment. In: ASRU. (2001)
3. Metze, F., Jin, Q., F ügen, C., Laskowski, K., Pan, Y., Schultz, T.: Issues in meeting transcription - the ISL meeting transcription system. In: ICSLP. (2004)
4. Schultz, T., Waibel, A.: Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication **35** (2001)
5. Fügen, C., Westphal, M., Schneider, M., Schultz, T., Waibel, A.: LingWear: A mobile tourist information system. In: HLT. (2000)
6. Metze, F., McDonough, J., Soltau, H., Langley, C., Lavie, A., Levin, L., Schultz, T., Waible, A., Cattoni, R., Lazzari, G., Mana, N., Pianesi, F., Pianta, E.: The NESPOLE! speech-to-speech translation system. In: HLT. (2002)
7. Levin, L., Gates, D., Lavie, A., Waibel, A.: An interlingua based on domain actions for machine translation of task-oriented dialogues. In: ICSLP. (1998)
8. Levin, L., Langley, C., Lavie, A., Gates, D., Wallace, D., Peterson, K.: Domain specific speech acts for spoken language translation. In: 4th SIGdial Workshop on Discourse and Dialogue. (2004)
9. : (http://isl.ira.uka.de/fame/)
10. Gavalda, M.: Soup: A parser for real-world spontaneous speech. In: 6th IWPT. (2000)
11. Tomita, M., Nyberg, E.: Generation kit and transformation kit, version 3.2, user's manual. In: Technical Report CMU-CMT-88-MEMO, Carnegie Mellon University, Pittsburgh, PA, USA (1988)
12. Arranz, V., Comelles, E., Farwell, D., Nadeu, C., Padrell, J., Febrer, A., Alexander, D., Peterson, K.: A speech-to-speech translation system for catalan, spanish and english. In: AMTA. (2004)
13. Xie, G., Zong, C., Xu, B.: Chinese spoken language analyzing based on combination of statistical and rule method. In: ICSLP. (2002)
14. Cao, W.: Approach to target language generation in spoken language translation (in Chinese). Master's thesis, Institute of Automation, Chinese Academy of Sciences (2004)
15. Vogel, S., Hewavitharana, S., Kolss, M., Waibel, A.: The ISL statistical machine translation system for spoken language translation. In: IWSLT. (2004)