From *Proceedings of 3rd CRIM–FORWISS workshop on Speech Understanding*, Montreal, 1996, pages 85–95.

# Language Identification in the Context of a Human Machine Dialog System

E. Nöth, S. Harbeck, H. Niemann
Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
Tel.: +49 (9131) 857888 Fax.: +49 (9131) 303811
e-mail: noeth@informatik.uni-erlangen.de

**Abstract**

We present two concepts for systems with language identification in the context of multilingual information retrieval dialogs. The first one has an explicit module for language identification. It is based on training a common codebook for all the languages and integrating over the output probabilities of language specific $n$–gram models trained over the codebook sequences. The system can decide for one language either after a predefined time interval or if the difference between the probabilities of the languages succeeds a certain threshold. This approach allows to recognize languages that the system can not process and give out a prerecorded message in that language. In the second approach, the trained recognizers of the languages to be recognized, the lexicons, and the language models are combined to one multilingual recognizer. Only allowing transitions between the words from one language, each hypothesized word chain only contains words from one language and language identification is an implicit by-product of the speech recognizer. First results for both language identification approaches are presented.

## 1 Introduction

In the past, language identification was a niche research topic and has had a "James Bond" flair. Speech research as a whole only dealt with monolingual recognition and thus most groups did not work on the subject of language identification since it was not necessary. On the other hand, if you uninvitedly want to listen to several hundred phone lines and want to only record conversations in some languages, you can of course not ask these phone users, what language they use and language identification is essential. This attitude towards language identification underwent a major change with the transition of speech research from laboratory systems to real life applications: consider an automatic speech understanding system for information retrieval over the telephone that is installed in Germany and that is intended to be used by the majority of the population. It will either have to be able to handle German with a wide variety of foreign accents or be able to handle German, Turkish, Greek, Italian, etc. or exclude guest workers as customers. Things get worse if the system is intended for travel information and foreign tourists are its potential customers.

In this paper we present our approach to language identification in the context of the multilingual and multifunctional speech understanding and dialog system SQEL (**S**poken **Q**ueries in **E**uropean **L**anguages). The system is being developed in the EC funded Copernicus project COP-1634. Partners are the Universities of Erlangen (Germany), Kosice (Slovak Republic), Ljubljana (Slovenia), and Pilsen (Czech Republic). The system is intended to handle questions about air flight (Slovenian system) and train connections (German, Slovak, and Czech system) in these four languages [1].

Basis of the system is the EVAR system, the architecture of which is based on the German SUNDIAL demonstrator (ESPRIT project P 2218) [3]. Even though major changes were made – especially in the *Linguistic Analysis* [5] and the *Dialog* module [2] – the general architecture of the SUNDIAL demonstrator was kept for the EVAR system. EVAR can handle continuously spoken German inquiries about the German IC train system over the telephone.

The rest of the paper is organized as follows: In section 2 we will explain the architecture of the national SQEL demonstrators by looking at the current EVAR system. Following this, we will motivate and introduce two different system architectures for the two versions of the integrated multilingual SQEL demonstrator. The main difference is that the first architecture (section 3.1) has an explicit language identification module, whereas in the second architecture (section 3.2), the language identification is a by–product of the speech recognition process. Following this we will explain the principle of the explicit language identification in section 3.3. In section 4 we will present preliminary results and conclude with an outlook to future work in section 5.

## 2 Architecture of the National SQEL Demonstrators

Figure 1 shows a system overview of the German SUNDIAL demonstrator as well as of the EVAR system and the intended national SQEL demonstrators. Each of the four demonstrators will handle one language and one application. Here we describe the EVAR system, since an improved version of it will be the German SQEL demonstrator and since it is the only SQEL demonstrator that is already fully functional. The main components of the system are:

- *Word Recognition:* The acoustic front end processor takes the speech signal and converts it to a sequence of recognized words. Ideally the recognized words are the same as what was actually spoken. Using state–of–the–art technology, the word recognizer module performs the steps signal processing, feature extraction and a search based on hidden Markov models (HMM). Signal processing techniques include channel adaptation and sampling of the speech signal. The well known mel–cepstral features as well as the first derivatives are calculated every 10 msec. Using a codebook of prototypes the first recognition step is a vector quantization. These features are used in a beam search operating on semicontinuous HMMs. Output of the module is the best fitting word chain. A description of the word recognition module can be found in [6, 9].

- *Linguistic Analysis:* The word string is interpreted and a semantic representation of it is produced. A UCG (unification categorial grammar) approach [4, 1] is used, to

[1] It will not be truly multifunctional in the sense that one can ask in one language questions about several applications and switch between applications during one dialog.

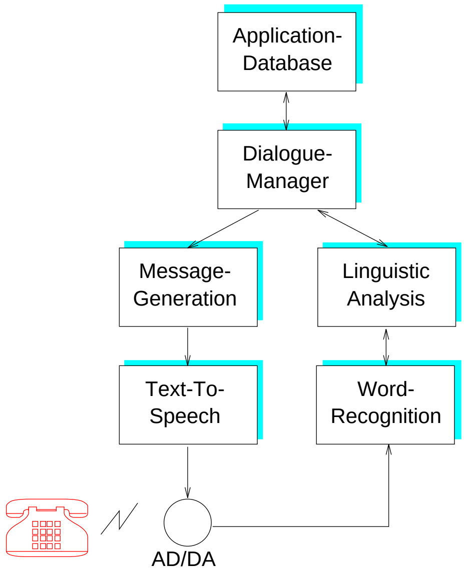


Figure 1: Architecture of the German SUNDIAL demonstrator.

model the user utterances. This approach utilizes partial descriptions, there is no need to have a complete interpretation spanning the whole utterance. Due to misrecognition or effects of spontaneous speech the system has to cope with linguistically ill-formed word sequences. The method of delivering partial interpretations is the key to enhanced robustness of the parser. A description of the linguistic analysis module can be found in [5].

- *Dialog Manager:* This module takes the semantic representation of the user utterance and performs the interpretation within the current dialog context. It decides upon the next system utterance. Specialized modules within the dialog manager for contextual interpretation, task management, dialog control and message generation are communicating via a message passing method. A description of the dialog manager can be found in [2].

- *Application Database:* The official German InterCity train timetable database is used. Ljubljana will use the Adria Airline database, Pilsen and Kosice will use the Czech and Slovak InterCity train timetable database.

- *Message Generation* and *Text–to–Speech:* In order to have a complete dialog system this module transforms the textual representation of the system utterance into sound. This sound is presented to the user. We use a simple concatenation of canned speech signals (All words that the system can say are recorded and stored as individual files).

In the next section we will describe the planned adaptation steps to build an integrated demonstrator that will be able to handle dialogs in all four languages.

# 3 Language Identification with Different Amounts of Knowledge about the Training Data

Of course, the best language identification module is a multilingual recognizer. In speech recognition this can be implemented in the following way: starting with the speech signal, run several recognizers in parallel. Each recognizer is specialized to one language, i.e. has an acoustic and a language model of one language. Then for each given point in time, one can identify the spoken language, based on the score (probability) for the best matching word chain in each of the recognizers. However, in this case the recognizers have to give comparable judgements. Also, if the system has to recognize $N$ languages then $N$ recognizers have to run in parallel, and $N$-$1$ recognizers do work that is unnecessary for the system. Another problem with this approach is that you can only recognize these $N$ languages.

Consider the situation that you want the SQEL system to be able to identify more than the four languages and react appropriately if a question is uttered in a language that can not be handled by the system. For instance, if the system identifies that an utterance was uttered in Polish, it can react with a prerecorded Polish utterance like

> The SQEL system detected a Polish utterance. Unfortunately, so far the system can only handle dialogs in Czech, German, Slovak, and Slovenian. Please ask your question again in one of these languages.

Clearly, the language identification module will not have the same quality of training data for additional languages. We might only have Polish speech samples where we know the language, but not what was said. Also, the samples might be from a very different domain, and the other necessary resources (pronunciation lexicon, stochastic language models) might not be available.

Our strategy for integrating the national demonstrators into one system is twofold:

- Build a system with explicit language identification. The only label of the training data for the language identification is the spoken language. The topic or the spoken words of the training utterance will not be known. We will describe the architecture of this system in section 3.1.
- Develop a multilingual recognizer for the four languages. In this case the same amount of labeled training data and resources (pronunciation lexicon, stochastic language models) has to be available for the languages to be identified as for the languages to be recognized. The language identification is done implicitly during the decoding of the utterance. We will describe the architecture of this system in section 3.2.

## 3.1 A System with Explicit Language Identification

Figure 2 shows a system overview of the intended final SQEL demonstrator with explicit language identification. As can be seen, the major changes affect the word recognition module and the information flow between the modules. Since we plan to use as many software modules as possible from the EVAR system, many of the internal changes can be implemented via switches for language specific resource files. To do this, the modules have to have a control channel in addition to the existing data channel. The control channel will be used

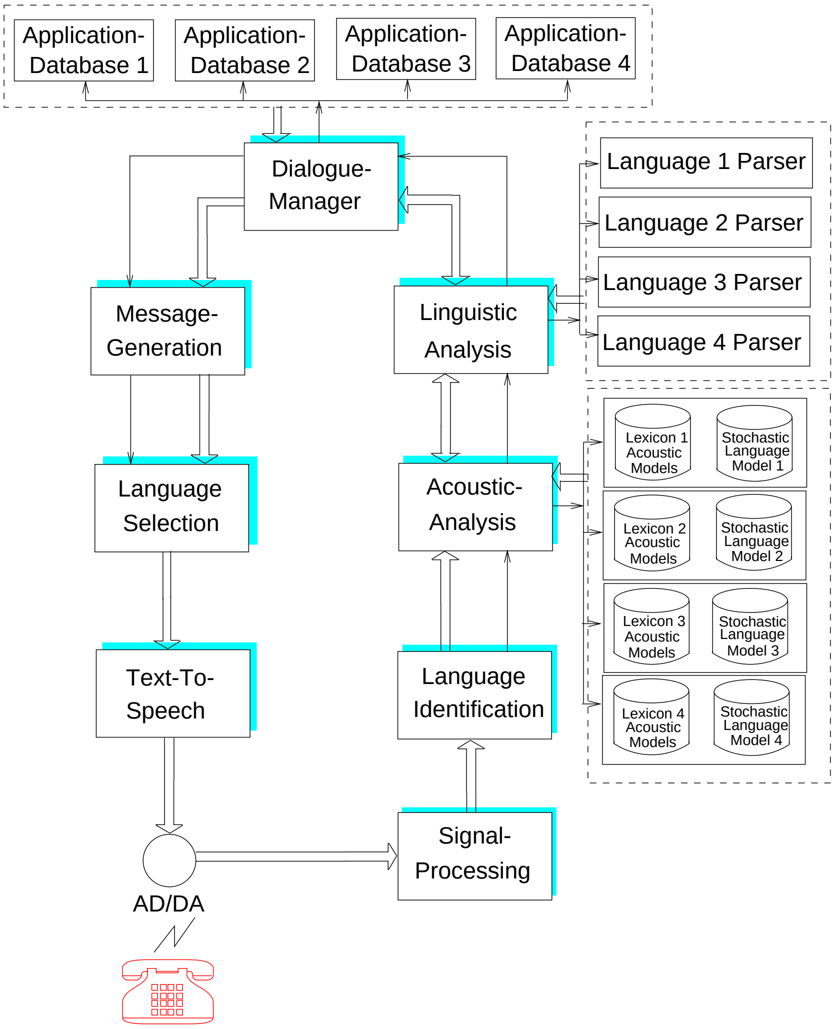


Figure 2: Architecture of the SQEL demonstrator with an explicit language identification module.

to pass messages like identity of the language and current application. The four–way arrows in Figure 2 indicate switches, the double arrows indicate data flow and the single arrows indicate control flow. The *signal processing* can be done independent of the language. The next steps — vector quantization and HMM search — need language dependent data. What is needed are language dependent codebooks, lexicons and stochastic language models. If the module has information about what language was uttered, it can simply switch to the resource files of the right language. Therefore a *language identification* module has to be added to the system that has to identify the language and pass a message to the remaining modules. The module will be activated at the beginning of the dialog. To save computation time, we use the same mel–cepstral features as the recognition module[2]. After a certain time interval a classification step between the four languages is performed. Typical time intervals

[2]Actually we only use a subset of the features, see section 3.3, but compute the whole feature set to be used in the speech recognition phase.

reported in the literature are one to five seconds of speech in order to decide between languages (for an overview of algorithms for language identification see [8, 11]). However, these results have to be verified with SQEL data, since it could well be that a longer interval is needed, if three Slavic languages (i.e. acoustically similar languages) are part of the test set.

During the analysis of further user utterances the language identification module simply passes on the extracted feature vectors and causes no delay.

Clearly, there is a tradeoff between recognition accuracy and delay time for the task of language identification: The longer the utterance, for which the sequence of feature vectors is computed, the more language specific sounds have been uttered by the caller and the better the automatic language identification will be. On the other hand, the recognition has to wait for the language identification decision, before it can start. As mentioned above, it is not clear yet, how long the utterance has to be for languages as close as Czech and Slovak, in order to be able to classify the language at an acceptable rate. This leads us to an alternative approach presented in the next section.

## 3.2 A System with Implicit Language Identification

Rather than running four recognizers in parallel, we intend to build one recognition module with all the words from all the languages. By using a stochastic bigram language model that only allows transitions between words of one language, each hypothesized word chain will only contain words of one language. Thus the language identification is done implicitly. It is implemented through a simple index lookup for the words of the best matching word chain and is done after the word recognition.

Figure 3 shows the alternative system architecture and Figure 4 shows the structure of the multilingual stochastic bigram model: the lexicon contains all the words from all four national systems. If a word from one of the languages is hypothesized, its successor has to be from the same language, since the transition probability

$$P(word_{language_i}|word_{language_j}) = 0 \quad \text{for } i \neq j. \tag{1}$$

One might argue that this approach will slow down the recognition, just like running four smaller recognizers in parallel, since we intend to quadruple the lexicon. This is however only true for the first couple words, since after this, the beam search [7] will cut off practically all the paths from the other languages. We will show in section 4 that at least in our first experiments with implicit language identification our expectation proved to be true that the increase in computational load is neglectable while the increase for running four recognizers in parallel would increase strongly: if one speaks a sentence in a foreign language into an automatic speech recognition system, the recognition time generally increases significantly, because nothing matches well and thus the dynamically adapted beam width [6, p. 120] goes up.

In the next section we will describe the language identification module that will be used in the first system architecture. The implementation of the implicit language identification for the second system is straight forward and we will not further elaborate on it.

## 3.3 Language Identification Based on Cepstral Feature Vectors

We want to build a module that only knows the identity of the training utterances, because we want to train additional languages. In order to be as efficient as possible, we want to use

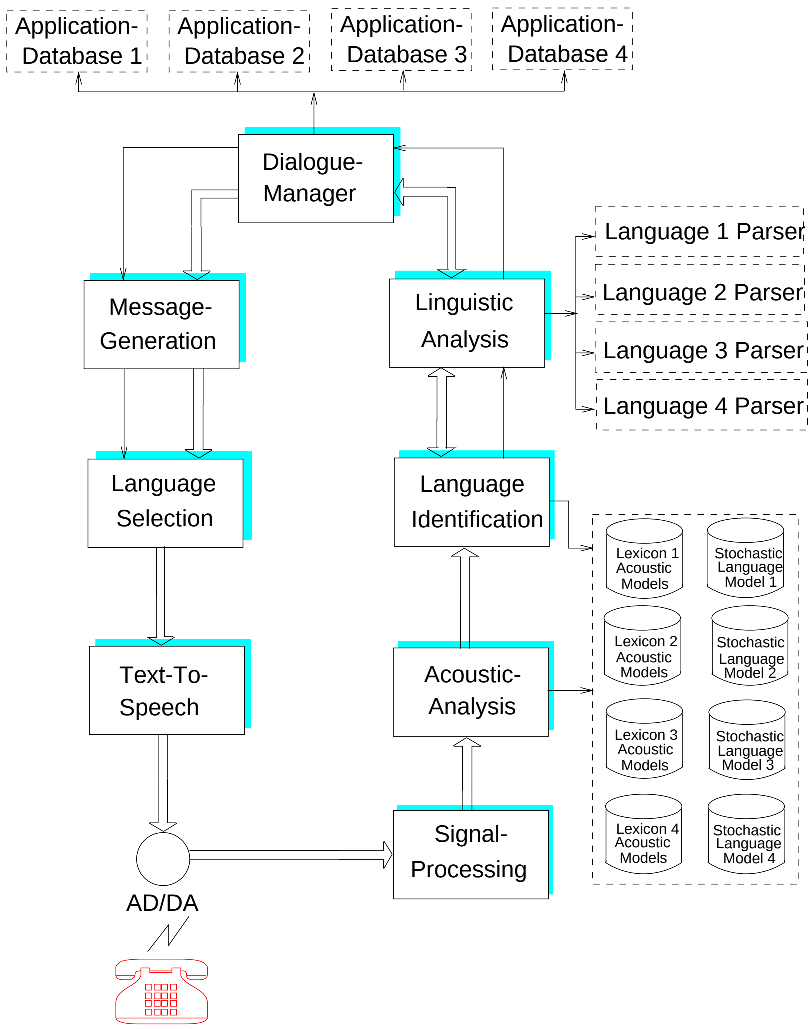


Figure 3: Architecture of the SQEL demonstrator with implicit language identification.

as many processing steps of our speech recognition system as possible. The following steps are performed:

- Extract the same mel-cepstral features and derivatives as for the recognition task. Thus after the identification no new feature extraction is necessary
- Take an appropriate subset of the features. The lower cepstral coefficients are more sound specific, i.e. language specific, whereas the higher coefficients are more speaker specific. In preliminary experiments [10] good results were achieved with using the first six mel–cepstral coefficients from three consecutive frames resulting in a feature vector of length 18.
- Train a vector quantizer with the training data from all the languages together. Output of the vector quantizer is a sequence of indices, i.e. we use a hard vector quantizer.
- Train $N$ $n$–gram language models over the symbol sequences for the $N$ languages.

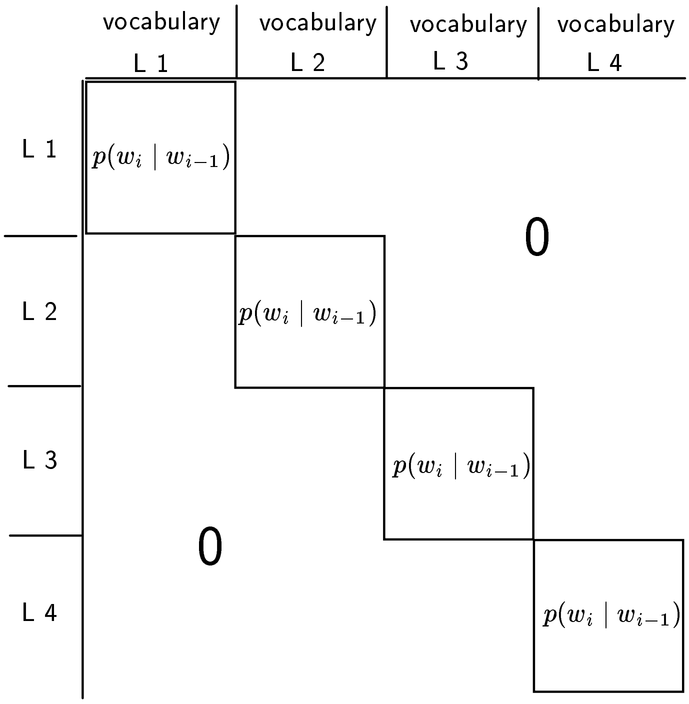


Figure 4: Structure of the bigram language model for multilingual speech recognition.

- To identify the language, calculate the sequence of vector quantizer symbols and calculate the $N$ $n$–gram probabilities in parallel. For each language sum up over the sequence of the negative logarithms of the $n$–gram probabilities. At any time the algorithm can decide for the then most likely language.

Note that for large values of $N$ a beam search can be used, i.e. after a certain interval, languages that are below a certain threshold, are discarded. Also, the module can decide for the language with the highest probability either after a fixed time interval or if the difference between the best and the second best alternative exceeds a certain threshold.

# 4 Results

In this chapter we want to present some preliminary results of explicit and implicit language identification experiments based on the SQEL corpus.

## Results for Explicit Language Identification

For the explicit language identification we tried to recognize the three Slavic SQEL languages Czech, Slovenian and Slovak. We trained the quantizer and the three $n$–gram language models with 4 hours of speech (1.7 hours from 42 Slovenian, 1.5 hours from 30 Slovak, and 1 hour from 23 Czech speakers).

Table 1 shows recognition rates of the explicit language identification module for 17 minutes from 9 independent test speakers (4 Slovenian, 3 Slovak, and 2 Czech speakers).

| rec. rate for Czech | rec. rate for Slovenian | rec. rate for Slovak |
|---|---|---|
| 91.47 % | 98.67 % | 93.65 % |

Table 1: Recognition rate for explicit language identification between three languages. Forced decision after 2 seconds (or at the end of the utterance, if it is shorter than 2 seconds).

We achieve even better results when using a transformation of the feature space with the *Linear Discriminant Analysis* (LDA) as shown in Table 2.

| rec. rate for Czech | rec. rate for Slovenian | rec. rate for Slovak |
|---|---|---|
| 96.76 % | 97.95 % | 98.67 % |

Table 2: Recognition rate for explicit language identification between three languages using LDA. Forced decision after 2 seconds (or at the end of the utterance, if it is shorter than 2 seconds).

Considering the small amount of training data, the similarity of these three languages and the time to decide, these results are very encouraging. However, it should be kept in mind, that so far we used high quality speech input and that the speech material is read speech from a restricted domain. Nevertheless, at least the restricted domain is realistic for an application in a human machine dialog system.

In [10] we obtained good results when trying to identify regional variations of German with the same module, which suggests that the proposed method can be used not only for language identification, but for speaker group adaptation within a monolingual speech recognition system.

## Results for Implicit Language Identification

The first results for our implicit language identification module concern a bilingual Slovenian/Slovak recognizer (the same Slovenian and Slovak speech data were used as for the explicit language identification). We only used these two languages, because at the time of the experiments the Czech transliterations were not available yet. In table 3 the word accuracy and the time behavior of normal monolingual and our multilingual speech recognition is shown.

These results are very encouraging. When using the multilingual speech recognition system the time for processing and the word accuracy is nearly the same as using a monolingual recognition system evaluated on sentences of the correct language only. This is due to the fact that after a short time where paths of both languages are in the beam, the multilingual recognizer "degrades" to a monolingual recognizer as explained in section 3.2.

| Recognizer | Slovenian sentences | | Slovak sentences | |
|---|---|---|---|---|
| Monolingual Slovenian | 91 % | 20 min | 0 % | 2 h |
| Monolingual Slovak | 0 % | 1 h | 86 % | 1 h |
| Multilingual | 90 % | 20 min | 85 % | 1 h |

Table 3: Word accuracy and time behavior of monolingual and multilingual speech recognition systems evaluated on Slovenian and Slovakian sentences.

Note also that when we run the monolingual recognizers with the "other" language, the computational load increases dramatically. This is due to the fact that the adaptive beam width stays wide because of a bad score for the best word chain.

# 5 Conclusions and Future Work

We presented two concepts for systems with language identification in the context of multilingual information retrieval dialogs. The first architecture is a straightforward integration of an explicit language identification module. It has the advantage of being able to recognize languages that can not be processed by the system and allows an appropriate reaction. It has the disadvantage of delaying the recognition process until the spoken language can be identified with a high accuracy. The alternative approach is to combine the monolingual recognizers to one recognizer. By forcing word transitions to stay within one language, the system identifies the language and decodes the utterance simultaneously. Since the beam search eliminates partial hypotheses with bad scores, the size of the search space approaches that of the monolingual recognizers. Thus, the delay caused by increased vocabulary size should be small. The approach utilizes the available speech data more efficiently than the explicit language identification, but can not identify additional languages.

For the explicit identification preliminary experiments with the three Slavic SQEL languages were presented that showed that the language can be identified with high accuracy after only two seconds.

For the implicit identification we presented first results with a bilingual recognizer for Slovenian and Slovak, indicating that the combined system can achieve the same recognition rates on both languages as the two monolingual recognizers. The time behavior also stayed the same as for the monolingual recognizers whereas the time behavior of the monolingual recognizers on the wrong language showed that our approach is superior to running $N$ recognizers in parallel for language identification purposes.

In the future we plan to do extensive experiments with the SQEL data (about seven hours of speech from 50 speakers for each language) with respect to accuracy and computation time for both approaches.

# References

[1] F. Andry, N. Fraser, S. McGlashan, S. Thornton, and N. Youd. Making DATR Work for

Speech: Lexicon Compilation in SUNDIAL. *Computational Linguistics*, 18(3):245–267, September 1992.

[2] W. Eckert. *Gesprochener Mensch–Maschine–Dialog*. PhD thesis, Universität Erlangen–Nürnberg, (to appear in 1996).

[3] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. G. Schukat-Talamazzini. A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In *Proc. European Conf. on Speech Communication and Technology*, pages 1871–1874, Berlin, September 1993.

[4] R. Evans and G. Gazdar. The DATR Papers: February 1990. Technical report, Cognitive Science Research Paper CSRP 139, University of Sussex, Brighton, 1990.

[5] G. Hanrieder. *Inkrementelles Parsing gesprochener Sprache mit einer linksassoziativen Unifikationsgrammatik*. PhD thesis, Universität Erlangen–Nürnberg, (to appear in 1996).

[6] T. Kuhn. *Die Erkennungsphase in einem Dialogsystem*, volume 80 of *Dissertationen zur Künstlichen Intelligenz*. Infix, St. Augustin, 1995.

[7] B. Lowerre and D.R. Reddy. The Harpy Speech Understanding System. In W.A. Lea, editor, *Trends in Speech Recognition*, pages 340–360. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.

[8] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE SIGNAL PROCESSING MAGAZINE*, pages 33 – 41, October 1994.

[9] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen*. Vieweg, Braunschweig, 1995.

[10] V. Warnke. Landessprachenklassifikation. Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, 1995.

[11] M. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4:31 – 44, January 1996.