# Multilinguality in Speech and Spoken Language Systems

A. Waibel, P. Geutner, L. Mayfield Tomokiyo, T. Schultz, M. Woszczyna

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA
University of Karlsruhe
Karlsruhe, Germany

## 1. Introduction

With the appearance of low-cost commercial large vocabulary dictation software for personal computers, speech recognition has truly come of age. Spoken language applications are transferred ever more rapidly into practical use and are beginning to affect our everyday lives. With these successes it is all too natural that there is a growing interest in expanding the reach of speech and language systems to international markets and bringing these technologies to consumers worldwide.

Unfortunately, such expansion is still associated with considerable effort. Modern speech and language systems increasingly employ automatic learning algorithms. While these algorithms reduce the painstaking development work, they do require large data resources such as texts, voice recordings, pronunciation lexicons, morphological decomposition information, and parsers. At present, adequate language resources have only been accumulated for a relatively small number of languages, particularly the more dominant languages of the world.

In the following sections, we will first discuss differences between languages and the resulting challenges for speech recognition. We introduce approaches to efficiently deal with the enormous task of even covering a small percentage of the word's languages by building speech recognition systems for multiple languages through model combination, bootstrapping, and adaptation techniques. Foreign accents, which present their very own challenge to speech recognition engines for all languages, are covered in section 3. Then, we will present an approach that allows speech recognition with virtually unlimited dictionary sizes, which is important for languages that are highly inflected or allow the generation of long compound nouns. The challenges of multilingual speech translation will be reviewed in the final section and conclude our overview of *Multilinguality in Speech and Spoken Language systems*.

# 2. Language Differences

In this section, we will highlight some of the differences between languages and the resulting challenges for speech recognition. Language differences that affect meaning, interpretation, and reference and the problems they present for speech understanding applications will be addressed in a section about *Multilingual Speech Translation*.

## 2.1 Scripts and Fonts

Many different character types are used in the world's languages (see figure 1). Writing systems fall into two major categories: ideographic and the phonologic. In ideographic scripts, the characters reflect the meaning rather than the pronunciation of a word. Examples for ideographic scripts are the Chinese Hanzi and the Japanese Kanji. Phonological scripts can be further divided into syllable-based scripts, like Japanese Kana or Korean Hangul, and alphabetic scripts which are used for most Indo-European languages, such as Greek script for Greek, or Latin script for English and German. In syllable-based scripts each grapheme reflects one syllable, whereas in alphabetic scripts graphemes correspond roughly to one phoneme.

Phonologic scripts are often easier to handle than ideographic scripts in the speech recognition framework, as in many cases rule-based grapheme-to-phoneme tools can be used to generate the pronunciation dictionary needed to guide recognition, while this is usually not possible for ideographic scripts. However, among the languages using alphabetic scripts, the grapheme-to-phoneme relationship varies considerably. It ranges from a nearly one-to-one relationship such as for Slavic languages like Russian and Serbian as well as some Romance languages like Spanish, up to languages like English and Gaelic that require complex rules and have many exceptions. Furthermore, in some languages the written script reflects only a part of the spoken phonemes. In Arabic, for example, only consonants are written out; vowels have to be filled in by scanning the context and understanding the meaning.


**Figure 1: Scripts for languages: Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, English, Greek, Hebrew, Hindi, Italien, Japanese, Korean, Rumanian, Russian, Serbian, Thai**

Usually, alphabetic scripts do not have more than 30 different characters. In this case, 8-bit character codes are sufficient to store all characters of this script. For ideographic scripts, however, 16-bit codes are required, since thousands of unique characters occur in written text, as in Chinese Hanzi and Japanese Kanji. For these languages multi-byte characters have to be used. Another issue, especially in case of multilingual text processing, is the direction in which text is written. Languages like Arabic are written from the right to the left, Indo-European languages are written from left to right, and for some languages the preferred direction is top to bottom.

## 2.2 Segmentation

English has a natural segmentation into words that can conveniently be used as dictionary units for speech recognition. The words are long enough to differ from each other in a sufficient number of phonemes, but short enough to be able to cover most material with a reasonable number of different word forms that occur frequently. This is important for the statistical analysis required by the automatic learning processes that modern speech recognition systems rely on. But other languages lack an adequate segmentation. In Japanese and Chinese, whole sentences are written in strings of characters without any spacing. In order to determine appropriate dictionary units, the transcribed speech data has to be segmented manually or by morphological analysis programs. Another group of languages, including Turkish and Korean, does have some segmentation within a sentence, but their morphology provides for agglutination and suffixing. The inflection, derivation, and other relationships between words in a sentence are expressed by concatenating multiple suffixes to the word stem. This results in rapid growth of the number of word-forms occurring in a given text. As a consequence, poor recognition results are achieved when using a certain set of word-forms as dictionary entries for speech recognition, and many new word forms are encountered in unseen speech material. The following example illustrates the morphological structure of the Turkish language (hyphens have been added to mark morphology. see [OGB94]).

| |
|---|
| Turkish: Osman-l•-laç-t•r-ama-yabil-ecek-ler-imiz-den-mi•-siniz<br><br>English: behaving as if you were of those whom we might consider not converting into Ottoman. |

## 2.3 Morphology

For agglutinative and highly inflected languages, splitting up the words into several morphemes provides a first solution to curb the rapid vocabulary growth and reduce the *Out-of-Vocabulary rate* (OOV) for speech recognition. This, however, reduces the effective reach of the language model, which can partly be counteracted by using higher-order n-gram language models. In general, a lower OOV-rate does not always reflect a better recognition performance, since the smaller units also suffer from a higher acoustic confusability.

Another possibility to reduce the OOV-rate is to allow a virtually unlimited recognition dictionary. This technique will be discussed in detail in section 4.

## 2.4 Prosodic Structure

Across the world's languages, the prosodic structure of words varies considerably [Cut97]. More than half of all languages belong to the group of tonal languages, in which the lexical items are distinguished by contrasts in pitch contour or pitch level on a single syllable. Simple tonal systems have only two different classes (high versus low pitch level), others have four our five tones like Mandarin Chinese and Thai, or even more, like Cantonese, with 6 tones.

The tone variations can affect either the semantics of a word, or its grammatical function. Mandarin is an example of a language in which tone changes the meaning. In the East-African language Twi, for example, tone variations are used to signal variations in tense (grammar).

In pitch accent languages like Japanese, pitch contrasts are drawn not between syllables but between polysyllabic words. In stress languages individual syllables in a polysyllabic word are stressed. In fixed stress languages the stress pattern always occur in the same position within a word, like in the Czech and the Finnish languages, where the first syllable is always stressed, or in Turkish, where it is the last syllable within a word. Fixed stress languages are easier to model than lexical stress languages like English and German, where the stress position varies across words.

# 3. Multilingual Speech Recognition

Multilingual speech recognition is required for tasks that use several languages in one speech recognition application. A very basic approach is to integrate several monolingual recognizers with a front end for language identification. Since storage requirements put a limit on this approach, we propose to combine individual recognizers into one multilingual engine, which can handle several different languages at a time. This concept requires a combined acoustic model that represents the sounds of all the languages involved. In this section we present several approaches that we developed in the framework of the multilingual speech recognition project ***GlobalPhone***.

## 3.1 Portability

A number of recognition systems developed initially for one language have been successfully extended to several languages, including systems developed by IBM [CDG97], Dragon [BCG96], Philips [DAK95], LIMSI [LAG95], CMU [OAM92], Karlsruhe [SW98], MIT [GFG95] and many more. The extension of English systems to such varied languages as German, French, Italian, Spanish, Dutch, Greek, and Mandarin Chinese illustrates that speech technology generalizes across languages, provided large transcribed speech databases are available. Results show that similar modeling assumptions hold for most languages, but there are some exceptions due to language differences highlighted above.

In general, however, the assumption that large speech databases are available for a given language does not hold for several reasons. About 400 of the world’s languages are spoken by at least 100,000 native speakers. Which of these languages are of interest for speech recognition applications can change very quickly with the political and economic situation. Since the collection of large databases requires a significant amount of time and considerable resources, it is difficult to provide databases on demand and impractical and wasteful to preemptively try to collect them for all languages. Finally, for combinatorial reasons, it is not possible to collect enough large databases to solve the problem of non-native speech recognition. As a result, our research must focus on the most effective and parsimonious ways to adapt existing recognition engines to new tasks and new languages with only very limited data.

## 3.2 Speech Recognition in multiple languages

The multilingual speech recognition project ***GlobalPhone*** at the Interactive Systems Labs [SW97, SW98a-c, SW99] investigates large vocabulary continuous speech recognition (LVCSR) systems in many languages. Data used for our investigations currently consists of read speech data for the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, Czech, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Put together with the English WSJ and French BREF databases, this covers 9 of the 12 most widespread languages of the world. In each of the languages about 15 hours of read speech was collected, spoken by 100 native speakers per language [SW98c].

Based on this data, we developed monolingual LVCSR systems in ten languages using the Janus Recognition Toolkit described below. For each language, the acoustic model of the baseline recognizer consists of a fully continuous 3-state HMM system with 3000 triphone models. A mixture of 32 Gaussians models each HMM-state. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power, and zero crossing rates. After cepstral mean subtraction a linear discriminate analysis reduces the input vector to 32 dimensions.

In table 1 we present the word error rates (ER), vocabulary sizes (vocab) and trigram perplexities (PP) for the ten monolingual recognizers. Though the core engines are the same across languages, differences in the recognition performance are not only due to inherent language-specific difficulties. They strongly depend on differences in quality and quantity of the data used, and on the expertise of the language experts. Moreover, the concept of a word is difficult to define for some languages (Chinese, Japanese, and Korean) as discussed above, making the comparison of word error rates awkward. In our opinion it is therefore misleading to infer language difficulties from the given word error rates.

In order to give a more reliable measure of the acoustic difficulties of the ten languages, table 1 presents the phoneme error rates, based on a phoneme recognizer without any phonotactic constraints. The results indicate significant differences in acoustic confusability between languages, ranging from 33.8% to 46.4% phoneme error rate. The Japanese language seems to be one of the easier languages with respect to acoustic confusability. This can be explained by its *mora* structure, and the resulting low phoneme trigram perplexity. Among the „easier“ languages we also find French, Korean, and Croatian. The low phoneme error rate for French stems from the frequent usage of a set of mono-phonemic words that correspond to the same phoneme. For example, the phoneme /ɛ/ can stand for *ai, aie, aies, ait, aient, hais, hait, haie, haies, es*, or *est* [LAG95]. It is important to note, however, that precisely this property also increases *word* error rate, as all these words become indistinguishable on acoustic grounds. The good phone recognition results for the Croatian language reflect the near one-to-one relationship between graphemes and phonemes. In contrast to that, English seems to be the most difficult task, which is a result of the well-known weak grapheme to phoneme relation, as well as reductions and strong allophonic variations. Other „hard“ tasks are the Mandarin language, German, and Turkish. The Mandarin phoneme accuracy in this experiment is low because we chose a tone-dependent phoneme set with 141 rather than 48 phonemes. For German, frequent consonant clusters result in higher confusion rates. Overall, however, one should treat comparisons across languages cautiously as individual results also depend considerably on the condition and availability of appropriate training data and on the general maturity of development in any given language's recognizer.

| Language | Word- based | | | Phoneme-based | | |
|---|---|---|---|---|---|---|
| | ER | Vocab | PP | ER | Vocab | PP |
| Ch-Mandarin | 14.5 | 45K | 207 | 45.2 | 137 | 12.5 |
| Croatian | 20.0 | 15K | 280 | 36.7 | 30 | 9.6 |
| English | 14.0 | 64K | 150 | 46.4 | 43 | 9.2 |
| French | 18.0 | 30K | 240 | 36.1 | 38 | 12.1 |
| German | 11.8 | 61K | 200 | 44.5 | 43 | 9.0 |
| Japanese | 10.0 | 22K | 230 | 33.8 | 31 | 7.9 |
| Korean | 14.5 | 64K | 137 | 36.1 | 41 | 9.9 |
| Portuguese | 19.0 | 25K | 297 | 46.8 | 46 | |
| Spanish | 20.0 | 15K | 245 | 43.5 | 40 | 8.2 |
| Turkish | 16.9 | 15K | 280 | 44.1 | 29 | 8.5 |

**Table 1: LVCSR Systems in 10 different languages[1]**

## 3.3 Acoustic Model Combination

For the integration of monolingual speech recognizers into one global multilingual engine we propose the combination of acoustic models. Here we share acoustic models for similar sounds across languages. Those similarities can be either derived from international phonemic inventories like Sampa or IPA, which classify sounds based on phonetic knowledge, by data-driven methods, or by a combination of both. For this paper, we investigated a combined procedure for multilingual context-dependent acoustic modeling. Based on the phonemic inventory of several monolingual systems, we can define a combined phoneme set. Sounds of different languages that are represented by the same IPA symbol share one phonetic unit. Combining 5 languages in this manner reduces the size of the phoneme inventory by 41%; nine languages yield a reduction of about 50%. Half of the phonetic units consist of phonemes only belonging to one language.

For monolingual systems, modeling wider contexts has been shown to increase recognition performance significantly. Extending context dependent models to a multilingual setting requires algorithms that can automatically construct them. In a multilingual system, we build context dependent models by initially assigning one model to each phonetic unit and training this model by sharing the data of all languages belonging to this phonetic unit. We then use a divisive clustering algorithm that creates context querying decision trees. As selection criterion for dividing a cluster into sub-clusters, we use the maximum entropy gain on the mixture weight

[1] Word-based error rates for Mandarin and Korean are given in characters, for Japanese in hiragana

distributions. This clustering approach provided significant improvements across different tasks and languages [FR97].

We investigated two variations on building the decision trees: Either all training data is shared without using language information or the information about which language the data belongs to is provided to the algorithm. In the latter case, adding questions about the language or language group to which a phoneme belongs enhances the set of context questions for the decision tree clustering. The decision of whether language information should be included with the phonetic context information is therefore performed on a case-by-case basis and depends only on the training data.

When recognizing data from a language that is part of the training set, our results show that acoustic model combination achieves better results if the language information is preserved. This observation is consistent with results from other studies [CDG97], [Koe98]. However, blind data shared models perform better if the recognition experiments are performed on languages that are not in the training set, which can be explained by an augmented language robustness achieved through sharing all information across languages.

## 3.4 Language Adaptation

Currently, one of the major time and cost factors for developing LVCSR systems for new languages is the need for large amounts of transcribed audio data for training accurate acoustic models. To accommodate potential variations in the amount of training data available for the target language, we address three topics of research:

| | |
|---|---|
| - Cross-language transfer | no data |
| - Language adaptation | very limited data |
| - Bootstrapping | large amounts of data |

The term *cross-language transfer* refers to the technique of using a recognition system from on language on a new language without having ever seen any training data in the new language. Research in this area investigates whether cross-language transfer between two languages of the same family performs better than across family borders [CC97], and whether the number of languages used for training the original acoustic transfer models influences the performance on the target language [GG97], [SW98b]. Results indicate a relation between language similarity and cross-language performance [CC97], [BKI97]. Furthermore our own work as well as of others [BKI97] have clearly shown that multilingual transfer models outperform monolingual ones [SW98a].

In a *language adaptation* technique, an existing recognizer is adapted to the new target language using very limited training data. Ongoing research ([WKA94], [Koe98], [SW98c]) concentrates on two issues: first, the amount of adaptation data needed to get reasonable results, and second, finding suitable initial acoustic models . As expected, language adaptation performance is strongly related to the amount of data used for adaptation. The results in [WKA94] demonstrate that the number of different speakers used for training is more critical than the number of utterances. In [SW98c] we investigated the issue of finding suitable initial models, comparing the effectiveness of multilingual acoustic models to monolingual models. Again, it could be shown in our own work (and confirmed by [Koe98]) that multilingual models outperform monolingual ones [SW98c].

The key idea in a *bootstrapping* approach is to initialize the acoustic models of the target language recognizer by using seed models that have previously been developed for other languages. After this initialization step, the resulting system is completely rebuilt using large amounts of training data from the target language. We have applied this approach in [OAM92] to bootstrap a German recognizer from English. Work by [GFG95] and [WKA94] confirms that cross-language seed models perform better than flat starts or random models. In more recent work, we could demonstrated the advantages of *multilingual* phonemic inventories and *multilingual* phoneme models as seed models [SW97].

We exploited the LVCSR performance of multilingual acoustic model combination by porting a multi-lingual recognition engine to new target languages comparing cross-language transfer, language adaptation, and bootstrapping. Our results indicate that language adaptation clearly outperforms bootstrapping and cross-language transfer. Bootstrapping performs better than cross-language transfer, even if only a very small amount of training data (about 10 minutes) is available.

Assuming that only a small amount of adaptation data is given, the performance on a new target language is mainly impaired by a considerable mismatch between the models built to match phoneme contexts observed during training on multiple languages and the actual phoneme contexts occurring in the new target language. Therefore, the high gain in performance achieved by language adaptation results from the specialization of these wide context models to the new target language [SW99]. Our results emphasize the importance of even a small amount of data for acoustic model adaptation and context specialization.

# 4. Non-native Speech in Multilingual Systems

Though multilingual systems handling a number of major languages broaden the reach of speech recognition technology to consumers around the world, it is to be expected that many users are not native speakers of the input language they have chosen to use. Not all languages can be covered by a multilingual system, so speakers of unavailable languages would need to use a second language. Even if the native language of a speaker is available as input language, he/she may prefer or need to use a second language. This can be the case for professionals in specialized fields who are not accustomed to using their native language at work, or for users who simply wish to use more than one language. Non-native speech is encountered in travel and business situations, for travelers and visitors to foreign countries, or in business or technical collaboration across national boundaries. Last not least, the recognition of non-native speech is required for educational applications like language tutoring.

Non-native spoken input can be a major challenge for speech recognition. Pronunciation, disfluencies, lexical choice, use of filler words, syntactic structure, and pragmatic goals can deviate considerably from the patterns that are found in native speech.

In this section, we focus on the problem of foreign-accented speech, looking at both non-native pronunciation (phoneme realization typical of a specific speaker group) and pronunciation errors (phonotactic errors and other speech errors associated with the articulation of an unfamiliar phoneme sequence). We then describe several approaches to acoustic modeling for non-native speech. Finally, we touch on the higher-level issues of word and structural choice, discussing the effects of non-native usage on language modeling and natural language understanding.

## 4.1 Characteristics of foreign-accented speech

Human listeners can adapt to accented speech. Most native speakers of English living in the United States, for example, can understand Spanish-accented English without difficulty, perhaps subconsciously performing a phonemic mapping. Even young children are able to imitate foreign accents, showing an ability to detect and identify common phonemic substitutions present in accented speech. Since many of the features found in speech vary considerably between native speakers, it is difficult to identify a boundary beyond which such variations are perceived as foreign accent. What are some of the dimensions along which native and non-native pronunciation can be distinguished?

### 4.1.1 Phoneme realization

Since stress patterns and durations play only a minor role in most speech recognition systems, the most important difference between non-native pronunciation and native pronunciation is in phoneme realization. Phonemes in a language that is not native to the speaker (the target language) can be placed into one of two categories: phonemes for which there is an obvious counterpart in the speaker's native language (the source language), and phonemes for which there is not. The perception that a source-language counterpart for a target-language phone exists is often based on acoustic similarity, but can also be influenced by orthography; a speaker may tend to substitute a phone that is quite dissimilar to the target phone acoustically but is represented by the same symbol in text.

When there is an obvious counterpart in the speaker's native language, phoneme realization errors can sometimes be attributed to linguistic transfer, although many studies have indicated that this is the case less frequently than it may seem (see [Bee80], e.g.). Transfer effects in pronunciation can range from slight deviations in place or manner of articulation to exact substitution of a native language phone for a target language phone; in some cases, the speaker does not even perceive a difference between the source and target language phones. Even when speakers of the same native language consistently substitute a specific source language phoneme for a specific target language phoneme, variation among those speakers' articulations of the source language phoneme can be significant, meaning that a seemingly straightforward mapping can be quite complex to model. Non- exact substitutions are even more difficult to model, as individual speakers' realizations can fall anywhere between the source and target phone, and may exhibit features that are not present in either the source or target phone.

When there is no obvious counterpart for a target phone in the source language, the speaker must approximate it as best he can. This can result in a realization that is unsystematic both within one speaker's speech and across speakers.

### 4.1.2 Articulation of phonemes in context

If the only deviation in the non-native speaker's pronunciation is in the realization of individual allophones, high-quality recognition can often be achieved with speaker adaptation [Sch97]. However, many non-native speakers differ from native speakers in the way phonemes are articulated in certain contexts. Native speakers of German speaking English, for example, may tend to devoice consonants at the end of a word in places where a native English speaker would not. As many speakers are unaware of allophonic variation in their own native speech, generating the correct allophones in context in the target language can be very difficult. In a recognition system in which phonetic contexts have already been clustered based on allophonic variations observed in native speech, codebook adaptation will not perform optimally as the contexts cannot be adapted separately.

### 4.1.3 Phonotactic constraints

A third source of errors in target language production lies in the phonotactic constraints of the source language. Different languages allow different sequences of phonemes, and attempts to pronounce phoneme sequences to which one is not accustomed can fail. Many of the phoneme combinations that appear in English are difficult for native speakers of other languages, and although it is possible to learn to pronounce them, it is common to make use of other strategies. Insertion of vowels (known as *epenthesis*) is one way to make a consonant clusters easier to articulate; Japanese speakers may pronounce *try* [T OW R AY], which is very confusing for a native English listener. Epenthesis is not limited to cluster-internal position: source language constraints on consonants in word-initial or word-final position may cause speakers to insert vowels in those positions, as in the Japanese-accented [B AE G G U W] (*bag*) or Spanish-accented [EH S K UW L] (*school*).

Even when a phoneme sequence is pronounceable, and is realized correctly in careful speech, the timing with which articulation of sequential phonemes is initiated, a largely subconscious process, can be incorrect and cause such phenomena as phoneme inversion in conversational speech. Words like the German '*sprichst'*, for example, can be problematic for English speakers, some of which tend to invert the final two consonants.

## 4.2 Acoustic modeling for non-native speech

One way to increase the robustness of a recognition system with respect to foreign accent is to develop accent-specific or accent-tolerant acoustic models. The former may be desirable when the source-target language pair is known and sufficient training data is available. The latter may be more appropriate when the system must handle a variety of different foreign accents.

### 4.2.1 Non-native models

If the source-target language pair is fixed, and enough non-native training data is available, models representing non-native pronunciation can be explicitly trained. This approach is most appropriate for a system designed to accurately recognize the speech of a specific non-native speaker group. This is essentially a bootstrapping approach, and brings with it the advantages (accuracy) and disadvantages (data requirements) discussed in section 2.2.

### 4.2.2 Bilingual models

An alternate way to allow pronunciation that is typical of a particular non-native speaker group is to include models from both the source and target languages in the acoustic model set. If the speaker's articulation of an /r/, for example, is much closer to a phoneme in the source language than the intended target phone, allowing the system to recognize the source phone may result in improved overall accuracy. With a bilingual acoustic model set, two sets of models are trained separately on different data. Target-language models can be taken from an existing target-language system. Source-language acoustic models can be taken from an existing source-language system, as in [Kaw99], or can be trained with data from heavily accented speakers, as in [RNF97]. Criteria must then be defined for allowing transitions between target and source language models, and ensuring that model sets are compatible.

### 4.2.3 Model merging

When source and target language models from compatible systems are available, it has been observed that "merging" the models can significantly improve recognition of non-native speech [WY99]. Witt and Young have reported that by combining Gaussian mixtures from corresponding source and target language models into a new model with twice as many mixture components, an increase in performance can be achieved that is greater than that given by creating new models composed of linear combinations of source and target model components.

### 4.2.4 Dictionary modification

A straightforward way to allow for non-native errors is to include common phonemic substitutions in the pronunciation dictionary. This approach can be implemented in off-the-shelf recognition packages, which may not be otherwise easily modified. Auberg et al. report success in modifying the IBM ViaVoice system to create an application which teaches users to discriminate, identify, and produce sounds that are recognized as being problematic for Japanese learners of English [Aub98]. Dictionary modification can also be used to model systematic phonemic shifts among speakers of different varieties of the same language, as discussed in [HW98]. This strategy can be used when no acoustic training material for the source language is available but basic pronunciation mappings can be derived.

## 4.3 Beyond accent: addressing non-native usage

Although modeling non-native pronunciation in the acoustic model can help to increase recognition accuracy on non-native speech, idiosyncrasies in non-native speech do not stop with pronunciation; non-native usage at the lexical and phrasal levels will need to be modeled to achieve accurate recognition of non-native speech. In this section we report on a series of experiments comparing linguistic features of native and non-native spontaneous and read speech.

### 4.3.1 Perplexity and frequent trigrams

Although it is difficult to make a judgement about grammaticality non-native conversational speech (even native conversational speech is often ungrammatical), measuring the perplexity and common phrases of a transcribed spoken corpus can help to quantify the ways in which non-native speech is unique. For a test corpus of tourist queries posed in English by native speakers of English, Japanese, and Chinese[2], the trigram perplexities for the two non-native speaker groups were significantly lower than those for the native speakers. The trigram hit rates were similar, but the set of most frequently used trigrams was quite different, suggesting that while non-native speakers are using phrases that are indeed common in native speech, they are not the ones the native speakers would use in a particular semantic and pragmatic context. This has implications not only for language modeling but also for parsing and translation, for which query formats favored by non- native speakers will need to be represented.

| Speaker group | Perplexity | Trigram hit rate | Common trigrams |
|---|---|---|---|
| Japanese | 66.5 | 55.8 | can i get, do you know |
| Chinese | 74.4 | 52.9 | the name of, can i go |
| English | 102.6 | 48.6 | i need to, you tell me |

### 4.3.2 Disfluencies in spontaneous speech

It has been observed that native speech contains many instances of abandoned words, stutters, restarts, filler words, and other disfluencies, some of which occur systematically enough to warrant incorporation in the language model ([SS96], e.g.). Disfluencies often occur when the speaker is searching for the right word or expression, or is pronouncing a word that is difficult to articulate. In our study, such situations arose more often for the non-native speakers than for the native speakers, and examination of their disfluencies showed a high incidence of both incomplete words and filler words, although much more so for the Japanese than the Chinese speakers.

| | % of stumbles[3] | | % of filler words | |
|---|---|---|---|---|
| Speaker group | spontaneous | Read | spontaneous | read |
| Japanese | 1.46 | 2.48 | 4.37 | 0.25 |
| Chinese | 0.83 | 1.31 | 1.46 | 1.31 |

[2] Non-native speakers tested at the novice-to-intermediate level; all had studied English for more than 8 years, but had been in an English-speaking environment for less than one year. The database is discussed in greater detail in [MB99].

[3] A stumble is a completed fragment, usually due to difficulty in articulation or visual decoding (...many env= environ= environmentalists oppose the law...) as opposed to an abandoned fragment (...many env= many researchers oppose the law...)

# 5. 5. Dynamic Lexical Adaptation

The quality of a speech recognizer is heavily influenced by the correspondence of the recognition dictionary used and the actual vocabulary of the utterances to be recognized. If a high percentage of the words to be recognized is not included in the dictionary, a large number of misrecognitions is triggered. This especially applies to open domains like dictation systems or the automatic transcription of broadcast news where the recognition dictionary cannot be constrained to a predefined vocabulary. Instead, an unlimited vocabulary is required. If the language to be recognized has a large number of inflections and composita, like German, Serbo-Croatian and Turkish, for example, the vocabulary grows even faster and the problem of new and unknown or out-of-vocabulary words worsens.

This section first introduces some morphological properties of languages and also possible solutions to the problem of out-of-vocabulary words. Finally, we introduce methods to improve the reliability and performance of speech recognition systems for continuous speech on large vocabularies by overcoming the limitation of the recognition dictionary to a certain size N. Even though the recognition vocabulary is still finite, the methods presented here allow for a virtually much larger vocabulary by dynamically adapting the dictionary to the speech data at hand [Geu99a]. Based on the idea of vocabulary adaptation, a multipass strategy called **H**ypothesis **D**riven **L**exical **A**daptation (HDLA) is developed and results on Serbo-Croatian, German and Turkish broadcast news data are presented [Geu98a].

## 5.1 Morphological Properties of Languages

As described in section 1.1, two major groups of languages can be distinguished when comparing their morphological properties: languages like English that show an exceptionally simple morphological structure, and morphologically rich languages like German, Serbo-Croatian and Turkish. Whereas English only has a small variety of different inflection endings both for verbs and nouns, highly inflected languages have a very large number of distinct verb conjugations and noun declinations. Taking the German word "kommen" ("to come") as an example, the difference between the two language groups becomes clear: whereas in English the present-tense conjugation of this verb consists of just two distinct endings, the number is twice as large for German where there exists a different ending for almost every person in the singular and the plural. Table 4.1 illustrates the differences between German and English for this example.

| **German** | | **English** | |
|---|---|---|---|
| Ich | komm-e | I | Come |
| Du | komm-st | you (sg.) | Come |
| er/sie/es | komm-t | he/she/it | come-s |
| Wir | komm-en | We | Come |
| Ihr | komm-t | you (pl.) | Come |

| | | | |
|---|---|---|---|
| Sie | komm-en | they | Come |

**Table 5.1 Examples of inflection endings for German verbs.**

Additionally, the German language has an uncountable number of compound words. Formation of these composita is not only possible for nouns but also for verbs. Several prefixes can be attached to every verb, each time creating a new word. The same applies to noun composita. Nouns can be concatenated to long noun chains, every chain creating a word with a new meaning. Naturally, these characteristics of morphologically rich languages lead to a much faster vocabulary growth over the same amount of training data than morphologically simple languages like English. Figure 4 shows a comparison of the vocabulary growth for Chinese, Serbo-Croatian, Japanese, Portuguese, Russian, Spanish, Turkish, German and English on broadcast news transcripts and newspaper articles.

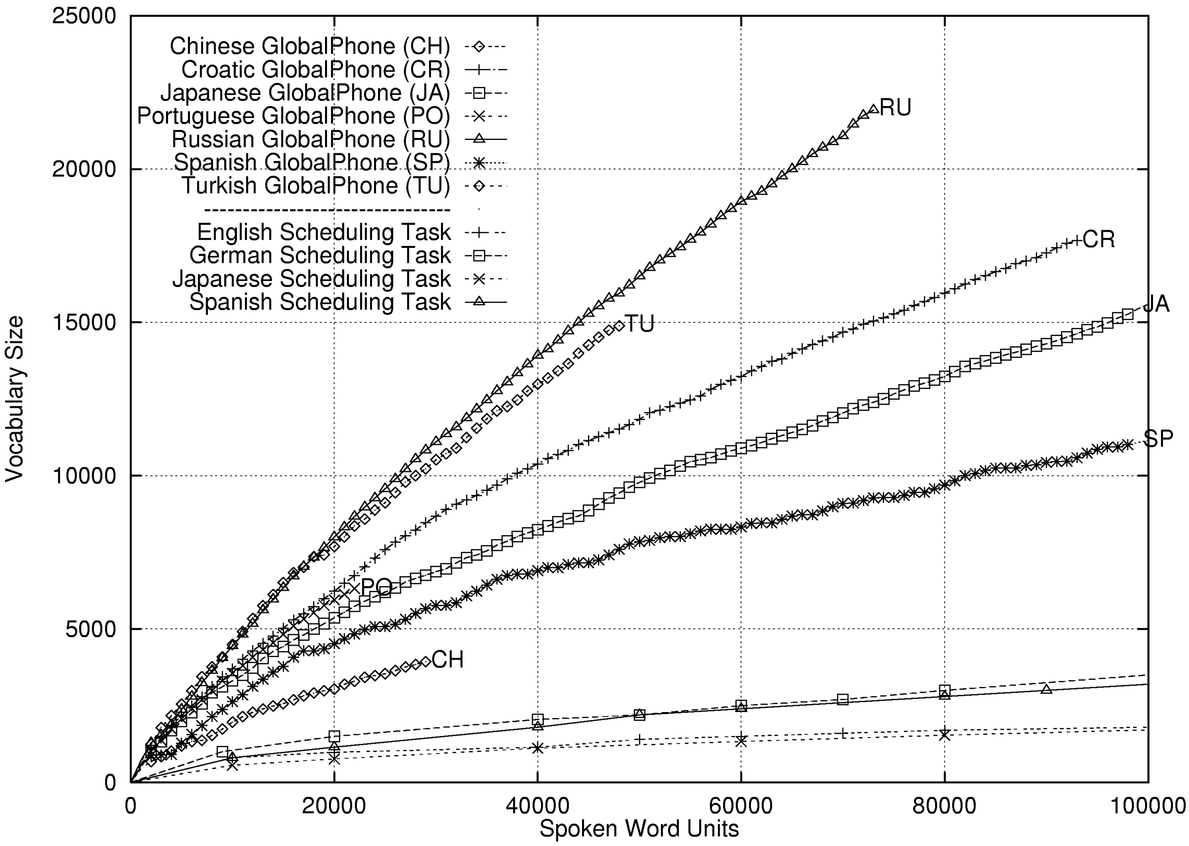



**Figure 2: Vocabulary Growth for several languages**

A consequence of a very fast vocabulary growth is the resulting large out-of-vocabulary rate for a given dictionary size. For a task like broadcast news, the out-of-vocabulary rate for English using a dictionary with 60,000 words is less than 1%. Much higher rates are encountered for languages like German, Serbo-Croatian or Turkish: a broadcast news recognizer for the Serbo-Croatian language with a comparable dictionary size shows an out-of-vocabulary rate of about 8%. As each out-of-vocabulary word causes one or more recognition errors, high out-of-vocabulary rates significantly worsen recognition performance.

## 5.2 Approaches to the Out-Of-Vocabulary Problem

One possibility to counteract both a fast vocabulary growth and high out-of-vocabulary rates is the usage of base units other than words for the recognition process. To this end, syllable-based as well as morpheme-based decompositions of words have been used as recognition units. Instead of a dictionary of words the underlying recognition lexicon consists of subword units. The coverage of such a dictionary by subword units is significantly better than the coverage of a dictionary of the same size comprised of conventional words. However, recognizers built on top of these units suffer a severe degradation in the performance measured at the word level, because many now hypothesized morpheme sequences do not map to legal words. To make matters worse, short morphemes (suffices, prefixes) are also more confusable than long words. As an alternative approach, the idea of a dynamic expansion of the recognition dictionary has been investigated. Words are still considered as the dictionary units for recognition. But instead of having a static dictionary of those words, a dynamic dictionary is introduced which has the same fixed size as the static dictionary but is tailored on the fly to each utterance. Since the recognizer uses a different customized dictionary for every single utterance, the total size of the recognition dictionary is virtually unlimited.

## 5.3 Dynamic Lexical Adaptation

Hypothesis Driven Lexical Adaptation (HDLA) is a technique for dynamically adapting the dictionary of a speech recognizer [Geu98b]. It still treats the size of a dictionary as finite, but allows for a larger number of 'virtual' words to be recognized. This is done by abandoning the notion of a fixed static dictionary. Instead, we exchange vocabulary entries from the recognition dictionary, dynamically, depending on the actual speech input. A two-pass recognition procedure is the basis for this vocabulary adaptation strategy. The first pass provides the necessary information needed to exchange the vocabulary entries of a general baseline dictionary by words similar to the actually uttered or hypothesized words. The second performs another recognition run on the adapted vocabulary that has a lower out-of-vocabulary rate, resulting in a lower word error rate. The dictionary used for both recognition runs has a fixed size, but the individual vocabulary entries are exchanged. Through this approach the lexicon is adapted to the actual speech utterance and an optimal vocabulary is created for each recognition subtask. Simultaneously, any size limitations of the dictionary imposed by implementational issues or computing resources are overcome and speech recognition on a virtually unlimited vocabulary is possible.

For the selection of the vocabulary entries incorporated into the recognition dictionary for the second recognition run, knowledge about morphological and phonetic affinity of actually uttered and hypothesized words is incorporated into the adaptation procedure. The expectation is that a dynamically adapted recognition dictionary, constituting an utterance-specific vocabulary for the speech segment to be recognized, reduces the number of out-of-vocabulary words and thereby improves recognition performance. Especially when transcribing broadcast news, this should keep the out-of-vocabulary rate limited and thus improve the word error rate.

### 5.3.1 The HDLA Algorithm

The algorithm below describes the steps of the Hypothesis Driven Lexical Adaptation (HDLA):

1. A first recognition run on a general domain-specific recognition dictionary generates word lattices and an utterance-specific vocabulary list.
2. This vocabulary list is then used to look up all similar words in a large background lexicon that contains words from large available text corpora.
3. All similar words are then incorporated into the original recognition vocabulary by replacing the least relevant words that did not show up in the lattice, so that the dictionary size of the recognizer does not change.
4. In an automatic procedure a new dictionary and language model are created and used to perform a second recognition run.

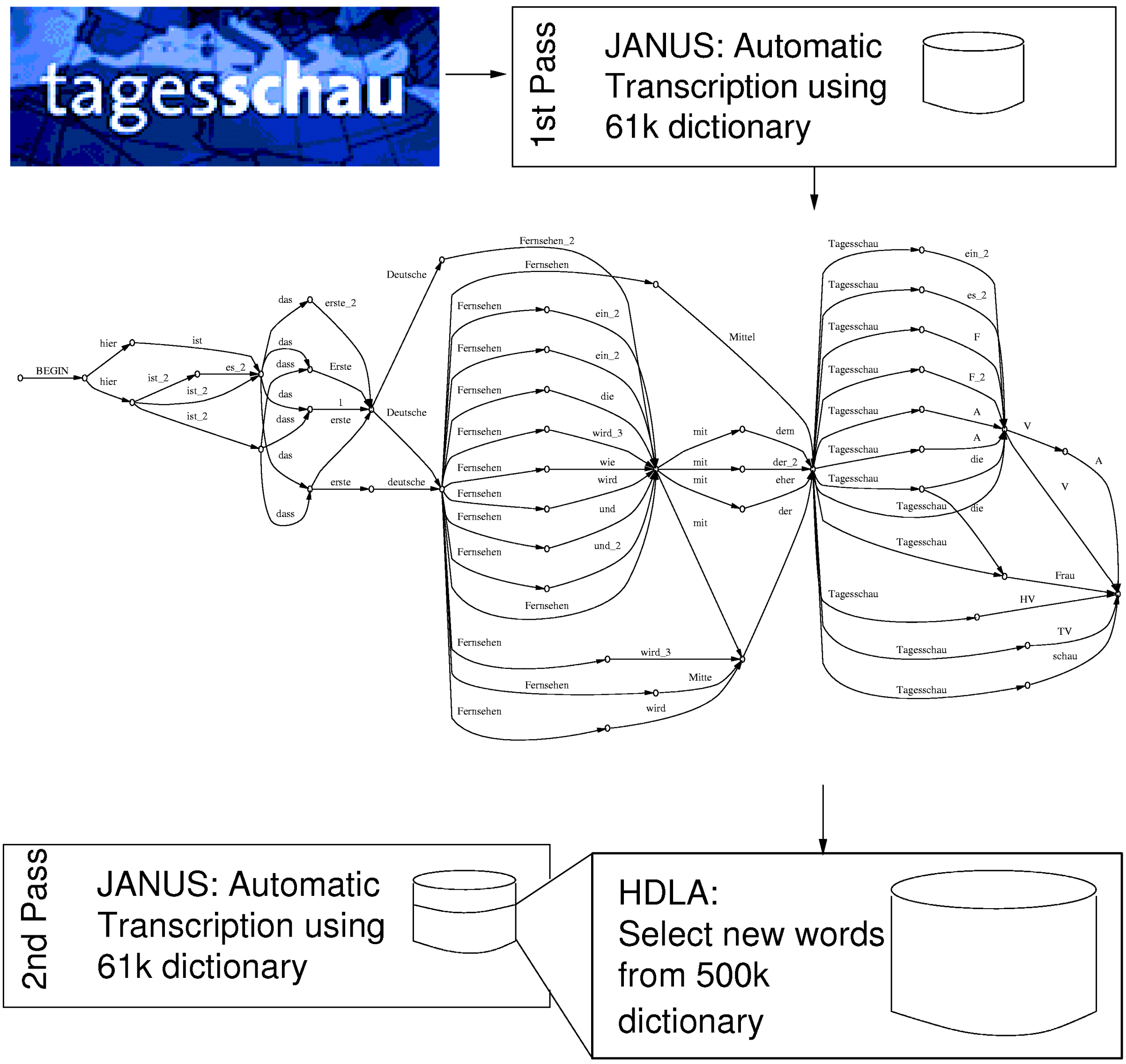



**Figure 3: The HDLA framework**

Figure 3 illustrates HDLA applied to German broadcast news. Applied to Serbo-Croatian and German broadcast news data it yields significant improvements both in out-of-vocabulary and in word error rate.

### 5.3.2 Different Selection Criteria

Various criteria for selecting entries for the adapted vocabulary can be applied [Geu99b]. Figure 4 summarizes the ideas and methods that have been used to generate customized dictionaries:

1. Selection from large dictionaries based on morphological similarity::
   For the morphology-based approach, two words are considered similar if they share the same word stem and only differ in their inflections (*morphological similarity*). Similarity is determined linguistically by morphemic rules.

2. Selection from large dictionaries based on orthographic of phonetic similarity:
   To estimate phonetic similarity, we introduced various distance measures that are either based on the letter sequence of words (grapheme-based) or their phoneme sequence (phoneme-based) [Geu98c]. For the phoneme-based approach three different methods of calculating the phonetic distance were used: the equality of two phonemes, the Hamming distance with respect to a binary vector of acoustic features, or the acoustic confusability of phonemes. In this approach, compounds can also be taken into consideration when determining word distances.

3. Creation of artificial large dictionaries and selection based on phonetic similarity:
   If no large database is available for a given task or language, language-specific morphological rules for generating inflections can be applied to create an artificial fallback lexicon. We then compute the phonetic distance between its entries and the entries of the vocabulary list from the first recognition run. Candidates beneath a certain threshold are included in the adapted dictionary.

4. World-Wide-Web-Based Retrieval for dictionary creation:
   Last not least, information retrieval on the World-Wide-Web has been applied to retrieve texts that are similar to the hypothesized output in order to create suitable customized dictionaries. Two approaches have been evaluated: the first employs a search engine to retrieve similar texts; the second uses the topicality of a news show to retrieve similar texts.

Based on these methods, we have implemented several algorithms to select the customized vocabulary for the second recognition.

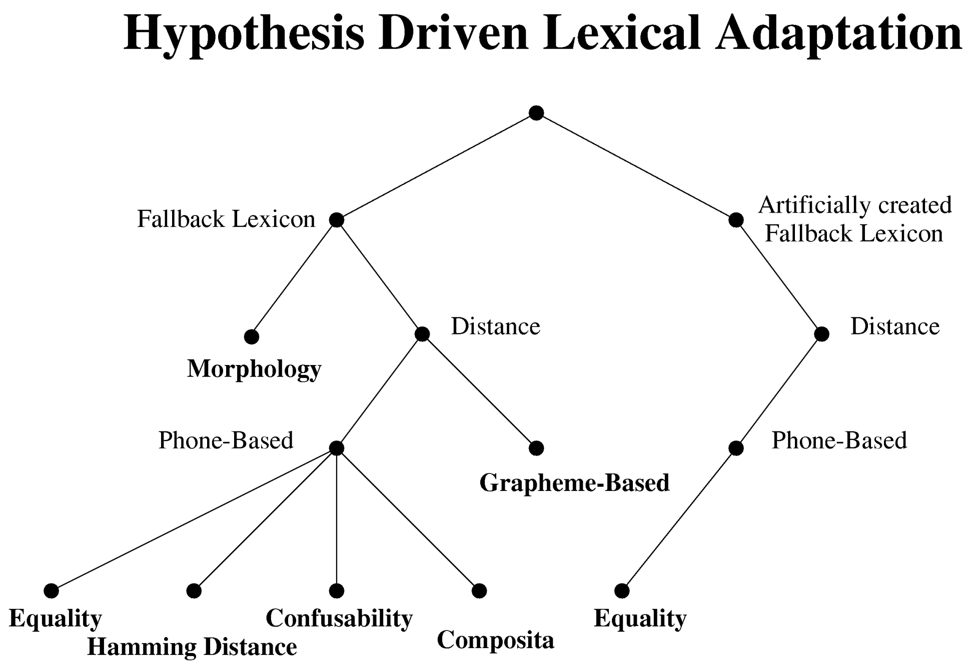



**Figure 4: Selection Criteria**

### 5.3.3 Results

Depending on the special characteristics of a language to which HDLA is applied, different procedures lead to optimal performance. Table 5.2 summarizes the results achieved by applying the HDLA- algorithm to Serbo-Croatian, German, and Turkish recognition. Note that the aim of these experiments is to establish relative improvements. Absolute error rates are higher than comparable systems in English in part due to the limited language resources available in these languages.

| | **Serbo-Croatian** | **German** | **Turkish** |
|---|---|---|---|
| | **OOV-Rates** | | |
| Baseline | 8.7% | 4.4% | 14.9% |
| Morphology-based | 4.8% | 2.9% | 10.9% |
| Grapheme-based | **4.0%** | -- | -- |
| Equality | **4.0%** | 3.1% | -- |
| Hamming distance | 5.4% | -- | -- |
| Acoustic confusability | 4.4% | -- | -- |
| Phoneme-based (Composita) | -- | 2.1% | -- |
| Artificially created fallback lexicon | 5.8% | -- | -- |
| Information retrieval based | -- | 3.2% | -- |
| Topicality-based | -- | **1.9%** | -- |

**Table 5.2 OOV-rates for Serbo-Croatian, German, and Turkish data**

The recognition experiments for both Serbo-Croatian and German show that the reduction in out-of-vocabulary rate leads to significant reduction in word error rate. Applying the HDLA procedure to Serbo-Croatian broadcast news, the percentage of new words decreases from originally 8.7% to 4.0%. This results in a reduction in word error rate from 29.5% to 25.4%.

| | Vocabulary Size | OOV-Rate | Word Error |
|---|---|---|---|
| Baseline | 49,000 | 8.7 % | 29.5 % |
| Morphology-based HDLA | 49,000 | 4.8 % | 26.0 % |
| Phoneme-based HDLA | 49,000 | 4.0 % | 25.4 % |

**Table 5.3 Serbo-Croatian recognition results based on adapted vocabularies**

For German, a 57% reduction in the number of unknown words from 4.4% to 1.9% can be achieved. The baseline recognition result of 24.7% word error rate can be improved to 23.1%.

| | Vocabulary Size | OOV-Rate | Word Error |
|---|---|---|---|
| Baseline | 61,000 | 4.4 % | 24.7 % |
| Topicality-based HDLA | 61,000 | 1.9 % | 23.1 % |

**Table 5.4 German recognition results based on adapted vocabularies**

## 5.4 HDLA Conclusion

Speech recognition systems for conversational speech have to be able to handle very large vocabularies, as spontaneous speech input cannot be restricted to a predefined vocabulary or domain. Therefore, unforeseen words can always occur and cannot be included in a static recognition dictionary. Each of these out-of- vocabulary words will automatically lead to one or more recognition errors and thus worsens recognition performance significantly. As an indefinite expansion of the size of the actual recognition dictionary is not possible, other ways have to be found to reduce the out-of-vocabulary rate of a speech recognizer. The HDLA approach presented above is a successful method for this purpose.

When looking at the resulting number of out-of-vocabulary words and recognition performance, it is interesting to see that different methods and selection criteria are better suited for different languages. While either grapheme or phoneme based distances turn out to be optimal for Serbo-Croatian, German (excluding compounds) appears to improve optimally using a morphology-based approach. Errors due to compound nouns (common in German), by contrast, can be improved using a phoneme-distance based distance selection. These results demonstrate that it is helpful to consider the special characteristics of a language when trying to find useful selection criteria for the lexical adaptation procedure.

In our effort to control unmanageable vocabulary growth in heavily inflected languages, Hypothesis Driven Lexical Adaptation was shown to be an effective means for reducing the rate of out-of-vocabulary words. Using a two-pass recognition strategy for German and Serbo-

Croatian Broadcast News transcription, a significant reduction of up to 57% could be achieved in the out-of-vocabulary rate, resulting in word error rate reduction by up to 14%.

# 6. Multi-lingual Speech Translation

Perhaps the most challenging task for multi-lingual speech and language processing is the automatic translation of spontaneous speech. Possible applications include international e-commerce, help desks, mobile translation systems for travelers, automatic generation of television subtitles, and the translation of telephone conversations.

In the following, we will highlight some of the challenges in speech translation, and review present current approaches and solutions. A brief overview of the C-STAR speech translation consortium and the Janus speech translation system will also be given.

## 6.1 Challenges

Many of the known problems of bilingual text translation, such as dealing with lexical ambiguity, anaphora, and idiomatic expressions, occur also in multi-lingual speech translation. A number of problems, however, are specific to the translation of spoken language and to the requirements of providing speech translation for multiple languages.

### 6.1.1 Translating Spoken Language

Many of the problems in automatic speech translation are introduced while transforming the input speech to tokens that can be used for translation[4]. The most obvious of these problems are recognition errors. Since dialogues usually contain spontaneously spoken utterances that are less well formed than those found in read speech, word error rates around 10 to 40% are still to be expected. This implies that there is at least one recognition error in every other utterance. Ignoring recognition errors, grammar coverage for the translation of completely correctly recognized utterances is typically between 70 and 90%. If the speech translation process were approached as speech recognition with subsequent text translation, the errors introduced by the individual steps would accumulate to overall unacceptable end-to-end performance.

Another set of difficulties is introduced by the fact that spoken language in dialogues differs considerably from written language. Ungrammatical utterances ("I mean would you?"), colloquial expressions, isolated fragments ("To Boston at ten.") and the lack of punctuation cause traditional text translation engines to fail. Since spoken utterances are less carefully planned, they can even be self-contradicting, as in the following example from our user studies:

| | |
|---|---|
| **question**: *can you book a flight*? | **answer**: *no, that's not a problem* |

[4] Translating from a string that is similar to the orthographic transliteration of the speech is not the only option: it is also conceivable to run a translation engine on a recognized phoneme string.

Participants in a spoken dialogue are more likely to refer to common experience than an author who does not know his readers. This introduces additional levels of ambiguity. However, spoken dialogues contain many clues that are missing in written language, such as prosody, timing, and references to the current visual context. To efficiently translate speech, this information has to be integrated.

Finally, the situations in which speech translation can be used impose constraints on the translation time and on the amount of data available for disambiguation: usually, an utterances by one speaker has to be translated before the other speaker's turn. Therefore, near real-time processing is important, and only data from earlier utterances can be used for disambiguation. If the setup allows no text output (telephone), the target language output has to be intelligible when spoken by a text-to-speech program (short sentences, prosodic hints).

### 6.1.2 Translating Speech in multiple languages

For multi-lingual speech recognition, a single multilingual engine (as described in section 2) or a set of monolingual recognizers can be used. For adding a new language to the system, the effort is limited to providing a recognizer for that language. For the translation phase, however, each language pair has to be considered. If any components in the translation system depend on both, source and target language, the effort for adding a new language increases with the number of languages already in the system. Later in this section, we will show that there is a tradeoff between the effort to add languages and the ease of expanding a system to new tasks.

Due to the structural differences between language groups, appropriate analysis and generation algorithms differ between languages. The Japanese language, for instance, does not have blank spaces in the written form, which makes the definition of a dictionary unit for recognizer and parser difficult. The Japanese •••••••••••• (Heyawoyoyakushitainodesuga), is an unbroken string of characters approximately equivalent to "I would like to reserve a room, but…". Studies have suggested that rule-based (e.g. [Pal97]) and statistical (e.g. [MR98]) algorithms can be used to automatically extract units from unsegmented text that are appropriate for both, recognition and semantic parsing, but determining the segmentation which leads to optimal translation accuracy remains a challenge. The problem of base unit determination is not limited to languages without spaces in written text. Turkish and Korean, as described in section 1.1, are agglutinative and must be segmented further than they are in their written form for speech recognition and translation. Some languages, for example Spanish, are written with relatively short words but require extensive morphological analysis. For other languages, such as English, the few inflected forms can be enumerated in the analysis grammar. For target language generation, languages with extensive agreement requirements (Spanish, German) require additional attention over languages where cases requiring agreement are rare.

Another problematic aspect of dealing with multiple languages lies in the cultural differences. In some languages for instance, it is considered impolite to say *‘no’*, and native speakers will rather say things such as *‘that may be difficult’* or just switch the topic. In a machine-interpreted dialog, the implication may not be clear to the English speaker unless the system finds a way to point it out. In the other direction, translating a flat *‘no’* into Japanese may be considered as very impolite. Other cultural differences include task- and language-dependent expressions. The *'queen sized bed',* while ubiquitous in American hotels, is a concept that does not exist in Germany and France, where beds come in the equivalent of twin and king-size only. In Germany, there is a room-rate called *'Halbpension'*, which includes breakfast and dinner, but there is no adequate equivalent in English. In such cases a translation system may have to insert brief explanatory sentences in order to be understood.

Many ambiguities that exist on a semantic level are not perceived as such by the speaker of the source language, but the missing disambiguation information can cause trouble when generating a target language that requires it. One example for this are numbers in Japanese. Consider translating the one-word utterance 'two' into Japanese: *two* as in 'two long objects' is "ni-hon" while *two* as in 'two flat objects' is "ni-mai' and *two* as in 'two people' becomes "futari". A similar example is the explicit mention of the subject in English, which may be missing in a normal Japanese utterance. Further information that is often missing when translating dialogues is the gender of the speaker and listener as well as the social relationship between them.
For some language pairs, however, maintaining the level of ambiguity present in the source language can help to avoid clumsy, confusing, or (when the ambiguity is incorrectly resolved) inaccurate translations.

## 6.2 Multi-lingual Speech Translation approaches

Given the problems outlined in the previous section, speech translation may seem an impossible task. A look at possible scenarios, however, indicates that while a "universal speech translator" may be beyond our current reach, speech translation in some very useful limited domains is feasible. The most important group of scenarios involves goal-oriented dialogues such as shopping, getting information, and scheduling events.

The assumption that the translation domain is constrained to such a scenario leads to several simplifications of the speech translation task: there will be less ambiguity within one task, and the conversation will be more polite and less colloquial; ill-formed spontaneously spoken utterances can be interpreted exploiting the semantic constraints of the domain despite syntactic deficiencies of the spoken utterance. Finally, the relationship between the participants is usually clear from the scenario (salesperson-customer). Proper names will be used more cautiously and may even be spelled. Utterances in such dialogues can be classified by their "domain action", that is by their achievement towards the dialogue goal (for instance giving information about a flight). Moreover, task-dependent idiomatic expressions abound in such dialogues. Therefore, semantic representations become an important tool for the translation of goal driven conversations.

### 6.2.1 Interaction between Recognition and Parsing

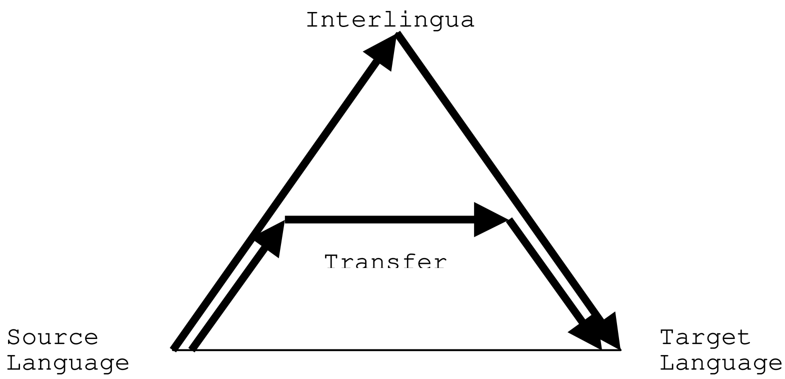
A common bottleneck in speech translation systems lies between the speech recognition output and the translation or analysis step. A number of methods are used to reduce the accumulation of errors at this point. Parsers used for speech translation are designed to accommodate repetitions, hesitations, and speech recognition errors by skipping input words or by parsing word graphs that include alternatives to the most likely recognition hypothesis. Some unknown words (unexpected proper nouns not known to the recognizer) can be represented by their phonetic transcription.

Unless the speaker asks for the exact spelling, it is often irrelevant in spoken dialogue translation[5]. When parts of the utterance cannot be translated, it can be helpful to skip them and provide a translation for the remaining parts of the utterance. Another approach for a close integration of recognition and analysis is to have the recognition engine use the robust analysis grammars to restrict the recognition search space. In this case, the parsing grammars have to provide for all ungrammaticalities. For utterances that are not covered by the parsing grammar, the next most likely interpretation will be recognized; to avoid this problem, confidence measures must be introduced.

### 6.2.2 Semantic Representations and Interlingua

Many systems perform a syntactic or syntactic/semantic analysis to extract a source language dependent representation, perform a transfer step from the source language representation to the target language representation, and then generate the target language. The number of transfer rules is usually proportional to the number of input languages times the number of output languages.
An Interlingua is a target- and source-language-independent representation of the content of an utterance. For multi-lingual translation systems, an Interlingua makes it possible to add translation between a new language and all existing languages by simply providing translation from the new language to the Interlingua and from the Interlingua to that language.

Ideally, we would like an Interlingua that is unambiguous, at least with respect to the current task. Therefore, natural languages are not very suitable as Interlingua. Moreover, goal-oriented tasks contain many idiomatic expressions that should not be translated literally. Mapping them on a language independent semantic Interlingua avoids misleading awkward translations. Special parsers have been developed to extract semantic Interlingua structures from recognition output. They work with semantic grammars that can be written by native speakers without linguistic training.




By translating from the source language to a language-independent semantic Interlingua and back into the source language, we also obtain a new feature: paraphrasing. Through the paraphrase, the

[5] proper nouns that are common enough in many countries so they have received a language dependent spellings (e.g. Muenchen, Munich), should be represented explicitly.

user can see whether the Interlingua structure that was extracted from the (possibly ambiguous) input corresponds to the intended meaning without actually looking at the Interlingua.

For many tasks, semantic Interlingua structures can also be transformed into database queries to obtain information for the task at hand or to resolve ambiguities or conflicts in translating the current utterance (next Friday is the 13th, Kyoto is in Tokyo, there is no train from New York to Frankfurt, etc.).

While systems that are based on semantic representations are easier to expand to new languages, porting to new tasks with new semantic concepts but similar syntactic forms is more expensive. For goal-driven dialogues, however, the number of semantic concepts is comparatively small and reusable grammar components (specific phrases for requesting information, e.g., times & dates, places, addresses, currencies and amounts, etc.) can be used to limit the effort.

Since an Interlingua should contain all information necessary to generate all target languages, much of the disambiguation work has to be done in the analysis step of such systems. Once a language-independent Interlingua representation has been established, the disambiguation work has to be done only once for every new source language.

### 6.2.3 Incomplete Information

Extracting a complete semantic Interlingua structure from a single utterance is not always possible. Consider the following example:

| | |
|---|---|
| **question:** *how many will be travelling* | **answer:** *two* |

While it is clear from the task that the question is about how many *people* will be travelling, it is not clear from the answer alone. Missing information can be extracted from multiple sources (dialogue, prosody, databases, default assumptions) and added to the Interlingua representation. Since there is often a source language dependent default for missing values (such as the speaker as subject in a Japanese sentence), a default can also be provided by the analysis step. It is, however, a good idea to add confidence measures and to mark information that was not in the original utterance. This way, it is possible to retain ambiguities where resolving them is not required for the target language rather than risking a possibly wrong interpretation of the input. As a last resort, the user of a speech translation system can be asked to disambiguate in critical cases ("eleven fifty five" = 1,155 or $11.55 or 11:55am).

### 6.2.4 Statistical Translation

Another well-known approach to enhance portability and to reduce the impact of multiple language pairs on the development effort is to use statistical translation approaches. For these systems to work in a multi-lingual environment, a bilingual corpus for the task in question has to be available for each language pair. Alternatively, a multi-lingual corpus for all languages can be used. This corpus should contain sufficient amounts of original, task dependent examples for each source language. Statistical methods and semantic Interlingua can be used together by training statistical systems to segment the input utterance into domain actions and concepts.

## 6.3 C-STAR and Janus

The Consortium for Speech Translation Advanced Research (C-STAR) was founded in 1991 as a forum for research groups focusing on speech translation to collaborate and meet the challenges of multi-lingual speech translation. C-STAR-I began with 4 members and demonstrated in 1992/3 the feasibility of speech translation using a rather limited prototype for German, English, and Japanese. At the end of the second phase of C-STAR in July 1999, the consortium, now expanded to 20 partner and affiliate members, demonstrated a much more powerful joint arrangement to translate between German, English, Japanese, Italian, Korean, and French on a travel planning domain. Each member in the consortium built a speech recognition system for its own language, and provided for translation from their native language either into multiple other languages or to and from a common Interlingua, called the 'Interchange Format' (IF). To perform multi-lingual translation, several systems running at the individual sites are connected through the Internet and create a distributed translation engine. The loose form of the consortium allows each member to do its own research with very little overhead, at the same time avoiding redundant development and making optimal use of the local resources in all member countries. The resulting distributed system allows for world wide cross-language experiments that none of the partners would have been able to perform each on its own. More information on C-STAR can be found on the consortium's web page: `http://www.c-star.org`.

### 6.3.1 C-STAR-II Interchange Format (IF)

While using English words to describe concepts, the C-STAR-II Interlingua is designed to work for all six C-STAR languages. Some target language dependent requirements were found during the development of the translation systems, leading to amendments in the IF.

A C-STAR-II IF consists of five components: speaker tag, speech act, concepts, arguments and argument values. In the case of the travel domain, the speaker tags are "a:" for travel agent, and "c:" for customer. Since the IF was designed for translation in goal-driven dialogues, the speech acts represent the intent of the utterance with respect to the dialogue goal: "*I want to book the cheap room*", therefore, is a request to book the room (`request-action`), while in a different context it could be considered as simply giving information. The most common speech acts in the C-STAR-II Interlingua are: `give-information`, `request-information`, `request-action`, `greeting`, and `closing`.

The concepts are used to specify the domain-specific intent of the utterance. In the sentence "*I want to book the cheap room*", the additional concepts are "reservation", "price", and "room". Speech acts and concepts together create the domain action, in this case `request-action+reservation+room`. Specifications beyond that level are made by argument-value pairs, in our example "price=cheap" and "who=I". The total IF representation for this utterance would therefore be:

```
c:request-action+reservation+room(who=I, price=cheap)
```

Due to the distributed nature of the IF development, most of the discussions had to be done by email, favoring solutions that are simple and easy to communicate. While the current IF works very well for the travel task, substantial enhancements are planned for the next C-STAR Interlingua in order to improve portability to new tasks. Planned extensions center around the current separation between concepts and arguments and the representation of optional information.

### 6.3.2 The Janus System

The ***Janus*** speech translation system as used in the 1999 C-STAR experiment [WBG98] provides a modular platform for combining and comparing multiple translation approaches. In the default setup, the Janus Recognition Toolkit (JRTk) is used for German or English speech recognition. The SOUP parser, using manually written, modular semantic grammars, parses the recognition output. The grammar for each language is modularized into one sub-grammar per sub-domain, such as hotel reservations and booking flights. Rules for actions that are required for multiple scenarios, such as requesting names and telephone numbers, are put into a cross-domain grammar. Common components such as time expressions reside in a shared grammar that acts as a library of non-terminals accessible to all other grammars. This structure makes the grammars more consistent and easier to maintain and it facilitates porting to new domains. The parser output is mapped to the C-STAR IF by means of a Perl script to maintain a high level of re-usability of parsing grammars with respect to changes in the IF. From the IF, the system can generate English, Japanese, German, Korean, French, Italian and Spanish output. Janus also supports a multi-engine approach, that permits combination of a number of alternate approaches for translation, including an example-based approach (PANGLOSS) [NIR95] and a statistical hidden understanding model (SALT) [Mun99] to automatically extract and label utterance segments corresponding to IF speech acts, concepts and arguments.

### 6.3.3 Language Portability of Speech Translation Systems

Porting speech translation systems to new languages has become a considerable concern, when considering the large number of world languages. Adding new languages to the ***Janus*** speech translation system for any given task requires the several steps. First, a speech recognizer has to be built. It can be efficiently bootstrapped either from a recognizer for that language from a different task, or from a multi-lingual speech recognizer as described in section 1. In either case, at least 50,000 to 100,000 words of text data in the domain are required as a development database for language modeling and translation.

Since the interchange format is language independent, the next task is to write semantic grammar rules that cover likely expressions for the core part of the interchange format. These rules then have to be expanded to cover all likely ways to express every concept covered by the interchange format. This is done by analyzing user data and developing grammar rules that cover the development data in a way that is likely to generalize to unseen data, while at the same time avoiding over-generalization. This part of the grammar development requires practice and skill. In order to get reasonable flexibility large grammars have to be coded.

The effort to developing a semantic parsing grammar clearly depends on the number of concepts in the domain. For the scheduling task, the main concepts centered around suggesting a time, accepting a time and rejecting a time. This resulted in compact parsing grammars that could be developed by a single person in a few months. The travel domain with its many sub-domains is considerably more complex. Careful modularization and reuse of existing structures allow development of a grammar with reasonable coverage for this task by one person in about 12 to 18 months. To add a new output language, the only new requirement is a generation grammar for the new language. Since it is sufficient to come up with a single way to express each concept, the effort for developing generation grammars is smaller than the effort for parsing grammars. When adding Japanese output to a system that already contained a fairly large generation grammar for English, we found that starting out by manually translating the core English rules to Japanese and then refining the grammars reduces the development time considerably. However, characteristics of certain languages (noun/verb agreement, as in German) can require the use of more complex generation systems. One person could develop a grammar with moderate coverage for the travel domain in approximately 6 months.

Several dialog models have been tried within the framework of our C-STAR systems. The currently most successful model is using the a priori likelihood of dialog-acts (directly integrated into SOUP's transition probabilities), as well as the likelihood of dialog-acts given the dialog-acts of the proceeding utterance (of the other speaker).

### 6.3.4 Evaluation Procedures and Results

To get realistic data to evaluate and improve speech translation systems, user studies are required. The data from each study was first used to evaluate the system, then for error analysis, and finally for development. The subjects were also given a questionnaire on user interface issues, which was evaluated to improve HCI aspects of the system. The subjects involved in all user studies had little or no previous exposure to speech recognition or speech translation. They were seated in a moderately noisy office and asked to play the role of a traveler booking a trip to Germany or, in the case of the latest user study, to Japan. The travel agents (researchers from our group) were placed in a different office. The only means of communication between the "client" and the "agent" were by way of our speech-to-speech translation system translating from English via IF back to English, a multi-modal interface allowing for handwriting recognition and sharing web-pages, and a muted NetMeeting video-conference (no audio). During the entire duration of the user study, the subjects were observed and videotaped by a researcher. Instructions on how to best use the system and interventions in case of problems were kept to a minimum.

Sentence-based Janus MT evaluations are run as end-to-end evaluations of translation output from speech input. Bilingual graders compare the source language input and target language output for each sentence. The grades assigned are `OK`, `bad`, and `perfect`. `OK` translations contain all the information from the source language sentence with no extra misleading information. `Perfect` translations meet this criterion and are, in addition, fluent sentences in the target language.

Table 5.1 reports the results of a recent evaluation. The evaluation was conducted on a set of 132 utterances (all previously unseen by the system developers). Each utterance contains one or more sentences. The data was taken from our latest user study of subjects trying to book a trip to Japan.

| Experiment | Method | Output language | % OK + perfect | % Perfect |
|---|---|---:|---:|---:|
| 1 | Recognition only | English | 78 | 62 |
| 2 | Soup on transcription | English | 74 | 54 |
| 3 | Soup on recognition | English | 59 | 42 |
| 4 | Soup on transcription | Japanese | 77 | 59 |
| 5 | Soup on recognition | Japanese | 62 | 45 |
| 6 | Soup on transcription | German | 70 | 39 |
| 7 | Soup on recognition | German | 58 | 34 |

**Table 6.1 Translation grades for English to English, English to Japanese, and English to German using the Soup parser**

Experiment 1 in Table 5.1 shows the quality of the speech recognition output measured by the same criteria as the output of the translation engine: `OK` for retaining all relevant meaning and `Perfect` for being fluent. For about 22% of all utterances, some important change of meaning had occurred due to a recognition error in the best matching hypothesis. Preliminary experiments using word graphs rather than first best hypotheses indicate that for about half of these utterances even a small word graph contains a hypothesis of the correct meaning. Experiments 2 and 3 give the performance of the system for paraphrasing back into English from transcribed text (Experiment 2) and speech recognition output (Experiment 3). An error analysis showed that only 8% of all utterances did not get a correct translation because of speech recognition errors. Another 20% of all utterances did not get correct translations because of coverage limitations of the interchange format or grammars. Experiments 4 and 5 give the performance for English-to-Japanese translation from transcribed English input (Experiment 4) and recognized English input (Experiment 5). The slightly better result in comparison to English-to-English paraphrase reflects the subjective nature of the grading process more than the actual performance. The results for translation into German (Experiments 6 and 7) mostly reflect the extremely short development time for the German generation grammars (7 weeks at the time of the evaluation).

# 7. Conclusion

In this paper we have reviewed several strategies for the development of multilingual speech recognition and understanding systems. While most modern systems are trained on large speech databases, careful design and long development times are still required to achieve good performance in multilingual spoken language systems. In this paper, we have described several difficulties of multilinguality and offered solutions to:

- the problem of portability across languages
- the problem of foreign accented speech
- the problem of morphology, or: lexicon size and confusability
- multilingual spoken language and translation systems

# 8. Bibliography

[Aub98] **S. Auberg, N. Correa, V. Locktionov, R. Molitor, M. Rothenberg:** *The Accent Coach: An English Pronunciation Training System for Japanese Speakers*. Proc. Speech Technology in Language Learning (STiLL), 1998.

[BCG96] **J. Barnett, A. Corrada, G. Gao, L. Gillik, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin**.: *Multilingual Speech Recognition at Dragon Systems*. Proc. ICSLP, pp. 2191-2194, Philadelphia, PA 1996.

[Bee80] **L. M. Beebe**: *Myths about Interlanguage Phonology*. Interlanguage Phonology: The Acquisition of a Second Language Sound System. G. Ioup and S.H. Weinberger (ed), Cambridge, MA, 1980

[BKI97] **U. Bub, J. Köhler, and B. Imperl**: *In-Service Adaptation of Multilingual Hidden-Markov-Models.* Proc. ICASSP, pp. 1451-1454, Munich 1997.

[CC97] **A. Constantinescu, and G. Chollet**: *On Cross-Language Experiments and Data-Driven Units for ALISP.* Proc. ASRU, pp. 606-613, St. Barbara, CA 1997.

[CDG97] **P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward**: *Towards a Universal Speech Recognizer for Multiple Languages*. Proc. ASRU, pp. 591-598, St. Barbara CA, 1997.

[Cut97] **A. Cutler**: *The comparative perspective on spoken-language processing*. Speech Communication 21, pp. 3-15, 1997.

[DAK95] **C. Dugast, X. Aubert, and R. Kneser**: *The Philips Large-Vocabulary Recognition System for American English, French, and German.* Proc. Eurospeech, pp. 197-200, Madrid, 1995.

[FR97] **M. Finke and I. Rogina***: Wide Context Acoustic Modeling in Read vs. Spontaneous Speech.* Proc. ICASSP, pp. 1743-1746, Munich, Germany, 1997.

[FKW98] **T. Fukada, D. Koll, A. Waibel, and K. Tanigaki***: Probabilistic Dialogue Act Extraction for Concept Based Multilingual Translation Systems*. Proc. ICSLP, Sydney, Australia, 1998.

[GW98] **M. Gavalda and A. Waibel:** *Growing Semantic Grammars*. Proc. COLING/ACL, Montral, Quebec, Canada, 1998.

[Geu98a] **P. Geutner, M. Finke, P. Scheytt, A. Waibel and H. Wactlar***: Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation*. Proc DARPA

Workshop on Broadcast News Transcription and Understanding, Lansdowne, Virginia, February 1998.

[Geu98b] **P. Geutner, M. Finke, and P. Scheytt***: Adaptive Vocabularies for Transcribing Multilingual Broadcast News*. Proc ICASSP, Seattle, Washington, May 1998.

[Geu98c] **P. Geutner, M. Finke, and A. Waibel***: Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News*. Proc ICSLP, Sydney, Australia, December 1998.

[Geu99a] **P. Geutner**: *Adaptive Vocabularies in Large Vocabulary Conversational Speech Recognition*. PhD Thesis, University of Karlsruhe, Germany, February 1999.

[Geu99b] **P. Geutner, M. Finke, and A. Waibel***: Selection Criteria for Hypothesis Driven Lexical Adaptation*. Proc ICASSP, Phoenix, Arizona, March 1999.

[GFG95] **J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue***: Multi-lingual Spoken Language Understanding in the MIT Voyager System.* Speech Communication (17), pp. 1-18, 1995.

[GG97] **S. Gokcen and J. Gokcen**: *A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition.* Proc. ASRU, pp. 599-603, St. Barbara, CA 1997.

[HW98] **J. J. Humphries and P. C. Woodland***: The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training*. Proc. ICASSP, Seattle, 1998.

[Kaw99] **G. Kawai**: *Spoken Language Processing Applied To Nonnative Language Pronunciation Learning.* PhD thesis, University of Tokyo, 1999.

[Koe98] **J. Köhler**: *Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks*. Proc. ICASSP, pp. 417-420, Seattle, 1998.

[LAG95] **L. Lamel, M. Adda-Decker, and J.L. Gauvain**: *Issues in Large Vocabulary Multilingual Speech Recognition.* Proc. Eurospeech, pp. 185-189, Madrid, 1995.

[MB99] **L. Mayfield Tomokiyo and S. Burger**: *Eliciting Natural Speech from Non-Native Users: Collecting Speech Data for LVCSR.* Proc. ACL workshop in Computer-Mediated Language Assessment and Evaluation in Natural Language Processing, College Park, Maryland, 1999.

[Mun99] **M. Munk**: *Shallow Statistical Parsing For Machine Translation*, Diploma Thesis, Carnegie Mellon University, May 1999.

[Nir95] **S. Nirenburg***: The Pangloss Mark III Machine Translation System.* A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT, April 1995. Issued as CMU tech report CMU-CMT-95-145, 1996.

[OGB94] **K. Oflazar, E. Göçmen, C. Bozşahin***: An Outline of Turkish Morphology.* Report on Turkish Natural Language Processing Initiative Project, 1994.

[OAM92] **L. Osterholtz, C. Augustine, A. McNair , I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna**: *Testing Generality in JANUS: a Multi-lingual Speech Tranlslation System.* Proc. ICASSP, volume 1, 1992.

[RNF97] **O. Ronen, L. Neumeyer, and H. Franco**: *Automatic Detection of Mispronunciation for Language Instruction*. Proc. Eurospeech, Rhodes, 1997.

[SW98c] **T. Schultz and A. Waibel**: *Language Independent and Language Adaptive LVCSR.* Proc. ICSLP, pp. 1819-1822, Sydney 1998.

[SW97] **T. Schultz and A. Waibel**: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets.* Proc. Eurospeech, pp. 371-374, Rhodes 1997.

[SW98b] **T. Schultz and A. Waibel:** *Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages.* Proc. Specom, pp. 207-210, St. Petersburg, Russia 1998.

[SW98a] **T. Schultz and A. Waibel**: *Multilingual and Crosslingual Speech Recognition*. Proc. DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne, VA 1998.

[SW99] **T. Schultz and A. Waibel***: Language Adaptation Through Polyphone Decision Tree Specialisation.* Proc. Multilingual Interoperability Speech Technology Workshop, Leusden, The Netherlands, 1999.

[Sch97] **R. Schwarz, H.Jin, F. Kubala**: *Modeling those F-conditions - or not.* Proc. DARPA Speech Recognition Workshop, 1997

[SS96] **E. Shriberg and A. Stolcke**: *Word Predictability after Hesitations.* Proc. ICSLP, 1996.

[WKA94] **B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy**: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language*. Proc. ICASSP, pp. 237-240, Adelaide 1994.

[WY97] **S. Witt and S. Young**: *Language Learning Based on Non-Native Speech Recognition.* Proc. Eurospeech, Rhodes, 1997.

[Wit99] **S. Witt and S. Young**: *Off-line Acoustic Modeling of Non-Native Accents*, Proc. Eurospeech, Budapest, 1999.

[WBG98] **M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, and A. Waibel** : *A Modular Approach to Spoken Language Translation for Large Domains*. Proc. AMTA. 1998.

Proceedings of the C-STAR Workshop in Schwetzingen, Germany, September 1999 (will be made available under www.c-star.org)