Adaptive Vocabularies in Large Vocabulary Conversational Speech Recognition

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
der Fakultät für Informatik
der Universität Karlsruhe (Technische Hochschule)
genehmigte

Dissertation

von

Petra Geutner

aus Mannheim

Tag der mündlichen Prüfung:	12. Februar 1999
Erster Gutachter:	Prof. Dr. A. Waibel
Zweiter Gutachter:	Prof. Dr. R. Rosenfeld

Berichte aus der Informatik

Petra Geutner

Adaptive Vocabularies in Large Vocabulary Conversational Speech Recognition

Shaker Verlag
Aachen 2000

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Geutner, Petra:
Adaptive Vocabularies in Large Vocabulary Conversational Speech Recognition / Petra Geutner.
Aachen : Shaker, 2000
(Berichte aus der Informatik)
Zugl.: Karlsruhe, Univ., Diss., 1999
ISBN 3-8265-7925-9

Printed in Germany.

ISBN 3-8265-7925-9
ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 1290 • D-52013 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • eMail: info@shaker.de

Acknowledgements

I would like to thank my advisors Prof. Alex Waibel and Prof. Roni Rosenfeld for giving me the opportunity to begin and finish this thesis. I know that for both of them advising from all over the world wasn't always easy, but as multimedia people we managed :-)

Thanks to Peter Scheytt and Michael Finke, the two of them gave me back the belief that "teamwork" is not just a word .... I definitely agree with a statement Peter once made: together we had fun as well as success.

Thanks to all colleagues that contributed to this work, ranging from working together on the same task and trying to reach the same goal up to "Could you install the cyrillic font for me?". Whereas this already means Frank, he has to be mentioned again, as a lot of computer-related things probably wouldn't have run as smoothly without his support. Also, a big "Thank you" to Silke who helped wherever she could, from organizing things up to moral support and encouraging words – especially in the last three months of this work that turned out to be rather hectic.

Thanks to all my friends who were with me in a difficult time. Their love and patience definitely made things easier.

Traditionally people seem to thank their parents in the very end of acknowledgements, maybe because these are usually the two persons having had the most influence in their lives: Danke dafür, daß in unserer Familie immer die Überzeugung herrschte, es gäbe nichts, was man nicht auch erreichen könne. Nun war es eben kein Dr. der Biologie ... :-)

Abstract

Performance of large vocabulary speech recognition systems for continuous speech has steadily improved in the recent past. In spite of this ongoing development one weakness in the reliability of speech recognition systems still remains: 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.

Especially for domains like dictation systems or the automatic transcription of broadcast news the recognition dictionary cannot be constrained to a predefined vocabulary. Instead, an unlimited vocabulary is required. Within a news broadcast, for example, several different topics are treated and as time progresses, contents of the news will vary. Changes of subject may occur from one month or week to another, sometimes even within a few hours: new topics appear, like Olympics and world championships, or totally new names and places that were never mentioned before, like new presidential candidates, locations of wars or natural disasters show up in the vocabulary of the anchor speaker or foreign correspondent. As the vocabulary of a recognizer is always considered to be limited, performance depends on the correspondence of the search dictionary with the vocabulary of the topic in question. If the language to be recognized has a large number of inflections and composita, like German and Serbo-Croatian for example, the vocabulary grows even faster and the problem of new and unknown or *out-of-vocabulary* words gets worse.

This work presents various innovative methods to improve the reliability and performance of speech recognition systems for continuous speech on large vocabularies significantly and thus establish an improved basis for a wide dissemination of these systems. All methods achieve performance improvements by overcoming the limitation of the recognition dictionary to a certain size N. Even though the recognition vocabulary practically is still considered finite, the methods presented here virtually allow for a much larger vocabulary by dynamically adapting the dictionary to the speech data to be recognized.

Based on the idea of vocabulary adaptation, a multipass strategy called **H**ypothesis **D**riven **L**exical **A**daptation (HDLA) is developed where two recognition runs are performed. The first run generates a word lattice[1] which is used to create an utterance-specific word list. Starting from this word list, the recognition dictionary for the second recognition run is adapted to the utterance to be recognized.

When dealing with out-of-vocabulary words many misrecognitions are no "real" errors, but the word simply could not be recognized at all because it was not included in the recognition dictionary. Normally one or more acoustically similar words are recognized instead. Based on this observation and a rapid vocabulary growth due to a large number of inflections the major sources of error are: first, often only the word ending is wrong, but the stem was recognized correctly. Second, frequently the erroneously hypothesized word is phonetically very similar to the correct one. Both observations triggered the development of the following adaptation methods, where different selection criteria are used to choose the dictionary for the second recognition run:

1. Usage of linguistic knowledge about the morphology of a language.
2. Usage of grapheme similarity between word pairs.
3. Usage of phonetic similarity between word pairs.

[1] The term "word lattice" will be explained in chapter 6.

4. Usage of phonetic similarity between word pairs based on an artificially created fallback lexicon.

5. Usage of world-wide-web-retrieval of related texts.

All of these criteria are used for selecting new vocabulary entries from a very large fallback lexicon which is either created by taking words from the largest available database or based on linguistic knowledge. Applied to test sets of German and Serbo-Croatian broadcast news shows, HDLA yields a significant reduction in the rate of out-of-vocabulary words and results in an improvement of recognition performance. Using linguistic knowledge about morphology and inflection endings for the dynamically adapted dictionary the achieved reduction of the out-of-vocabulary rate is 45%, using phonetic similarity even up to 54%. The decrease of out-of-vocabulary words in the dictionary is reflected by an increase in recognition performance of up to 14%.

Whereas the usage of morphology to reduce the rate of out-of-vocabulary words and improve recognition performance requires linguistic knowledge about a language, the second and third method offer the distinct advantage of being easily applicable to any new language without the need of expert knowledge on it. However, all methods considering knowledge about morphology or using grapheme respectively phonetic similarity must fail if the out-of-vocabulary words are names or places (*named entities*) which are neither similar in stem nor phonetics to already known words. To address this problem the already proposed methods have been enhanced by a fifth possibility of vocabulary adaptation: beside morphologically or phonetically similar words also totally new and unknown words from the world-wide-web that refer to actual events of the day are incorporated into the dictionary of the second recognition run. Thereto the recognized hypothesis of the first run is used as input to search engines of the world-wide-web to retrieve texts dealing with the same topic. Likewise, simple retrieval of texts of a specific date can be carried out as these are most likely to contain the desired words. The expectation is to find words that did not even appear in the very huge database already collected up to that point. Filling up the dictionary with these words to a certain percentage guarantees that new words referring to current events will be considered that cannot be found by any of the other methods. By combining morphology, acoustic confusability and information about the change of topics over time, the main causes for out-of-vocabulary words can be omitted: new inflections of already known words, composita and named entities.

Zusammenfassung

Die Leistung von Systemen zur Erkennung natürlicher Sprache konnte in den letzten Jahren immer weiter verbessert werden. Dabei bleibt trotz stetiger Weiterentwicklung insbesondere bei der Erkennung von kontinuierlicher Sprache auf großen Wortschätzen ein gravierender Schwachpunkt in der Zuverlässigkeit von Spracherkennungssystemen bestehen: die Güte eines Erkenners wird wesentlich durch die Übereinstimmung des verwendeten Erkennungslexikons und dem tatsächlich zu erkennenden Vokabular beeinflußt. Sind viele zu erkennende Wörter nicht im Lexikon vorhanden, kommt es zu einer großen Anzahl von Fehlerkennungen.

Dieses Phänomen ist insbesondere in Domänen, wie z.B. Diktiersystemen oder dem automatischen Transkribieren von Nachrichtensendungen zu beobachten, wo der Wortschatz des Systems nicht mehr auf ein bestimmtes Vokabular eingeschränkt werden kann, sondern quasi unbeschränkt sein muß. Innerhalb einer Nachrichtensendung werden nicht nur viele verschiedene Themengebiete behandelt, sondern über einen Zeitraum von mehreren Monaten, Wochen oder sogar schon Stunden ändern sich auch die Inhalte der Tagesnachrichten. Dabei kommen neue Themengebiete hinzu, wie z.B. Olympiaden und Weltmeisterschaften, oder neue noch nie zuvor genannte Namen und Orte, wie z.B. Kandidaten für Präsidentschaftswahlen, Kriegsschauplätze und Orte von Naturkatastrophen, tauchen im Vokabular des Nachrichtensprechers bzw. Auslandskorrespondenten auf. Da der Wortschatz eines Spracherkenners jedoch als endlich angenommen wird, hängt die Erkennungsleistung auch wesentlich von der Übereinstimmung des im Suchprozeß verwendeten Lexikons mit dem Vokabular des jeweiligen Themengebiets ab. Ist die zu erkennende Sprache zusätzlich durch eine große Anzahl möglicher Inflektionen und Komposita geprägt, wie es z.B. im Deutschen oder Serbokroatischen der Fall ist, beobachtet man ein noch schnelleres Vokabularwachstum und das Problem einer großen Menge neuer und unbekannter Wörter wird noch verschärft.

Im Rahmen dieser Arbeit werden verschiedene innovative Methoden vorgestellt, die die Zuverlässigkeit und Leistungsfähigkeit von Systemen zur Erkennung natürlicher Sprache auf großen Wortschätzen maßgeblich erhöhen und somit eine verbesserte Basis für eine breite Anwendung dieser Spracherkennungssysteme schaffen. Allen Methoden gemeinsam ist, daß sie die Beschränkung des tatsächlichen Wortschatzes des Erkenners auf eine bestimmte Größe N aufheben. Trotz nomineller Beschränkung lassen sie virtuell ein viel größeres Vokabular zu, indem sie jeweils das Erkennungsvokabular an die zu erkennende Äußerung dynamisch anpassen.

Aufbauend auf dieser Idee wurde ein mehrstufiges Verfahren namens **H**ypothesis **D**riven **L**exical **A**daptation (HDLA) entwickelt, bei dem innerhalb des Erkenners zwei Erkennungsläufe durchgeführt werden. Im ersten Erkennungslauf wird ein sogenannter Worthypothesengraph[2] erzeugt. Die sich aus diesem Graphen ergebende Wortliste wird dann benutzt, um nach verschiedenen Kriterien das Wörterbuch des Erkenners zu verändern und für einen zweiten Erkennungslauf an die zu erkennende Äußerung anzupassen.

Bei einer hohen Anzahl unbekannter Wörter ist die Ursache vieler Fehlerkennungen kein „echter" Erkennungsfehler, sondern das richtige Wort kann vielmehr nicht erkannt werden, weil es nicht im Erkennungslexikon enthalten ist. In der Regel werden an der Stelle des unbekannten Wortes ein oder mehrere akustisch ähnliche Wörter erkannt. Zusammen mit der Beobachtung rapiden Vokabularwachstums aufgrund einer hohen Anzahl von Inflektionsendungen treten daher vor allem folgende Fehlerquellen bei der Erkennung auf: zum einen wird häufig nur die Endung eines Wortes falsch, der Wortstamm aber richtig erkannt. Zum anderen handelt es sich bei dem erkannten Wort sehr oft um ein dem korrekten Wort

[2]Der Begriff „Worthypothesengraph" wird in Kapitel 6 näher erklärt.

phonetisch ähnliches. Beide Beobachtungen wurden für die Adaption des Vokabulars ausgenutzt, und folgende Kriterien zur Auswahl des im zweiten Erkennungslauf verwendeten Vokabulars verwendet:

1. Verwendung linguistischen Wissens um die Morphologie einer Sprache.
2. Verwendung von Graphem-Ähnlichkeit zwischen Wortpaaren.
3. Verwendung phonetischer Ähnlichkeit zwischen Wortpaaren.
4. Verwendung phonetischer Ähnlichkeit zwischen Wortpaaren unter Berücksichtigung eines künstlich erstellten Hintergrundlexikons.
5. Verwendung thematisch ähnlicher Texte des World-Wide-Web.

Jede dieser Methoden wird zur Auswahl neuer Vokabulareinträge aus einem sehr großen Hintergrundlexikon benutzt, das entweder aus der größten zur Verfügung stehenden Datenbasis oder basierend auf linguistischem Wissen künstlich erstellt wird. Angewandt auf Testmengen deutscher und serbokroatischer Nachrichtensendungen führen alle HDLA-Verfahren zu einer signifikanten Reduktion der Anzahl unbekannter Wörter und einer daraus resultierenden Verbesserung der Erkennungsleistung. Benutzt man linguistisches Wissen über Morphologie und Inflektionsendungen zur Erzeugung des dynamisch adaptierten Wörterbuches, kann eine Reduktion der Anzahl unbekannter Wörter um 45% erreicht werden, bei Ausnutzen der phonetischen Ähnlichkeiten sogar von 54%. Die Verringerung der Anzahl unbekannter Wörter im Vokabular zeigt sich im Erkennungsprozeß durch eine Steigerung der Erkennungsleistung um bis zu 14%.

Während das Verwenden von Morphologie zur Reduktion der Anzahl unbekannter Wörter und Verbesserung der Erkennungsleistung linguistisches Wissen um eine Sprache voraussetzt, können die zweite und dritte Methode einfach und schnell auf jede neue Sprache übertragen und angewendet werden, ohne Expertenwissen über diese zu benötigen. Alle Methoden, die Wissen um Morphologie oder Graphem- bzw. phonetische Ähnlichkeiten verwenden, müssen jedoch zwangsläufig scheitern, wenn es sich bei neuen Worten um Namen oder Orte handelt, die weder in Wortstamm noch Phonetik Ähnlichkeit zu bereits bekannten Wörtern haben. Zur Lösung dieser Problematik wurden die bereits vorgestellten Verfahren um eine fünfte Möglichkeit der Vokabular-Adaption ergänzt: neben morphologisch bzw. phonetisch ähnlichen Wörtern werden vom World-Wide-Web völlig neue und unbekannte Wörter, die sich auf die aktuellen Ereignisse eines Tages beziehen, mit ins Vokabular des zweiten Erkennungslaufes aufgenommen. Dazu wird die erkannte Worthypothese des ersten Erkennungslaufes als Eingabe in Suchmaschinen des World-Wide-Web benutzt, um Texte über dieselben Themengebiete zu suchen. Eine weitere Möglichkeit ist, ausschließlich nach Texten eines bestimmten Tages zu suchen, da diese die gesuchten Wörter mit der höchsten Wahrscheinlichkeit enthalten. Man erwartet, auf diese Weise Wörter zu finden, die bis dahin auch in bereits vorher gesammelten riesigen Datenbasen noch nie gesehen wurden. Durch Auffüllen eines bestimmten Anteils des Lexikons mit solchen Wörtern sollen neu auftauchende Worte berücksichtigt werden, die sich auf aktuelle Ereignisse beziehen und mit keiner der anderen Möglichkeiten zu finden sind. Durch die Kombination von Morphologie, akustischer Verwechselbarkeit und Information über den Wandel von Themengebieten über die Zeit sollen die Hauptursachen von unbekannten Wörtern vermieden werden: neue Inflektionen bereits bekannter Wörter, Komposita und Namen bzw. Orte.

Contents

Chapter 1

Introduction

Research in the field of speech recognition already started in the late fifties. Ever since various problems and issues have been studied, and significant progress has been made concerning the performance and usability of systems for transcribing speech input. The following chapter gives an introduction into the development of speech recognition systems from early systems on a fairly constrained vocabulary to present-day systems for large vocabularies. Special emphasis is put on the task of transcribing broadcast news. The difficulties inherent in doing large vocabulary speech recognition are summarized, especially the problem of words unknown to the speech recognizer. Classical solutions pursued so far are briefly mentioned, and the main ideas of this thesis are sketched. An outline of this work concludes the chapter.

1.1 Speech Recognition: An Overview

The problem of speech recognition has been actively studied since the 1950s [Waibel & Lee 1990] when first pure digitalized speech signals were processed. Later on, preprocessing steps were performed on the signal, and the frequency domain instead of the time domain was used as feature space, analyzing spectograms instead of the pure signals to distinguish between digits [Davis et al. 1952] or act as a phonetic typewriter [Olson & Belar 1956].

Following this approach, one of the very first steps that were made in the area of speech research were experiments on the recognition of vowels [Forgie & Forgie 1959]. Over the years speech recognition systems were able to recognize larger speech units, like for example words [Itakura 1975]. Still, these early systems [Reddy 1976] were only able to handle a very limited number of about 100 to 200 vocabulary entries, such as the digits from 0 to 9, the letters of the alphabet or certain keywords [Rabiner & Levinson 1981]. They were mainly used for machine control and most of them had to be trained by

their potential users through speech samples. Therefore they were also called *speaker-dependent.*

During the 1980's the first *speaker-independent* systems for isolated letter [Cole et al. 1983] or word recognition of up to 130 dictionary entries evolved [Rabiner et al. 1979] [Wilpon et al. 1982] [Furui 1986]. Also, the vocabulary size increased to systems with dictionaries that contained several 1,000 words [Averbuch et al. 1985]. However, speakers were still forced to pause in between words when uttering sentences or sequences of words, as only recognition of isolated words or digits was possible. As time progressed, more and more systems evolved that were able to also handle continuous speech. An example are dictation systems where now not only isolated words could be dictated, but the user was able to utter continuous sentences in a natural way without artificially pausing in between words. One of the first high-performance speaker-independent speech recognition systems for continuous speech on a vocabulary of 1,000 words was the SPHINX system [Lee 1988] [Lee 1989]. It introduced a number of new techniques that lead to a recognition performance comparable to the best speaker-dependent systems without the need to train the system for each speaker individually.

A further development took place when applications of speech recognition systems were extended from continuous dictation to the recognition of conversational or spontaneous speech [Godfrey et al. 1992] as it is for example used in human-to-human conversations [Wahlster 1993] [Suhm et al. 1995]. Spontaneous speech is mainly characterized by the occurrence of ungrammatical sentences and disfluencies, like hesitations, stuttering, false starts within a sentence, breathing noises, lip smacks, laughter and similar features. All of these phenomena aggravate the quality of recognition [Liu et al. 1998]. Thus, in order to guarantee acceptable performance, systems handling spontaneous speech had to be limited to a certain domain.

The size of manageable recognition vocabularies has also undergone major developments throughout the last years. Currently commercially available dictation systems have already reacted to the requests of their customers for quite some time and are easily applicable to various domains and unrestricted vocabularies. Whereas they allow unrestricted speech input by being able to dynamically extend the recognition dictionary [Baker 1993], it is only recently that even for conversational speech recognition systems researchers try to overcome the limitation to a specific domain or vocabulary. Thus, present-day state-of-the-art dictation systems have already been extended to be able to handle almost unrestricted vocabularies, while this is still uncommon for systems transcribing spontaneous speech input. Some of the problems that are inherent in this development to large vocabulary systems for unrestricted natural language input, such as a fast vocabulary growth and a large number

of new words, are the subject of this thesis.

1.2 Automatic Transcription of Broadcast News

One variation of large vocabulary speech recognition systems for spontaneous speech are systems designed for the automatic transcription of broadcast news. Since 1996 evaluations on this task are organized by the Defense Advanced Research Projects Agency (DARPA) every year. Automatic transcription of broadcast news shows is without doubt one of the most challenging problems to date. Compared to large vocabulary dictation systems, e.g. read newspaper texts like articles from the English newspaper "Wall Street Journal" (WSJ) [Paul & Baker 1992], automatic transcription of broadcast news is further impeded by the following additional problems:

- spontaneous speech effects
- dynamic change between anchor speaker and foreign correspondent
- overlapping of noise and music
- unsegmented speech
- telephone quality channel

At first glance this task also seems to be limited to a certain domain, but when considering the varied contents of news broadcasts this impression quickly vanishes. News topics can vary from political news to reports on wars or natural disasters. Some television stations even interrupt their news transmissions for advertising, so that the latest sports results are broadcasted in the same way as advertisements. As a consequence, the vocabulary of a broadcast news recognizer cannot be limited to a predefined set of words of a certain size, but virtually has to be able to recognize everything that has been transmitted during a broadcast news show.

Whereas this does not pose a problem for some languages, others show certain characteristics where even the usage of very large recognition dictionaries is not able to fulfill this demand. However, especially for tasks like transcribing broadcast news the usage of unrestricted vocabularies is required to keep the number of words unknown to the recognizer to an acceptable size. Research within this work focuses on solutions to provide the required unrestricted vocabularies for the recognition process of spontaneous speech input when automatically transcribing broadcast news shows.

1.3 The Problem

When trying to recognize speech on limited tasks only, usually a task-specific vocabulary can be defined. This predefined vocabulary is used as a static dictionary during the recognition process. When turning to conversational or spontaneous speech, it is impossible to restrict potential users to a fixed set of words or phrases. This applies to dictation tasks where the user of a dictation system is not willing to limit its use to certain topics or idioms, as well as to tasks like e.g. recognizing conversational speech or transcribing broadcast news shows.

As a consequence the unrestricted natural language input to these speech recognition systems will always contain a large number of words that are not covered by the recognition dictionary. Of course 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. Every word that is not included in the dictionary cannot be recognized correctly and usually even triggers one or more additional errors around it. A high percentage of new unknown words causes a large number of misrecognition errors, thereby significantly worsening recognition performance.

Two main reasons can be identified for a large number of words unknown to the recognition dictionary, also referred to as *out-of-vocabulary* words:

1. The usage of languages that show a fast vocabulary growth.

2. Speech recognition in domains that cannot be restricted to a limited and fixed vocabulary.

For a language like English, where the vocabulary growth is very slow (see also chapter 3 for details on several different languages), a recognition dictionary of 64,000 words is enough to guarantee an out-of-vocabulary rate of less than 1% when transcribing broadcast news shows [Gauvain et al. 1997a]. Thus, the first reason of out-of-vocabulary words predominantly applies to highly inflected languages like e.g. German and Serbo-Croatian, where the out-of-vocabulary rates run up to 4.4% and 7.8% respectively for a lexicon of comparable size. To achieve out-of-vocabulary rates of less than 1%, a vocabulary size of 200,000 to more than 500,000 words would be necessary. In both languages every verb has a variety of different conjugation endings. Every noun can have several different declension endings, where adjectives usually also follow the pattern of the affiliated noun. Moreover, some languages, like for example German, allow to form composite forms out of several independent nouns, therewith also contributing to an exceptionally high vocabulary growth. Thus, due to the large number of inflection endings and composita,

some languages have an unusually high out-of-vocabulary rate, aggravating the speech recognition process in terms of quality.

The impracticality of using a limited vocabulary concerns a variety of domains, such as conversational speech, dictation of newspaper articles and the transcription of broadcast news shows. All of these tasks are extremely difficult to be restricted to a predefined vocabulary. They cannot be constrained to a certain topic, but subjects of interest will vary. Topic shifts may occur very rapidly, especially when dealing with current news topics. Already within the news broadcast of only one day several different topics are dealt with over the time. As time progresses, contents of the news shows will constantly change: reports on new events will turn up, such as stories on Olympic Games or world championships. Totally new places like location of wars or natural disasters that were never mentioned before will appear. Also, names that are unknown so far, like e.g. a new presidential candidate, might show up in the vocabulary of the anchor speaker or foreign correspondent.

Both facts introduced above are essential reasons for a rapid vocabulary growth with no saturation on a certain task. They are also responsible for the resulting high out-of-vocabulary rates that in most cases significantly influence recognition performance. To limit this degradation in performance to a minimum, a high correspondence of the vocabulary used as search dictionary within the recognition process and the utterance to be recognized is required.

1.4 Importance of the Out-of-Vocabulary Problem for Speech Recognition

In the course of developing speech recognition technology over the past decades the evaluation paradigm used to be the word error rate. A fixed dictionary size was not considered a weakness of speech recognition systems, as long as the number of words in the dictionary ensured a reasonably good coverage of the unseen corpora tested on. Errors due to out-of-vocabulary words were considered insignificant for large vocabulary recognizers, like for example English broadcast news systems, where the out-of-vocabulary rate is less than 1% for vocabularies of 64,000 words.

As speech technology starts hitting the market, the focus changed quite a bit and other means of evaluating the ease of use and the performance of speech recognition systems are applied. The satisfaction of respective users of commercially available systems comes to the fore. Customers of dictation systems or systems offering information services like train or flight timetables will be frustrated very quickly if a large share of their speech input is recognized erroneously. To them a misrecognition error is no different whether the poor

behaviour of the system is due to an unsatisfactory recognition accuracy in general, or the word could not be recognized by the dictation engine because it was not included in the recognition dictionary, thus constituting an out-of-vocabulary word.

Also, recognition performance so far has always been measured in word error rate. One effect of an improved performance is therefore a higher percentage of words that are recognized correctly. However, evaluation criteria, especially for the task of transcribing broadcast news, are shifting more and more towards information retrieval tasks [Hauptmann et al. 1998]. New metrics are being defined that go beyond extracting words from speech only and aim at providing a standard understanding evaluation measure [Boros et al. 1996]. More importance is attached to correctly recognize important keywords that can be used for information retrieval on large multimedia databases than to simply count the number of correctly recognized words. Relevant keywords are words that can be used for named entity extraction [Kubala et al. 1998b] or topic classification [Wayne 1998] [Allan et al. 1998]. The term *named entity* refers to words such as proper names and places. As a consequence, special emphasis is put on the recognition of these named entities and the recognition of content words, mostly nouns. Inherent in the continual occurrence of new topics within news broadcasts, these words constitute a major part of out-of-vocabulary words when transcribing broadcast news shows.

In addition, the application of other languages than English for the recognition process also creates a new problem for the quality of speech recognition systems: whereas English, even for spontaneous speech tasks like transcribing broadcast news, shows a percentage of less than 1% of unknown words, other languages like German and Serbo-Croatian for example, show out-of-vocabulary rates of 4.4% to 8.7%. High out-of-vocabulary rates like this of course degrade the recognition performance of a speech recognizer. As speech recognitions systems are used for an increasing variety of tasks and are also applied to a growing number of different languages, more and more importance gets attached to the problem of restricting the number of words unknown to the recognizer to a minimum.

Considering innovative user requirements, the shift in evaluation priorities as well as the wide dissemination of speech technology, the problem of being able to recognize new and unknown words becomes more significant than it used to be. In summary the following three criteria contribute to the justification for an improved handling of out-of-vocabulary words:

- User satisfaction

  What used to be an insignificant source of errors feels very different to the casual user of a speech recognition product. User satisfaction with

the product will quickly degrade if too many words are unknown and as a consequence are not recognized correctly. Even though most products try to make up for this deficiency by allowing for user-specific dictionaries, an improved handling of new words is necessary.

- Information retrieval and understanding

  Approximately 15% of out-of-vocabulary words are named entities and thus crucial for retrieval and understanding. Especially when dealing with automatic transcription of broadcast news it is the nature of the application to have new words and names showing up that the recognizer needs to account for.

- Use of multiple languages

  So far most of the research effort in developing recognition technology has been spent into English. Other languages show much higher out-of-vocabulary rates due to a complicated inflectional structure and the resulting rapid vocabulary growth. Increasing usage and application of those languages leads to the need of an improved treatment of out-of-vocabulary words to be able to guarantee acceptable recognition performance.

1.5 Related Work

Up to the present, the dictionary of a speech recognition system usually has been limited to a certain size N. For the majority of tasks and domains so far also the vocabulary used throughout the recognition process has been fixed to a predefined set of words. As more and more applications of speech recognition systems require unrestricted natural language input, the recognition dictionary of a speech recognizer would have to be unlimited to allow for arbitrary speech input. In order to alleviate the effect of the inherent vocabulary growth, but simultaneously restrict the appearance of words unknown to a recognizer to a minimum, four approaches can be pursued:

1. Increase of the size of the recognition dictionary.

   The most straightforward method of decreasing the out-of-vocabulary rate of a speech recognizer of course would be to simply extend the number of entries in the recognition dictionary to whatever size is needed for the respective speech application problem. However, huge recognition dictionaries increase the search space of a recognizer, thereby sincerely slowing down recognition speed and often also worsening recognition performance. For many existing speech recognition systems this

approach is also impractical considering their current implementation schemes. Moreover, often even very large recognition lexica are not sufficient to account for all out-of-vocabulary words and thus to restrict the number of words unknown to the speech recognizer to a minimum. Especially for highly inflected languages and tasks allowing spontaneous speech input, a static dictionary of fixed size and predefined entries will never be able to cover all potential word candidates included in a speech utterance, even if its size is very huge.

2. Decomposition of words into smaller base units.

   Using smaller units than words for the recognition process has one major advantage: the speed of vocabulary growth is limited and thus the number of words unknown to the recognizer is decreased. A new token that would have constituted a new vocabulary entry on word level might merely consist of parts already known to the dictionary when being decomposed into smaller units.

   When using other base recognition units than words, the question on how to find the size of an optimal base unit is still an unsolved problem [Geutner 1995] [Mayfield Tomokiyo & Ries 1997]. Research has been performed on phoneme-, morpheme- and syllable-based recognizers for Korean [Hwang 1997], or string patterns as base units for the language model of a Japanese speech recognition system [Ito & Kohda 1996]. For German and Serbo-Croatian, a morpheme-based approach has been used (see also chapter 5). Although these techniques limit the vocabulary growth of the respective languages, they suffer from deficiencies that normally lead to recognition results worse than the traditional word-based approach.

   The work of Deligne et al. [Deligne & Bimbot 1995] [Deligne et al. 1995] describes speech recognition on base units of variable length. Language is viewed as a stream of words where certain dependencies between words can be found. These dependencies are modeled by variable-length sequences of words. The resulting multigram approach is found to be a competitive alternative approach to the conventional n-gram models in terms of language modeling. As the joint multigram model provides a framework that can be applied to various pattern recognition problems, the application of the multigram approach to other issues as e.g. the automatic definition of speech recognition units is suggested. Possible future work might investigate the ability of the proposed model to derive acoustically motivated subword units and test the relevance of these units with respect to contextual coarticulation.

The work of [Hwang 1997] also tries to find lexical units that are more suited for speech recognition than the notion of words. On a Korean database first all words in the vocabulary are split into phonemes, syllables or morphemes. Then the pairs with the highest bigram frequencies are rejoined again. This process is repeated until a certain threshold is exceeded. Thereby the vocabulary size is adjusted to an operational point between the number of distinct subword units and the number of words. Although the percentage of out-of-vocabulary words is decreased by using smaller base units than words throughout the recognition process, no improvements concerning recognition performance are reported.

Similar results have been obtained for most languages. In most cases speech recognition experiments show that the traditional word-based approach outperforms recognizers both on morpheme and syllable level. However, improvements in recognition performance have been reported [Ries et al. 1996] when using algorithms that automatically determine word phrases or multi-words as base units.

Experiments [Geutner 1995] have shown that useful morphologically based decompositions are hard too find, but the decomposition of compound words, as found in the German language, shows promising results. Following this conclusion a number of publications are concerned with the segmentation of composite words into their individual parts.

[Spies 1995] proposes another method to handle compound words differently to other words within the language model component of a recognizer. Experiments are performed on a German database, as compound words are a special characteristic of the German language. Here, the last component of a compound word is considered to be the grammatically determining constituent. The other elements of compound words are modifiers of this constituent. Based on this fact the words are split into their individual constituents and the probability of word sequences is estimated depending on the position of the compound constituents. The implementation of this linguistically based approach within the framework of a speech recognizer is not trivial. No comparative experimental results are given, so that it is unclear if the implementation effort would be worthwhile.

[Geutner 1995] already showed that a complete morphological decomposition reduces the vocabulary size of a German speech recognition system significantly, but suffers from overgeneration of illegally inflected words due to the shortness of many inflectional and derivational prefixes and suffixes. As a consequence, the work of [Berton et al. 1996] focuses on compound word decomposition only. Composite words are split into their

individual components reducing the vocabulary size of the used speech recognition system significantly by 24%. The new base units are then used for speech recognition experiments. After the recognition process, the components of a word are rejoined again performing a lexical search on the word graph provided by the recognizer. However, the achieved recognition results are still worse than the baseline results using the larger vocabulary. This is due to the fact that not all components of a composite word could be recognized correctly by the speech recognition system and thus cannot be recomposed to a legal German word. As an alternative the decomposition of very frequent compound words is suppressed. Although the reduction in vocabulary size is smaller, recognition performance can be slightly improved.

Decomposition of compound words into their individual parts is also pursued in [Carter et al. 1996]. Within a spoken translation system from Swedish to English, compound words are decomposed and the individual parts are incorporated into the lexicon as well as the language model. After the recognition is done, they are rejoined again. The remapping on words is done on word string level only, not on word graphs which represents a very simple and easy-to-implement method but of course also is very error-prone. Slight improvements of up to 0.5% word error rate are reported concerning the recognition performance of the system and also the end-to-end performance of the whole system is improved.

The work of [Lüngen et al. 1996] concentrates on compound words that are uttered in separate parts, often divided by filled pauses. It is found that this kind of splitting mainly takes place at morphological boundaries, like e.g. "April-Hälfte". Based on these so-called split compounds three methods to model the compounds are proposed, depending on the characteristics of the segment inserted between the individual parts. The algorithm theoretically is able to rejoin incorrectly inflected hypotheses to a correct compound word. Small improvements in word accuracy measured on word graphs are reported.

Common idea of all the approaches presented above is to decrease the fast vocabulary growth of highly inflected languages and simultaneously reduce the resulting high out-of-vocabulary rates. Instead of using a word-based recognition dictionary, a lexicon consisting of subword units is used. The coverage of such a dictionary in terms of word units by subword units or concatenations thereof is significantly better than the coverage of a dictionary of words of the same size. However, the majority of recognizers built on top of these units suffer a severe performance degradation, as some hypothesized subword concatenations do not map

to legal words, still leaving word-based speech recognition systems as the best choice for most languages.

3. Special treatment of new words.

   When sticking to words as base units, the so-called *new word problem* is usually solved by trying to detect that the user has uttered an out-of-vocabulary word and identify the new word [Fetter 1998]. After obtaining a phonetic transcription for the new word, it is added to the vocabulary of the system [Asadi et al. 1990] [Asadi et al. 1991] [Kemp & Jusek 1996]. Other approaches use the same acoustic model for all new words [Boros et al. 1997].

   Analysis of out-of-vocabulary words has not been necessary for a long time, as their occurrence did not influence recognition performance significantly. However, as out-of-vocabulary words are mainly semantically important words, their correct recognition is essential for certain tasks and applications. One of the first efforts aimed at investigating the special characteristics of new words were studies on the Wall Street Journal (WSJ) corpus [Suhm et al. 1993]. The distribution of out-of-vocabulary words concerning their grammatical category were examined and word length studies were performed. No significant difference of out-of-vocabulary words and words included in the recognition dictionary concerning the number of phonemes was found, but a major proportion of out-of-vocabulary words constituted proper names as well as inflections and concatenations of already known words.

   Thorough studies can also be found in [Hetherington & Zue 1993] where four different speech corpora were investigated and studies of vocabulary growth and out-of-vocabulary rates were performed. In addition, the phonological properties of out-of-vocabulary words were examined: e.g. their length distribution concerning the number of phones or syllables in a word as well as their acoustic novelty which means the average number of new syllables per new words. The effect of new words on recognition performance are quantified in [Hetherington 1995] and methods to detect, locate and incorporate new words into a speech recognition system are presented.

   There are several different methods that have been proposed and examined to handle the appearance of unknown words properly. Most of them pursue the idea to first detect and identify an out-of-vocabulary word. After detection and classification a usable phonetic transcription is generated to be included in the recognition dictionary. The work of [Fetter et al. 1995] [Fetter et al. 1996] [Fetter 1998] for example exam-

ines two approaches to solve the problem of out-of-vocabulary words: the first idea tries to prevent the problem itself by optimizing the used vocabulary based on information extracted from other corpora and application domains. Preventing the problem means following up the idea of using morpheme-based base units instead of the traditionally used word-based ones as proposed first by [Geutner 1995] and already introduced above. By doing this the number of expected out-of-vocabulary words can be reduced and the vocabulary coverage is significantly improved. The second approach adapts the respective recognition system to the occurrence of out-of-vocabulary words by incorporating explicit knowledge of new words into the system models. To this end the acoustic models are complemented with additional garbage models for out-of-vocabulary words. After the identification of unknown words, those words that have occurred with a certain minimum frequency are incorporated into the vocabulary and a phonetic transcription is generated. This information is extracted from the word-based search. Using this by-product of recognition only little additional computation is required, whereas the results are almost as accurate as using phoneme recognizers at higher cost.

When trying to identify out-of-vocabulary words and process them in a different way than other words, a filler model that covers these unknown words during the recognition process is desirable. The acoustic modeling of such filler models has been subject of several publications, mostly in the context of word-spotting applications [Kemp & Jusek 1996] [Klemm et al. 1995]. However, beside a special acoustic modeling of new and unknown vocabulary words, considering the language model information seems similarly important for the recognition of out-of-vocabulary words. As a consequence the appropriate integration of such words into statistical language models is also a problem to be solved. A single out-of-vocabulary label for all unknown words cannot incorporate much language model information as it has to cover fundamentally different classes of out-of-vocabulary words such as fragments or proper names. Thus, [Gallwitz et al. 1996] proposes a solution that integrates information about the presence of out-of-vocabulary words into statistical language models. The approach allows both the detection of out-of-vocabulary words by the recognizer as well as the assignment of a semantic word category to each occurrence. The various categories can either be constructed manually based on linguistic knowledge or automatically. The category information is used to estimate an approximation of the out-of-vocabulary word emission probability for each word category. Experiments are conducted on the EVAR corpus, a task where spontaneous speech recognition data was collected by a spoken dialogue

system [Schukat-Talamazzini et al. 1994]. The system is able to answer inquiries about German Intercity train connections. Common to similar applications this is a spontaneous speech task where the recognition vocabulary cannot cover all occurring words and the amount of out-of-vocabulary words is significant. The new approach for the integration of out-of-vocabulary words into statistical language models yields a 6% reduction in word error rate with a lower out-of-vocabulary rate of 5%.

Especially in natural language spoken dialogue systems misrecognitions often result in possibly irreparable misunderstandings between the user and the system. Therefore it is desirable to have the system detect unknown words and inform the user about them, so that the error can be corrected. Based on this observation [Boros et al. 1997] describes a methodology to detect, classify and process out-of-vocabulary words in an automatic train timetable information system [Eckert et al. 1993]. Acoustic information as well as language model information is taken into account for classifying out-of-vocabulary words into different word categories. The same acoustic model is used for all out-of-vocabulary words, only the language model information contributes to the assignment of a category to each. For a vocabulary size of about 1,000 words and an out-of-vocabulary rate of 5% a word error rate reduction of 5% relatively could be achieved compared to not giving out-of-vocabulary words a special treatment. As, beside the detection of out-of-vocabulary words, also the categorization of these words is important, the word category is assigned correctly in 94% of all cases.

A large number of practical applications does not require perfect recognition of the speech input signal. Especially for tasks like directory assistance services, flight or train timetable information systems and other information services where open input is allowed, relevant keywords or phrases are often embedded in longer utterances than the speech recognition system expects. Hence high-quality recognition of a small subset of certain keywords is already sufficient to achieve good system performance and guarantee a high task completion rate. In such word spotting applications the problem of identifying out-of-vocabulary words is a critical issue to be solved. As most of the existing word spotting systems are based on **H***idden* **M***arkov* **M***odels* (HMMs)[1], the main difference between them is the implementation of the out-of-vocabulary words or *garbage models* [Boite et al. 1993]. Some approaches use HMMs which are trained only with non-keywords and background noise to model the complete non-keyword sounds [Lleida et al. 1993]. Other solutions de-

[1]For a definition of HMMs see chapter 2.

fine garbage models with the same models that are used for the relevant keywords [Jeanrenaud et al. 1993]. [Klemm et al. 1995] also does not require any additional training for out-of-vocabulary words as they are covered by syllable-based concatenations of the standard subword units included in the system.

4. Dynamic Expansion of the lexicon.

   Another solution to reduce the number of out-of-vocabulary words is a dynamic expansion of the recognition dictionary as proposed in this thesis. The limitation of the dictionary to a fixed size is overcome and the usage of a virtually unlimited vocabulary is allowed. This is done by a multipass strategy that adapts the used recognition dictionary dynamically to the speech segment to be recognized [Geutner et al. 1998b]. As a consequence, the recognition vocabulary practically is still considered finite, but virtually a much larger vocabulary is applied. The development of this algorithm is the focal point of this work and will be introduced in more detail in the next section.

Of course lexical optimization is a common goal of all these approaches meaning that the optimized vocabulary is expected to be as small as possible while simultaneously reducing the number of occurring out-of-vocabulary words to a minimum. The definition of an appropriate vocabulary is always a fundamental problem when designing spontaneous speech recognition systems on large vocabularies for all practical applications. The optimal vocabulary should cover as much of the future user utterances to be recognized as possible. At the same time it should contain no words not necessary for the respective task, as they may lead to recognition errors and also increase computation time with a growing vocabulary. The optimization of a suitable recognition vocabulary is topic of [Rosenfeld 1995]. The North American Business (NAB) news corpus is used to study the effect of various types and amounts of data taken from varying time periods on the quality of the derived vocabulary. The effects of an increased vocabulary size on the accuracy of a speech recognizer is studied and the results are used to pick the optimal vocabulary size. The effects of seasonality, meaning the time of year from which the data is drawn, the amount, recency and source of the training data are examined. The expectation that seasonal effects might reduce the out-of-vocabulary rate when using training data from the same or adjacent months as the test data could not be confirmed. When measuring the correlation of out-of-vocabulary words it is found that, as expected, more training data results in a lower out-of-vocabulary rate, even though the effect slows down after 30,000 to 50,000 words. Concerning the recency of the training data it is shown that similar amounts of training data taken from different time periods make a difference,

albeit slowly. The source of the training data is also important and has more effect on the percentage of unknown words than recency or seasonality. In summary the best coverage is achieved when the speech data to be recognized deals with the same topic or comes from the same domain as the text material the recognition dictionary is derived from and the largest available database is used.

1.6 Thesis Contributions

First attempts to address the problem of rapid vocabulary growth and high out-of-vocabulary rates were made by decomposing words into smaller base units. Chapter 5 gives a more detailed report on the developed techniques and the performed recognition experiments. Here, a morphology-based approach has been chosen where words are decomposed into their word stems and inflection endings. As a result the recognition vocabulary grows much slower than on word basis. The number of out-of-vocabulary words on the respective task is reduced and, through a smaller dictionary, the recognition speed is also accelerated. Nevertheless, only very small performance improvements compared to conventional word-based recognition are achieved with a very specialized decomposition that mainly considers compound words only.

Of course the simplest way to counteract a high out-of-vocabulary rate would be to extend the recognition dictionary of the recognizer to the size actually needed. For the broadcast news tasks dealt with in this work even an astronomically big vocabulary would not be able to fulfill all necessities, as there will always appear new unknown words. Also, this approach is impractical for almost all existing speech recognition systems. As a consequence, this work has investigated another possibility to dynamically extend the size of the recognition dictionary. Main idea of this approach is to still consider the recognition vocabulary finite, but virtually allow for a much larger vocabulary. This is achieved by not using a fixed dictionary for all utterances to be recognized, but dynamically exchanging the recognition dictionary depending on the actual input. By adapting the dictionary of the used speech recognizer to the speech data to be recognized, the limitation of the dictionary to a certain size N is overcome. This is done by a multipass strategy called **H**ypothesis **D**riven **L**exical **A**daptation (HDLA).

Within the HDLA framework two recognition runs are performed. The algorithm uses a word list generated by a first recognition pass to create an adapted recognition dictionary for the second run. Different selection criteria are applied for choosing the vocabulary of the adapted dictionary:

1. Linguistic knowledge about the morphology of a language.

2. Similarity measures between word pairs based on grapheme distances.

3. Similarity measures between word pairs based on phonetic distances.

4. Similarity measures between word pairs based on phonetic distances using an artificially created fallback lexicon.

5. Retrieval of related texts on the world-wide-web.

All of these criteria, except the last one, are used for selecting new vocabulary entries from a very large fallback lexicon that can either be created from a large text database or based on linguistic knowledge. The HDLA procedure presented in this work reduces the number of unknown words significantly and successfully improves recognition performance.

1.7 Outline

Research conducted within this work focuses on the problem of dynamically adapting the recognition vocabulary of a speech recognizer to the speech input to be recognized, especially within the framework of large vocabulary speech recognition systems for continuous speech like transcribing broadcast news shows.

Chapter 1 The current chapter has given a short summary on the development of speech recognition research from early speech recognition systems up to present-day systems. The problem of new unknown words, which is the core of this thesis, has been introduced and possible solutions to the out-of-vocabulary problem have been briefly mentioned.

Chapter 2 The first part of chapter 2 gives an insight into the fundamental theory and basic model assumptions inherent in the principles of speech recognition. The general course of the recognition process is described, and a summary of the desirable features of current state-of-the-art large vocabulary speech recognition concludes the section. The second part presents a short overview over research and achievements in speech recognition applications in the past and today. Possible applications like information retrieval tasks, dictation systems and the recognition of spontaneous or conversational speech are introduced. Special emphasis is put on the automatic transcription of broadcast news, as most of the experiments presented in this thesis are performed on this task.

Chapter 3 Language characteristics of the languages experimented with in this thesis are introduced in chapter 3. The chapter describes the lexical properties of the Serbo-Croatian and the German language. In comparison the distinctive features of the English language are presented. Beside possible dialects and pronunciations, the writing system and spelling, phonology and morphology of each of these languages are introduced. The chapter especially focuses on the lexical properties of the three languages in regard to noun declensions, verb conjugations, and compound words or composita. Graphs on the vocabulary growth are illustrated, and the out-of-vocabulary rates for different vocabulary sizes are shown. The chapter concludes with an overall comparison of the vocabulary growth of the three languages introduced, complemented by the curve of a Turkish database.

Chapter 4 All speech recognition experiments performed within this work have been conducted either on a Serbo-Croatian or a German database. Chapter 4 outlines the design, development and training of two speech recognition systems for transcribing Serbo-Croatian and German broadcast news. For both languages the speech and text databases, the training of the speech recognition systems and first baseline results are presented.

Chapter 5 To counteract the fast vocabulary growth of highly inflected languages and thereby also decrease the number of out-of-vocabulary words, smaller base units than words are desirable to be used for the recognition process. Chapter 5 reports on several decomposition methods applied to the German and Serbo-Croatian language. All methods are either based on linguistic knowledge about the morphology of the respective language or employ clustering techniques that are based on similarity measures. Examples of the different decompositions are given and the recognition results that were achieved when conducting speech recognition experiments on these new base units are presented.

Chapter 6 Chapter 6 takes an alternative approach: words are still considered as the semantic content bearing units of recognition. But instead of having a static dictionary of those words, the concept of a dynamic dictionary is introduced. This dictionary has the same fixed size as the static dictionary but is tailored on the fly to each specific utterance to be recognized. Thus, with each utterance having its own customized dictionary the size of the recognition dictionary is virtually unlimited. The chapter introduces a newly developed technique of dynamically adapting a recognition dictionary, the Hypothesis Driven Lexical Adaptation (HDLA) algorithm. The algorithm allows to use several different criteria to select the vocabulary of the adapted recognition

dictionary. The following chapters 7, 8, 9 and 10 give an overview over these various selection criteria that can be applied within the framework of HDLA.

Chapter 7 In chapter 7 linguistic knowledge about morphology and inflection endings is used to dynamically adapt the recognition dictionary to the utterance to be recognized. Morphological similarity is applied as selection criterion for the adaptation process. The resulting morphology-based approach for lexical adaptation is described and recognition results on Serbo-Croatian and German broadcast news data are presented.

Chapter 8 Beside the application of morphological similarity, also similarity based on acoustic or grapheme distance measures can be used as selection criterion for the dynamic adaptation of a recognition dictionary. Chapter 8 on distance-based selection criteria for HDLA is divided into three sections: the first part describes a grapheme-based distance measure that is used to choose similar words for the adapted vocabulary. The second part introduces distance measures based on phonetic distances where three different methods of calculating the phonetic distances are investigated. Finally a third approach, also based on phonetic distances, is presented where special consideration is given to the phenomenon of composita. Decreased out-of-vocabulary rates as well as recognition results are presented both for the Serbo-Croatian as well as for the German language.

Chapter 9 All techniques presented so far rely on the availability of a large fallback lexicon for a specific language to select the word entries of the adapted vocabulary from. The creation of such a lexicon requires the existence of an enormous text database for a particular language. Chapter 9 introduces a method that has no need of a large amount of text data. Instead it is based on the formulation of morphological rules for the language in question and the possibility of measuring phonetic distances between words. New words are automatically generated according to the defined language-specific rules and these morphological variations are then incorporated into an artificially created fallback lexicon. Phonetic distances between word pairs are then determined by using this artificially created lexicon.

Chapter 10 In addition to the already presented selection criteria, also information retrieval techniques can be employed. The underlying idea is to retrieve texts from the world-wide-web related to the utterance to be recognized. By processing texts dealing with the same topic relevant words for the adaptation procedure can be found. These words might be neither linguistically nor

phonetically close to hypothesized words of the first HDLA recognition pass and thus cannot be found through the methods presented so far. Chapter 10 describes two approaches: the first is based on the Okapi similarity measure, the second uses the topicality of a news show to retrieve similar texts.

Chapter 11 Finally, chapter 11 summarizes the work performed here and reviews the achieved results.

Chapter 2

Large Vocabulary Speech Recognition

Current state-of-the-art speech recognition systems have evolved from speech recognizers on limited domains and vocabularies to large vocabulary systems. The first section of this chapter gives an overview over the basic principles of speech recognition. It describes the general course of the recognition process and concludes with a summary of current state-of-the-art large vocabulary speech recognition system techniques. The section is based on a review of Large Vocabulary Continuous Speech Recognition (LVCSR) that in more detail can be found in [Young 1996]. The second part introduces possible applications like information retrieval tasks, dictation systems, and the recognition of spontaneous or conversational speech. Special focus is put on the automatic transcription of broadcast news, as most of the experiments presented in this thesis are performed on this task.

2.1 Basic Principles of Large Vocabulary Speech Recognition

Following the very first speech recognition systems for tasks on limited domains with fairly constrained vocabularies, like e.g. speech recognition front-ends for the Air Travel Information Services (ATIS) task [Hemphill et al. 1990], systems for large vocabulary speech recognition were developed. In the beginning these systems were dictation systems that were only able to handle recognition of isolated words. The user had to pause briefly between each word of a sentence. Moreover, these early systems were also speaker-dependent, requiring a certain amount of training by the user before achieving optimal performance. Technology has improved steadily ever since. As a consequence current contin-

uous speech recognition systems that transcribe read speech recorded in clean laboratory environments are able to achieve word error rates between 5 and 10%. When speaker adaptation is applied, these error rates can even be improved to a better performance after an enrollment period of 10-20 minutes. In the meantime recognition on large vocabularies has been extended to conversational or spontaneous speech, thereby also decreasing recognition performance compared to read input. The following sections introduce the basic principles of speech recognition systems (see also [Jelinek 1997] [Rabiner & Juang 1993] [Schukat-Talamazzini 1995]) that are inherent to all kinds of large vocabulary applications from dictations systems to systems for spontaneous speech input.

2.1.1 The Speech Recognition Problem

The basic principles and algorithms speech recognizers are based on have changed very little since the early days of speech recognition systems. Basically a speech waveform is converted to a sequence of acoustic vectors by a front-end signal processor. Each vector is a compact representation of a short-time speech spectrum covering a certain time period, typically 10 milliseconds.

Given the stream of acoustic vectors and starting from a pattern recognition problem, the following Bayes rule is used to turn a classification problem into a modeling problem and determine:

$$\hat{W} = \arg\max_{W} P(W|A) = \arg\max_{W} \frac{P(W)P(A|W)}{P(A)} \tag{2.1}$$

That means, in order to find the most likely sequence of words $W = w_1, w_2, ..., w_n$, the word sequence that maximizes the product of $P(W)$ and $P(A|W)$ divided by $P(A)$ has to be found. Here, $P(A)$ is the average probability of observing the acoustic signal A. As this probability is fixed and independent of the maximization problem above, the aim of a recognizer is reduced to finding the word sequence W that maximizes the product of $P(W)$ and $P(A|W)$. $P(W)$ is the a priori probability that the word string W has been uttered independent of the signal. This probability is given through the *language model* component of the speech recognizer. $P(A|W)$ represents the probability of observing the acoustic signal A given a certain word sequence W. This probability is determined by an *acoustic model*.

2.1.2 Recognition Process

During the recognition process a certain word sequence $W = w_1, w_2, ..., w_n$ is hypothesized and the language model computes its probability $P(W)$. Each word of the hypothesis is converted into a sequence of basic sounds, the phones.

This mapping is done by means of a pronunciation dictionary where each word of the vocabulary is represented by a sequence of phones. Each phone in turn is modeled through a Hidden Markov Model (HMM) [Rabiner 1989] (see section 2.1.2.2). For an utterance to be recognized, a sequence of HMMs is concatenated and a single composite model is built. Based on this model the probability of the observed sequence A given the stream of words, $P(A|W)$, is calculated. This is done over all possible sequences of words that may have been uttered, and the most likely sequence is then picked as the recognizer output.

To be able to perform such a recognition run, the following components are necessary as the building blocks of a speech recognition system:

- a preprocessing component that extracts all the necessary information in compact form from the speech waveform,
- acoustic models that, in form of HMM models, accurately represent the distributions of each sound in each context they may occur in, and
- a language model that gives reliable estimates of the probability of a word based on its history.

Both HMM models as well as the language model also have to be able to predict data never seen in the training database. And, since enumerating all possible word sequences is clearly intractable, an efficient decoding is imperative. Potential word sequences need to be explored in parallel, discarding very unlikely hypotheses as early as possible.

2.1.2.1 Preprocessing

The preprocessing module of a speech recognizer extracts all the necessary information from the speech waveform in compact form. The incoming speech signal is regarded as stationary, i.e. the spectral characteristics are assumed to be relatively constant in intervals of a few milliseconds. As a consequence, the main task of the preprocessing component is to divide input speech into blocks from which smoothed spectral estimates can be derived.

Spectral estimates are either computed through Linear Prediction or Fourier analysis [Rabiner & Schafer 1978] plus a number of additional transformations that can be applied to generate the final acoustic vectors. *Mel-Frequency Cepstral Coefficients* (MFCCs), see [Davis & Mermelstein 1980] for example, are a method where the Fourier spectrum is smoothed by integrating the spectral coefficients within triangular frequency bins arranged on a non-linear scale called the Mel-scale. By appending first and second order differentials to the basic static coefficients to incorporate dynamical information about the signal, the problem is greatly reduced.

2.1.2.2 Acoustic Model

To be able to calculate the likelihood of any vector sequence A given a word w within a speech recognition system, an acoustic model $P(A|w)$ is used. As it is impossible to learn the required probability distribution by finding enough samples of each word in the training data and collecting the statistics of the corresponding vector sequences, word sequences are decomposed into smaller base units, the phones. Each phone is represented by a Hidden Markov Model (HMM), where each HMM consists of a number of states connected through arcs with probabilities attached. HMM phone models typically have three emitting states and a simple left-to-right topology as illustrated in figure 2.1. The entry and exit states are provided to allow to join multiple models together. A composite model is formed by merging the exit state of one phone model with the entry state of another one. Several phone models can then form a word, and the concatenation of a sequence of words is able to cover a complete speech utterance.

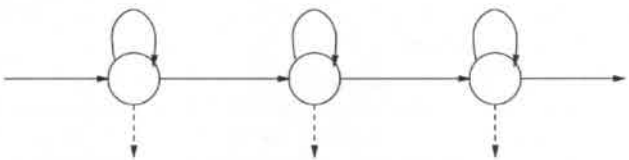

Figure 2.1: Example Architecture of a Hidden Markov Model (HMM).

A Hidden Markov Model is a finite state machine that randomly generates vector sequences. Every time unit the HMM state changes and an acoustic vector is generated with a certain probability dependent on the current state. Transition probabilities model the durational variability, the output probabilities the spectral variability. Modeling the temporal structure of the data through transition probabilities is a poor model for the duration of real speech. However, this model is always dominated by the output probabilities anyway.

Each HMM state provides a prototype acoustic vector, and the log output probability function provides a distance metric to measure the likelihood of the state generating the acoustic observation. The choice of the output probability function is crucial. It must be able to model the spectral variability in real speech both within and across speakers. State-of-the-art systems use parametric continuous density output distributions that model the acoustic vectors directly through mixtures of multivariate Gaussians.

To determine the likelihood of a sequence of acoustic vectors $P(A|W)$ the following problem arises: whereas the observation sequence is known, the state

sequence is hidden. This is the reason HMMs are called "hidden". The probability estimate $P(A|W)$ is found by calculating the sum over all possible state sequences. This is done by applying an efficient recursive method, the *Forward-Backward algorithm*. As this algorithm allows to calculate the probability of being in a specific model state at a specific time, Maximum-Likelihood estimates of the HMM parameter sets can be calculated by the Baum-Welch algorithm [Baum 1972]. Instead of summing up over all possible state sequences, also the state sequence that maximizes the likelihood $P(A|W)$ can be used. This is efficiently done using the *Viterbi algorithm* [Viterbi 1967] [Forney 1973].

Speech recognition systems for the English language usually differentiate between 45 phones. Therefore about 45 HMMs seem sufficient to model the acoustic likelihoods. But as contextual effects cause significant variations on pronunciation, different sounds are produced for the same phone in different phonetic contexts. In order to obtain a good phonetic discrimination, different HMMs are trained for different contexts. The most simple approach is to use *triphone models*, where every phone has a distinct HMM model for every unique pair of left and right neighbours. Even better modeling accuracy is achieved when cross-word triphones are used that span word boundaries. Gaussian mixture output distributions allow each state distribution to be modeled very accurately, but require too many parameters to be trained when simultaneously having too little training data. For this reason state-tying and phone-based component-tying have been introduced into most current systems. They reduce the number of trainable parameters and models by sharing the same probability distribution across different triphones.

2.1.2.3 Language Model

The language model within a speech recognition system provides estimates of the probability for a word w_k in an utterance, given the preceding word sequence $W_1^{k-1} = w_1, w_2, ..., w_{k-1}$. Usually *n-grams* are used, meaning that the probability estimate of the next word depends on the preceding $n-1$ words:

$$P(w_k|W_1^{k-1}) \equiv P(w_k|W_{k-(n-1)}^{k-1}) \tag{2.2}$$

N-Grams are able to encode syntax, semantics and pragmatics. However, they are only able to model local dependencies. Nevertheless, this makes n-grams a reasonable and surprisingly reliable model for English where word order is important and the strongest contextual effects come from the nearest neighbours of a word. As shown later statistical n-gram constraints might not be all that appropriate for languages other than English. But even for

English, statistical n-gram models show obvious deficiencies concerning long-range constraints, as for example subject-verb agreement. As a consequence, alternative language modeling techniques have been applied during the last years, such as:

- tree-based models [Bahl et al. 1989]
- trellis models [Waegner & Young 1992]
- trigger models [Lau et al. 1993]
- history models [Black et al. 1992]
- variable n-grams [Deligne & Bimbot 1995]

Also class-based language models have been used widely, either using predefined classes like e.g. function and content words [Geutner 1996], automatic clustering techniques [Kneser & Ney 1993] or stochastic class definitions like e.g. part-of-speech tags [Geutner 1997]. As for all these approaches only very little improvement at considerable computational cost could be achieved, conventional n-gram language models still dominate to date [Jelinek 1991]. Toolkits are publicly available and are used widely [Clarkson & Rosenfeld 1997].

N-Gram probability distributions can be computed directly from text data. No linguistic rules or formal grammars of a language are necessary. The n-grams are estimated from simple frequency counts stored in a look-up table. The most commonly used form of n-grams is the *trigram model*, where word predictions are based on the preceding word pair:

$$P(w_k|W_1^{k-1}) = P(w_k|w_{k-2}, w_{k-1}) \tag{2.3}$$

But even when only considering this relatively short history, a modest vocabulary of N=10,000 words leads to a very large number, namely $10,000^3$, of potential trigrams. This means that the data sparsity problem does not only apply to acoustic samples but also to language model data. Many trigrams do not appear in the training data at all. A large number of others show up only once or twice, thereby providing very poor estimates. To guarantee that probability estimates even for word sequences unseen in the training text are available, normalization procedures and some smoothing methods have to be applied.

The most widespread solution is to combine *discounting* and *backing off* techniques. Discounting means that the trigram counts of more frequently occurring trigrams are reduced and the excess probability mass is redistributed among less frequently occurring trigrams [Ney & Essen 1993] [Ney et al. 1994]. The technique of backing off [Katz 1987] is applied when

there are too few trigrams available to form a reliable estimate at all and thus the estimation falls back to a shorter history, namely bigrams.

As word accuracy or word error rate are used to determine the performance of a speech recognition system, the quality of a language model is measured by its *perplexity* [Bahl et al. 1977]. Perplexity is based on ideas of information theory and represents a measure of the average branching factor of a text as seen from the point of view of a language model. A more detailed description about the principles of lexical language modeling for speech recognition can be found in [Jelinek 1990] [Jelinek et al. 1991].

2.1.2.4 Decoding

Acoustic and language model are the main components of speech recognition systems. As already mentioned the problem of speech recognition consists of finding a sequence of words W that maximizes equation 2.1. It is the task of the decoder to solve this search problem. The two classical approaches for search problems are *depth-first* and *breadth-first search*. In depth-first search the most promising hypothesis is pursued until the end of the speech signal is encountered. Examples are stack- or A^*-decoders. When doing breadth-first search, all hypotheses are pursued in parallel and Bellman's optimality principle is exploited, which is also called Viterbi decoding.

Result of the recognition process in a large vocabulary speech recognition system is the hypothesized speech output of the recognizer representing the most likely uttered sentence. The recognized hypothesis is also called the *First-Best* hypothesis of the speech decoding process. However, the output of most current speech recognizers comprises not only the most likely uttered sentence, but a much richer representation of the search space as pruned by the recognizer. Word graphs or *word lattices* (see section 6.2 for a more detailed description) contain all possibly uttered word sequences for a certain speech input. From this structure, a list of the N most likely utterances hypothesized by the recognizer, also called N-*Best* list, can be provided.

2.1.3 Current State-of-the-Art Large Vocabulary Continuous Speech Recognition

On dictation-style large vocabulary recognition current speech recognition systems show an average word error rate of about 7%. Individual results for different speakers vary from 2.5% up to 20% word error. These figures suggest that better robustness and more effective adaptation techniques are needed. The performance described above was achieved on read speech under controlled recording conditions with known microphones and on a known domain many

times slower than real time. Before widespread deployment of large vocabulary recognition techniques is possible, several other issues will have to be resolved. These are:

- Effective speaker adaptation techniques

  The independence of a speech recognition system from the respective user is highly desirable. Systems should be able to be used "straight out of the box" and their training should be possible even when the speaker is not known in advance. Current research focuses on techniques that yield worthwhile performance gains with very little adaptation data, but also improve performance when a large amount of data has been acquired. Large vocabulary recognizers have a very large number of parameters. As adaptation is supposed to work even on a small amount of new data, no direct re-estimation of the parameters itself but some form of global transformation is performed (*Maximum Likelihood Linear Regression* (MLLR)).

- Environmental robustness

  Robustness to background noise and channel variability is required to guarantee decent recognition performance under changing acoustic conditions. Noise can be dealt with by removing it from speech at the front-end, by using noise-robust features, by masking the noise or mapping the features. All approaches only exploit knowledge of the global characteristics of the speech. An alternative is to make the pattern-matching process itself robust to noise. However, in large vocabulary recognition systems there exists a large number of overlapping classes, and maintaining maximal discrimination is essential. As a consequence, the best approach certainly is to adapt the recognizer to handle the corrupted speech signal directly. Many techniques for speaker adaptation are also capable of adapting to different noise environments. So far for large vocabulary recognition systems, noise-robustness is more or less an unsolved problem.

- Higher task-independence

  Normally the acoustic models of a speech recognizer are relatively task-independent. Unfortunately the language model is typically trained on task-specific material, resulting in a task-dependent system. This mismatch of the language model component increases the number of out-of-vocabulary words and lowers accuracy when moving to different domains. Better language model capabilities would be able to reduce this task dependency. Solutions are to use multiple-domain modeling, where several

individual language models are trained on a range of possible domains and combined as a mixture model. A second approach is to use adaptation, where a static language model is trained off-line and combined with a dynamic cache of n-grams that are continuously updated.

Looking at the problem of out-of-vocabulary words, the robustness and task-dependence of current speech recognition systems concerning new words also has to be increased. One possibility would be to simply add new words to an existing language model. A better solution is the usage of class-based instead of word-based models. Using word classes allows to train more compact and robust language models for very large vocabularies. The problem with this approach is that choosing the right classes is very important. Words should only be grouped into a class if there is little or no increase in perplexity, so efficient data-driven methods are needed.

- The ability to handle spontaneous speech

  Research so far has concentrated on transcribing read speech. In the recent past the research focus has shifted towards being able to transcribe spontaneous casual speech. Looking at the Switchboard database of human-to-human telephone conversations (see in more detail in section 2.2.3.1), the error rate of current systems increases substantially when applied to spontaneous speech. The reasons are unclear, but some of them certainly are: poor articulation, increased coarticulation, a highly variable speaking rate and various types of disfluencies, such as hesitations, false starts or corrections. To achieve satisfactory performance, state-of-the-art large vocabulary speech recognition systems have to be able to handle these effects inherent in spontaneous speech input.

- Real time recognition

  Concerning the requirement to operate under real-time conditions, there is also the need to keep up an acceptable level of recognition accuracy while using a computationally efficient implementation.

2.1.4 Conclusions

The previous section has described the state-of-the-art of current speaker-independent large vocabulary continuous speech recognition systems. So far no robust general purpose large vocabulary speech recognition systems are available, but they are only on the threshold of usefulness for practical applications. As long as the environment the system is supposed to be used in is reasonably controlled and the task well-defined, the technology is already

usable now. More and more off-the-shelf large vocabulary speech recognition systems appear on the market running in real-time on high-end personal computer class machines. Large vocabulary speech recognition system based services appear in form of remote servers in public telecom systems adding telephone-based transcription capability to information and personal management services. They can also facilitate multimedia information retrieval by allowing video soundtracks to be transcribed and searched for keywords and phrases.

2.2 Applications in Speech Recognition

Whereas the very first speech recognition systems normally had to be trained on a certain speaker, current systems are speaker-independent for most applications. Also, progress has been made concerning the level of liberalness for entering speech input to speech recognition systems. Starting off with systems that could only handle the recognition of isolated words and forcing users to pause in between words when uttering complete sentences, nowadays continuous speech input in a natural way is possible and no artificial pausing is required anymore. Inherent in this development, a transition from small vocabulary systems on limited domains to systems that can handle very large vocabularies has taken place.

The variety of such continuous speech recognition systems ranges from information retrieval tasks via dictation systems to the recognition of spontaneous speech input with no restriction whatsoever on the used vocabulary, word order or sentence structure. The following sections introduce different varieties of speech recognition systems and also present examples of special applications.

2.2.1 Information Retrieval Tasks

Initial attempts to compare various speech recognition systems for a certain domain on a larger scale were made when first official tests on the so-called Resource Management Task [Price et al. 1988] were conducted at the end of the eighties [Pallett 1991]. The scenario within which this test took place was a naval officer placing queries to a database containing information about a naval fleet. Requests to the system to obtain information about ships, their condition, location, the number of crew members, and properties like e.g. size or age of the respective ship could be placed verbally. Although inquiries to the speech recognizer could be made quite naturally by using complete sentences, the vocabulary of the underlying speech recognizer covered only about 1,000

words. Also, the uttered sentences had to follow a certain sentence structure dictated by a grammar and defined through a finite state automaton. The resulting task completion error rates were around 5%. Even though having achieved a high success rate, no further evaluations on the Resource Management Task were conducted. The task allowed only a very limited number of input phrases that could be handled properly with its limited vocabulary. Thus, it was rather a means to evaluate task completion on information retrieval tasks, but not suitable to conduct further evaluations on continuous speech input.

Following this very first comparative tests the Defense Advanced Research Projects Agency (DARPA) of the United States of America started to organize official evaluations for information retrieval systems as well as for dictation systems in the beginning of the nineties. Whereas one reason certainly was an effort to create an official common platform on which different speech recognition systems could be evaluated against each other, there was also the goal to provide a common forum where research groups working on speech recognition from all over the world could exchange their ideas and experience, and profit from each other.

The first task to officially evaluate speech recognition systems for information retrieval tasks against each other, was the Air Travel Information Services (ATIS) project in 1990 [Hemphill et al. 1990] [Dahl et al. 1992]. Systems developed within this task and participating in the DARPA evaluations from 1990 to 1994 again operated on a very small and restricted vocabulary of about 2,000 words [Price 1990] [Pallett et al. 1995]. They allowed their users only a fairly limited variety of queries to acquire information derived from a small relational database excised from the Official Airline Guide. This information included flight schedules, airfares, type of meals served on the flight as well as the type of the aircraft available between a restricted set of cities within the United States and Canada. Although spontaneous input within the limits of the task was possible, input requests were normally fairly planned, as people believed they were talking to a machine. Queries to the systems followed easy-to-formulate rules and also satisfied a predefined grammatical structure which was necessary to guarantee sufficient recognition performance and task completion. As for information retrieval tasks a very simple structure of requests is already sufficient, the restricted vocabulary and the very firm structure of input to such systems did not impede their usefulness.

In the meantime, a lot of different information retrieval systems are available, see e.g. [Zue 1997] for a summary on conversational interfaces. Applications range from the retrieval of travel-related issues, like train [Eckert et al. 1993] or flight information [Zue et al. 1994], to obtaining the weather forecast. The **A**utomatic **R**ailway **I**nformation **S**ystems for **E**urope

(ARISE) project [Lamel et al. 1998], for example, developed a prototype for automatic train schedule inquiry services that allows to get information about train timetables. The Multimodal Multimedia Automated Service Kiosk (MASK) project [Lamel et al. 1995] [Gauvain et al. 1997b] also offers train-related information. Within this project a prototype service kiosk was developed and underwent user tests located in a Parisian train station. Users of the system can obtain information on train schedules, services and fares. The kiosk does not only allow speech input but enables interaction through the coordinated use of multimodal input sources like speech and touch as well as multimedia output like sound, video, text and graphics. Another system providing travel-related on-line information is the GALAXY system that enables universal information access using spoken dialogue [Goddeau et al. 1994]. JUPITER, an offspring of the GALAXY system, offers information about the weather in more than 500 cities world-wide [Zue et al. 1997]. However, all these systems still have in common to operate on a very limited domain, can only handle a small variety of sentence structures and are restricted to a predefined and fixed set of vocabulary words.

Another application of speech input to information systems is the automotive environment. Whereas in the beginning speech recognition technology in car applications was restricted to the control of the telephone or entertainment components such as radio or CD-player, in the recent past utilization of speech technology has been extended to the navigation component of driver information systems [Van Compernolle 1997]. One application is the Voice Operated Driver Information System (VODIS) [Pouteau & Arévalo 1998]. Within this project speech is not only used to operate in-car devices such as radio, CD-player or telephone, but also the input component of a navigation system. The transition from traditional tactile interfaces to speech input provides easy-to-use and user-friendly interfaces that allow a higher user comfort and security when operating driver information systems. So far the usage of voice-operated systems in the car has been restricted to the input of a predefined set of keywords only. Hence, medium vocabulary recognition systems with more or less rigid dialogue structure have been sufficient for this task. It is only recently that current research activities go beyond command and control issues from keyword recognition on limited vocabularies to spontaneous speech input that allow the operation of a driver information system in a natural way [Geutner et al. 1998a].

2.2.2 Dictation Systems

Compared to systems that are used for information retrieval purposes only, dictation systems allow a much larger variety of speech input. First research activ-

ities on large vocabulary speech recognition systems for dictation applications were performed on the Wall Street Journal (WSJ) task [Paul & Baker 1992]. Similar to the ATIS project DARPA also organized annual evaluations on this task to bring together researchers from academia and industry to mutually exchange ideas and experience. Evaluations were carried out from 1992 to 1994 when a transition from continuous read speech to spontaneous speech input took place and first evaluations on spontaneous speech tasks started. The evaluations on dictation systems were carried out on read newspaper articles from the Wall Street Journal recorded by a number of speakers [Gauvain et al. 1995] [Bahl et al. 1995]. The system that performed best in the final evaluation on the WSJ task in 1994 achieved a 7.2% word error rate with a vocabulary of 60,000 words [Woodland et al. 1995].

In the meantime, a number of different dictation systems for automatic speech recognition are already commercially available [Baker 1993]. Usually the usage of such a system requires some hours of speech training to be used optimally by a certain speaker. However, whereas the very first systems that were available on the market were restricted to certain domains such as medical or legal applications for doctors or lawyers, current dictation systems are not restricted to a specific domain anymore. They allow an almost unlimited vocabulary and arbitrary grammatical construction of sentences. As a consequence, they have to be able to handle several thousand words of vocabulary. Still, users of currently available systems wish to use them for arbitrary purposes and domains, so that even large vocabularies of 60,000 words or more are not enough to guarantee a perfect (= 100%) coverage of the used input. Therefore, in commercially available dictation systems new words can always be taught to the system and added to the recognition dictionary. By that means state-of-the-art dictation systems can handle a vocabulary of more than 60,000 words and thus are classified as large vocabulary systems where the used vocabulary is normally connected to the respective person utilizing the system. In contrast to the restricted input for information retrieval systems where queries have to follow a certain predefined structure and also have to stick to a fixed vocabulary, dictation systems allow unrestricted speech input on an unlimited vocabulary. However, compared to conversational or spontaneous speech, input to dictation systems, often also referred to as "read" speech, is easier to recognize because of the following reasons: first, people tend to use well-structured grammatical sentences when dictating to a speech recognition system. Consequently, it is very rare that a recognition system dedicated to a dictation task has to be able to handle sentences not following a grammatical structure. Second, when dictating text, people usually do not act very spontaneous and do not make a lot of mistakes. So-called "spontaneous" speech is rare and effects like hesitations, stuttering, false starts, coughing, laughing

or breathing noises that influence recognition performance negatively, do not come up very often. This is different for applications that require the recognition of conversational or spontaneous speech. The phenomena of spontaneous speech input will be described in the following section.

2.2.3 Spontaneous Speech

As more and more applications arise that justify the usage of speech recognition, also the spectrum of domains broadens and the necessary dictionary size increases. Recently interest in Large Vocabulary Continuous Speech Recognition (LVCSR) research has shifted from read speech to speech data found in real world applications like conversational speech over the telephone or the transmission of broadcast news over radio and television. The recognition of conversational or spontaneous speech input creates a new challenge for traditional speech recognition systems as they encounter new and so far ignored speech phenomena. These effects have to be successfully handled in order to offer speech systems with satisfactory recognition performance.

2.2.3.1 Conversational Speech

Human-to-human or human-machine conversations are characterized through different features and follow different rules than continuous or read speech, as it is found when entering speech input into a dictation system. In conversational speech often ungrammatical sentences are uttered that do not follow a predefined sentence structure for a certain domain or application. Moreover, they do not follow any grammatical rules for a specific language and therefore are linguistically incorrect. Similar to dictation systems it is also impossible to restrict the user of a conversational speech recognition system to a predefined and fixed set of vocabulary words. Even worse than for dictation tasks where a person might stick to a certain topic, for example when dictating a business letter, it is a special feature of conversational speech that it cannot be restricted to a specific domain. Also, spontaneous speech effects like the ones described in the previous section: ungrammatical sentences, disfluencies like hesitations or stuttering, fragments like false starts or word fragments, heavy articulations such as breathing noises, lip smacks, coughing or laughing as well as other environmental noises are quite common here and have to be handled properly in order to facilitate the deployment of speech recognition systems in real world applications. In summary, recognition of conversational speech, compared to read speech, is aggravated through the following reasons:

- unlimited domain

- unrestricted vocabulary
- ungrammatical speech
- spontaneous speech effects

When trying to recognize conversational speech between two people as it is done in the Verbmobil project [Wahlster 1993] all of these effects will appear and worsen recognition performance significantly. The project is concerned with the development of a speech-to-speech translation system for face-to-face dialogues. The Verbmobil domain expects two people trying to schedule a meeting with each other. For data collection purposes both individuals are given their personal engagement diaries. They are then supposed to negotiate about time and place of the appointment according to this scenario. Furthermore, they also have to try to plan the trip that is required to let the meeting take place. Different scenarios for the engagement diaries were devised to record German, English and Japanese conversations and collect them within the German, English and Japanese Spontaneous Scheduling Task (GSST, ESST and JSST). Data collection was done in all three languages as the project does not only focus on the recognition of spontaneous conversational speech, but also on the translation between several languages. This is done by translating the utterance of one conversational partner into the mother tongue of the other and vice versa, or by using some kind of intermediate language that both partners do at least understand, like e.g. English. The Verbmobil project is a German project sponsored by the BMBF [1]. Throughout its duration, similar to DARPA projects, also annual evaluations were organized by the BMBF to compare different speech recognition systems developed and trained by several partners participating in the project. On behalf of the University of Karlsruhe a system [Geutner et al. 1995] trained with the Janus Recognition Toolkit (JRTk) that was used for all speech recognition experiments within this work, also participated and performed as one of the best each year evaluations were conducted [Finke et al. 1997b].

However, the Verbmobil project still acts on a very limited domain and the conversations collected in the project stay within the scheduling task. This is different for the Switchboard or Callhome project [Godfrey et al. 1992] [Jeanrenaud et al. 1995] where unrestricted telephone conversations between two people are treated. For this project, also sponsored by the American DARPA, a collection of English spontaneous conversations between two people via telephone with a pregiven topic out of a selection of 70 topics was recorded. Although each conversation was assigned a certain topic, people often did not stick to it and started to deviate from the subject. Spontaneous speech effects

[1] Bundesministerium für Forschung und Technik (BMBF)

are ubiquitous and decent recognition of the data is further aggravated by the noisy telephone channel the data is recorded on. Evaluations on this task have been conducted every year. Again the Janus Recognition Toolkit from the University of Karlsruhe was one of the best systems in the last two evaluations in 1996 and 1997 [Finke et al. 1997a] [Geutner et al. 1997b].

2.2.3.2 Automatic Transcription of Broadcast News

Transcription of broadcast news data and television news has added a number of additional challenges to traditional, mainly dictation-based large vocabulary transcription systems. Data contained in broadcast news shows is not homogeneous and is characterized by a bewildering variety of different acoustic conditions since the speaker, speaking style, channel and environment change frequently. A typical broadcast news show ranges from clear read, so-called "baseline" speech to conversational or spontaneous speech. The different speech styles include speech from native and non-native speakers, telephone speech, high or low bandwidth speech with or without background music or noise, and speech under degraded acoustic conditions. Speech recognition systems trained on read speech corpora such as the Wall Street Journal corpus show a high word error rate for applications like this.

Since 1996 the Broadcast News task is incorporated into the DARPA-sponsored Automatic Speech Recognition (ASR) Benchmark Tests where a number of speech recognition tasks is evaluated annually [Garofolo et al. 1997]. In the 1997 annual evaluation the best result reported for the Broadcast News Benchmark Test was a word error rate of 16.2% [Pallett et al. 1998] compared to a word error rate of 27.1% in 1996 [Pallett & Fiscus 1997]. However, this difference in word error rate does not mirror the actual improvements that have been made when recognizing broadcast news. Part of this apparent improvement is due to a different proportion of well-recognized data in the test data compared to the 1996 test set. The different balance of test data proportions was the result of an effort to balance the test pool to better match the properties of the training data whereas the 1996 test set was selected to maximize focus condition coverage. Also, the 1997 test set contained no markers indicating speaker or show changes. The evaluation systems of all participating groups had to rely on automatic methods of segmenting the audio data into manageable pieces. Additionally, no information was provided about channel conditions, speaker gender or accent, the presence of noise or music, or speaking style as it was done in the previous year. This very realistic scenario forced all participating recognition systems to intelligently and automatically cope with a variety of acoustic and linguistic conditions. A total of ten research groups from nine different sites participated and submitted results.

For the 1997 evaluation the BBN group with its Byblos system [Kubala et al. 1998a] focused its attention entirely upon those portions of the news broadcasts containing studio-quality, uncorrupted speech for native speakers of American English. The motivation for this decision was the observation that baseline speech recognition performance on clean wideband data is still unacceptably poor for most applications. As this kind of speech represents the majority of the content-rich portion of the news that would be most useful for information retrieval or extraction applications, it seems reasonable to concentrate research on this part of the data. Also, fundamental improvements in recognition accuracy will be most easily achieved on uncorrupted data first. The expectation is that some fraction of it will certainly translate to the corrupted conditions. The same idea was pursued by the IBM group where the 1997 evaluation system was also specifically built for handling clean baseline speech that is either read or conversational [Chen et al. 1998]. Also, new segmentation and clustering algorithms were used by almost all participants of the evaluations. LIMSI for example mainly addressed the problem of partitioning the continuous stream of acoustic data as well as improving and simplifying the acoustic models [Gauvain et al. 1998].

Current research of the group from Cambridge University [Woodland et al. 1998] concentrated on data where no segmentation is given and also the type of data is unknown. A data segmentation and classification scheme which incorporates clustering is developed where the data is first segmented into homogeneous segments of differing data types. Also, segments that contain no speech at all but e.g. background music only are rejected. Based on this automatic segmentation, acoustic modeling techniques are developed that do not rely on detailed hand-derived data classification. HMMs that are independent of individual data types are used and fit well with this automatic data segmentation and classification. The performance of models that do require knowledge of the data type was compared with condition independent models which are more suitable to automatically segmented data since fine classification is not required. As already previously shown, data condition independent models can give surprisingly good performance [Gauvain et al. 1997c] [Kubala et al. 1997a]. This fact also applied to the 1997 broadcast news evaluation where these models yielded at least as good performance as data type specific models that had been used in 1996 [Woodland et al. 1997]. The so-developed evaluation system performed best in the annual evaluation and yielded the lowest overall word error rate.

2.2.4 The Informedia Project

The Informedia Project at the Carnegie Mellon University (CMU) has been going on since 1993. Core of the Informedia system is the Informedia Digital Library [Christel et al. 1995], a multimedia database consisting of digital video, audio, images, text and other related material. It integrates speech recognition, image understanding and natural language processing technologies to transcribe, segment and index audio and video documents. Automatic processing of the input data enables users of the system to perform a full-content and knowledge-based search and retrieval [Hauptmann & Wactlar 1997]. To accomplish intelligent search and selective retrieval the same tools as for incorporating data into the multimedia database are used, thus enabling a fast user access to huge amounts of formerly unstructured heterogeneous data sources.

Especially within the domain of broadcast news, the amount and availability of information steadily increases. As a consequence search and retrieval of particular documents within this domain can be very difficult. Once a radio or television broadcast has been transmitted, access to specific information included in the news show is hard to be retrieved again. One solution would be to scan through all available audio and video data. However, this is a time-consuming and tedious task, and will in many cases not provide the desired results. Thus techniques providing efficient access to many different information sources and types are needed. Within the Informedia project this is done by automatically generating transcripts of news shows through a speech recognition system or using closed caption if available. When transcribing a news broadcast, the content of the show is extracted and a time alignment between audio and text data is generated. The content of the show is further structured by additional video processing: story boundaries are determined by analyzing changes in the acoustic conditions of the broadcast as well as in the video signal. Other algorithms make use of word frequency and statistical analysis to find text pieces belonging to the same topic. All these different techniques for extracting, structuring and indexing the news data are being applied to overcome the linear access restrictions, to ease and speed up the retrieval of relevant information.

Beside image understanding and information retrieval techniques, the Informedia system also comprises a speech recognition component. Speech technology is used both for the creation and exploration of the Informedia Digital Library [Hauptmann & Witbrock 1997b]. To this end the Sphinx speech recognition engine has been used [Huang et al. 1993] [Placeway et al. 1997] [Seymore et al. 1998]. When creating the digital video library, transcripts of the broadcast news shows are automatically generated. Concerning the exploration of the library by posing a query to the Informedia system, input into the

system is either entered by typing or by speaking into a microphone. A spoken query is first submitted to a similar speech recognition system as used for the automatic transcription process and then submitted to the database. Through information retrieval techniques information relevant to the query is searched within the multimedia database. Documents retrieved are then returned to the user satisfying the query submitted to the system.

2.2.5 The Multilingual Informedia Project

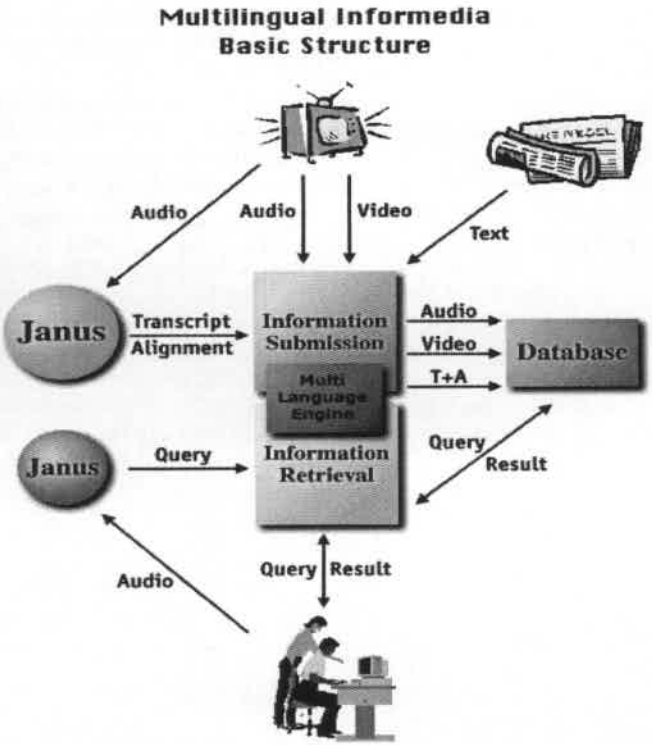

Figure 2.2: Integration of the Serbo-Croatian Recognizer into the Multilingual Informedia Project.

As an extension to the Informedia Project that had so far concentrated on automatic transcription and retrieval of English news documents only, in 1997 the project was enhanced by the goal of extending the Informedia system to multilinguality. Beside English the second language to be added to the multilingual structure of the project was Serbo-Croatian. The so-called Multilingual Informedia Project at Carnegie Mellon University is a collaboration of the Informedia group at CMU [Hauptmann & Witbrock 1997a] concerned with the development of the digital video library, the Language Technologies Institute (LTI) responsible for translation and cross language retrieval, and the Interactive Systems Laboratories (ISL) developing the speech recognition engine for the Serbo-Croatian language. To this end Serbo-Croatian broadcast news shows were recorded at the University of Karlsruhe (see also chapter 4) for the development of a Serbo-Croatian broadcast news recognizer based on the Janus Recognition Toolkit.

The resulting Serbo-Croatian speech recognition system was integrated into the Multilingual Informedia System as illustrated in figure 2.2. The automatically generated transcripts are inserted into the multilingual Informedia database. Together with the Serbo-Croatian broadcast video material, the transcripts and the recognizer allow for automatic content-addressable search and multimedia document retrieval across languages.

The extension of the Informedia database to more than one language does not only add to the diversity of information retrieved by monolingual queries, but also offers the possibility to phrase queries in several languages. By being able to retrieve news shows from various countries in several languages, the scope of potential users is extended to a broader view about national and international events. Thus, the development of our Serbo-Croatian recognizer [Scheytt et al. 1998] provides an instance of a potentially larger multilingual information resource.

Chapter 3

Lexical Properties Across Languages

Each of the world's major languages has certain characteristics that distinguish it from most of the others. These characteristics have to be identified and can be exploited when doing speech recognition on a specific language. The following chapter gives an insight into the languages on which recognition experiments were conducted within this work: Serbo-Croatian and German. In comparison the distinctive features of the English language are presented. Beside possible dialects and pronunciations, the writing system and spelling, phonology and morphology of each of these languages are introduced [Comrie 1987]. Special emphasis is put on the lexical properties of the three languages concerning noun declensions, verb conjugations, and compound words or composita. Graphs on the vocabulary growth of the respective languages are illustrated. Also, the out-of-vocabulary rates on selected test sets dependent on different vocabulary sizes are shown. The chapter concludes with an overview over the vocabulary growth of the three languages introduced before, complemented by the vocabulary growth curve of a Turkish database.

3.1 Serbo-Croatian

3.1.1 Dialects and Variations

The Serbo-Croatian language consists of three main dialects: Čakavian, Kajkavian and Štokavian. The literary language is based on Štokavian and this dialect itself has again two different major dialectic variations: the Ekavian and the Ijekavian dialect group. Ekavian is spoken in most of Serbia, Ijekavian is found in the western part of Serbia, Montenegro, the east of Bosnia and Hercegovina, and parts of Croatia. Whereas Ekavian is the basis of the east-

ern variety of the literary language with Belgrade as its centre, Ijekavian is the foundation of the western variant whose focal point is Zagreb. The eastern variety is traditionally written in Cyrillic, but nowadays also the Latin alphabet is often used, as it is the case for the western variety.

Another important difference is the representation of the sound combinations "ije", "je" or "i" that are distinguished in Ijekavian. In the corresponding Ekavian words they are written as a single "e". Table 3.1 gives some examples of both variations.

Ijekavian	Ekavian	Translation
rijeka, rjeka	reka	river
lijepo, ljepo	lepo	beautiful
riječnik, rječnik	rečnik	dictionary
vijetar, vjetar	vetar	wind
htio	hteo	he wanted to

Table 3.1: Ijekavian and Ekavian Variants of Standard Serbo-Croatian.

Finally there are some obvious differences in lexis. A lot of things are referred to by totally different and unrelated words in the two different dialects. The word "kruh", for example, means "bread" and is used in Ijekavian, whereas the word for "bread" is "hleb" in Ekavian.

There are fewer borrowings in the west and correspondingly more calques and neologisms. When looking at words borrowed from other languages, words that are used in the western variety predominantly come from German, Latin and Czech, whereas the ones used in the eastern variety are mostly borrowed from Turkish, Greek and Russian. Words that have been borrowed by both varieties may show differences in derivational morphology. Thus, the word "student", meaning "male student", is used in both dialects, whereas "female student" translates to "studentica" in Ijekavian and "studentkinja" in Ekavian.

Also, often the differences between both variations are not absolute but a matter of frequency of usage. Many features that are claimed to be characteristic for one of the dialects also occur in the other, though they are less common there. These features suggest to view Serbo-Croatian as one language with two varieties, whereas natives often consider it very important to recognize Croatian and Serbian as two distinct languages.

3.1.2 The Writing System

A new version of the Cyrillic alphabet was introduced in 1818 when the existing alphabet was simplified. Since this major reform a single letter is used for

each sound and a phonetically based orthography was adopted. The Serbo-Croatian Latin alphabet underwent the equivalent reform a little later, using diacritic symbols on the Czech model. It is with minor modifications still in use today. Unlike the Cyrillic alphabet it includes three digraphs: "lj", "nj" and "dž". As the individual combinations of the two letters of each of these digraphs occur very rarely or not at all, they cause only little problems. Except for these digraphs, the Latin alphabet also uses a single letter per sound and sticks to a phonetically oriented spelling. This basically means that the Serbo-Croatian orthography is based on the phonemic principle. Assimilations are indicated in spelling and the phonemic principle is applied at the expense of the morphological principle with unusual consistency. This characteristic of the Serbo-Croatian language is one of the features that influenced the development of the speech recognition engine for Serbo-Croatian described in the following chapter significantly. The close correlation of Serbo-Croatian orthography and pronunciation also proved to have an important influence on the research presented in this work and was exploited during the development and realization of some of the ideas and algorithms described in the following chapters. Table 3.2 shows a comparison of the modern Serbo-Croatian and Cyrillic alphabets [Corbett 1987].

Latin	Cyrillic	Latin	Cyrillic
A a	А а	L l	Л л
B b	Б б	Lj lj	Љ љ
C c	Ц ц	M m	М м
Ć ć	Ћ ћ	N n	Н н
Č č	Ч ч	Nj nj	Њ њ
D d	Д д	O o	О о
Dž dž	Џ џ	P p	П п
Dj dj	Ђ ђ	R r	Р р
E e	Е е	S s	С с
F f	Ф ф	Š š	Ш ш
G g	Г г	T t	Т т
H h	Х х	U u	У у
I i	И и	V v	В в
J j	Ј ј	Z z	З з
K k	К к	Ž ž	Ж ж

Table 3.2: Serbo-Croatian Latin and Cyrillic Alphabets.

3.1.3 Phonology

In Serbo-Croatian the inventory of segmental phonemes is one of the smallest in the Slavonic family because there exist much less palatalized consonants like e.g. in Russian. Generally a set of 25 consonants can be identified, and there are five vowels.

	bilabial	labio-dental	dental	alveolar	palato-alveolar	palatal	velar
Plain Stop							
voiceless	p		t				k
voiced	b		d				g
Affricate							
voiceless			c		č	ć	
voiced					dž	dj	
Fricative							
voiceless		f		s	š		x
voiced		v		z	ž		
Nasal	m		n			nj	
Lateral			l			lj	
Trill				r			
Semi-Vowel						j	

Table 3.3: Articulation of Serbo-Croatian Consonants.

3.1.3.1 Consonants

The table above presents the segmental phonemes for the set of Serbo-Croatian consonants and is taken from [Gvozdanović 1980] and [Langenscheidts Sprachführer 1997]. It explains the articulation type and position of the phonemes. For training the Serbo-Croatian speech recognition system used for all experiments conducted in this work, a larger set of acoustic and articulatory features to determine a set of phone classes was used [Scheytt 1997]. For a more detailed description of Serbo-Croatian phonology see [Gvozdanović 1980] or [Langenscheidts Sprachführer 1997].

3.1.3.2 Vowels

Tongue	front	center	back
high	i		u
middle	e		o
low		a	

Table 3.4: Articulation of Serbo-Croatian Vowels.

The Serbo-Croatian language comprises a set of five vowels. These vowels follow certain patterns of accentuation and thereby provide the most interesting feature of Serbo-Croatian phonology as described in more detail in the next section. The table above demonstrating the articulation of Serbo-Croatian vowels is also a synopsis of [Gvozdanović 1980] and [Langenscheidts Sprachführer 1997].

3.1.3.3 Accentuation

As Serbo-Croatian is a very melodic language, different accents based on stress, tone and length can be clearly identified. The representation of the various accentuation markers in table 3.5 is extracted from [Gvozdanović 1980]. Accentuation of vowels can vary according to length and pitch. They may be short or long, both in stressed position or positions after the stress. Pitch is differentiated only in initial stressed position, where there is a distinction between rising and falling tone. Serbo-Croatian stress symbols are used in dictionaries and grammars, but are not printed in ordinary texts.

		long	short
stressed syllables	falling tone	^	"
	rising tone	'	`
unstressed syllables		-	

Table 3.5: Serbo-Croatian Accentuation.

3.1.3.4 Pronunciation Examples

Table 3.6 summarizes some pronunciation examples to give an impression of the sound of the Serbo-Croatian language.

Phoneme	Serbo-Croatian	German	English
a	ona, rad	hatte, natürlich	father
b	dobar, Zagreb	Bein, heben	bag
c	centimetar, utakmica	zahm, sitzen	rats
ć	ćevapčići, kući	Mädchen	
č	četiri, či	Matsch, rutschen	church
d	dan, sedam	dumm, Adel	dog
dj	duvec, mlada		jar
dž	džezvica, udžbenik		jar
e	gleda, evo	fett, Männer	bed
f	fabrika, kafa	Vater, Affe	fun
g	govori, digao	Gift, Segen	get
h	Ahmetm, novih	ach, Loch	Loch Ness
i	pita, bilo	sieben, nie	police
j	jedan, rijeka	Jammer, Jahr	yes, boy
k	ko, istok	Kamm, Socke	ski
l	lampa, bilo	Lampe	like, let
lj	ljudi, nedjelja		million
m	malo, znam	Murmel, Dame	meet
n	narodni, stanica	nackt, wenig	note
nj	Njemačka, pičinje		onion
o	korzo, odgovara	Post, offen	port
p	petak, opet	pumpen, Post	spy
r	zar, odgovor, prvi		
s	sada, silazi	reißen, Riß	six
š	šeta, šta	schief, pfuschen	ship
t	ti, utorak	toll, Platte	stop
u	uči, fabriku	klug, Schuh	boot
v	voda, provesti	Vase, wo	very
z	zove, jezik	sagen, Eisen	zero
ž	možda, muž	Journalist, Genie	measure

Table 3.6: Serbo-Croatian Pronunciation Examples.

3.1.4 Morphology

In this work the notion of morphology is used to denote the decomposition of words into their word stems and inflection endings. Knowledge about morphology is used in the following chapters

- to conduct morphology-based recognition experiments (see also chapter 5), as well as
- to do vocabulary adaptation (see chapter 7).

Therefore this section gives a short insight into the morphological characteristics of the Serbo-Croatian language, as these characteristics will be exploited by the research approaches described in the subsequent chapters.

3.1.4.1 Nouns

In Serbo-Croatian we encounter seven different grammatical cases together with three genders which are distinguished both in the singular as well as the plural. Generally morphology is fusional and there is a strong correlation of gender with declensional class. However, there are some exceptions, as for example in the vocative case where a mutation of consonants for many masculine nouns can be found. Also, some consonant stems are preserved and certain suffixes may be added or lost in the declension of masculine nouns. Length and tone of the stressed syllable may change, as the position of stress may move. But in general, with very few exceptions, all Serbo-Croatian nouns are declinable. This means that there exists a vast number of noun declension endings. Table 3.7 gives examples of the declension endings of the feminine noun "žena" ("woman"), the masculine noun "zakon" ("law") and the neuter words "selo" ("village") and "stvar" ("thing"). These examples together represent the main types of Serbo-Croatian noun declensions.

Feminine a-Stem	**Singular**	**Plural**	**Translation**
Nominative	žen-a	žen-e	woman
Vocative	žen-o	žen-e	woman!
Accusative	žen-u	žen-e	(the) woman
Genitive	žen-e	žen-a	woman's
Dative	žen-i	žen-ama	(I gave it to the) woman
Instrumentalis	že-nom	žen-ama	(with the) woman
Locative	žen-i	žen-ama	(I went to the) woman

Masculine o-Stem	**Singular**	**Plural**	**Translation**
Nominative	zakon	zakon-i	law
Vocative	zakon-e	zakon-i	law!
Accusative	zakon	zakon-e	(the) law
Genitive	zakon-a	zakon-a	law's
Dative	zakon-u	zakon-ima	(to the) law
Instrumentalis	zakon-om	zakon-ima	(with the) law
Locative	zakon-u	zakon-ima	(to the) law

Neuter o-Stem	Singular	Plural	Translation
Nominative	sel-o	sel-a	village
Vocative	sel-o	sel-a	
Accusative	sel-o	sel-a	
Genitive	sel-a	sel-a	
Dative	sel-u	sel-ima	
Instrumentalis	sel-om	sel-ima	
Locative	sel-u	sel-ima	

Neuter i-Stem	Singular	Plural	Translation
Nominative	stvar	stvar-i	thing
Vocative	stvar-i	stvar-i	
Accusative	stvar	stvar-i	
Genitive	stvar-i	stvar-i	
Dative	stvar-i	stvar-ima	
Instrumentalis	stvar-ju/stvar-i	stvar-ima	
Locative	stvar-i	stvar-ima	

Table 3.7: Examples of Serbo-Croatian Noun Declensions.

3.1.4.2 Adjectives

Most of the numerals no longer decline, but adjectives, when used pronominally, correspond in case, number and gender with the noun. The accusative singular masculine form of an adjective even depends on the animacy of the noun. For animate nouns the accusative form of the masculine singular adjective is identical to the genitive, for inanimate nouns nominative and accusative are the same. Distinction of the respective cases is mainly done by the notion of definite and indefinite adjectives. These two forms of adjectives differ by inflection only in the masculine singular form:

- definite: "dobri čovek" ("the good man")
- indefinite: "dobar čovek" ("a good man")

Table 3.8 gives an example of the definite adjective "mladi" ("young") with optional forms given in brackets.

Feminine	Singular	Plural	Translation
Nominative	mlad-a	mlad-e	young
Vocative	mlad-a	mlad-e	
Accusative	mlad-u	mlad-e	
Genitive	mlad-e	mlad-ih	
Dative	mlad-oj	mlad-im(a)	
Instrumentalis	mlad-om	mlad-im(a)	
Locative	mlad-oj	mlad-im(a)	

Masculine	Singular	Plural	Translation
Nominative	mlad-i	mlad-i	young
Vocative	mlad-i	mlad-i	
Accusative	as Nom. or Gen.	mlad-e	
Genitive	mlad-og(a)	mlad-ih	
Dative	mlad-om(e)	mlad-im(a)	
Instrumentalis	mlad-im	mlad-im(a)	
Locative	mlad-om(e)	mlad-im(a)	

Neuter	Singular	Plural	Translation
Nominative	mlad-o	mlad-a	young
Vocative	mlad-o	mlad-a	
Accusative	mlad-o	mlad-a	
Genitive	mlad-og(a)	mlad-ih	
Dative	mlad-om(e)	mlad-im(a)	
Instrumentalis	mlad-im	mlad-im(a)	
Locative	mlad-om(e)	mlad-im(a)	

Table 3.8: Examples of Serbo-Croatian Declensions of the Definite Adjective.

3.1.4.3 Verbs

When looking at verbal morphology, also a large variety of forms can be found. Having moved from a system based on tense to one in which aspect has a central role, Serbo-Croatian preserves present tense, imperfect and aorist tense. In this work only the present tense will be considered. Present tense conjugations of the verbs "govoriti" ("to speak"), "imati" ("to have") and "tresti" ("to shake") are illustrated for all variations in singular and plural in table 3.9.

	Decomposition	Translation
Infinitive	govor-iti	to speak
1. Person Singular	govor-im	I speak
2. Person Singular	govor-iš	you speak
3. Person Singular	govor-i	he speaks
1. Person Plural	govor-imo	we speak
2. Person Singular	govor-ite	you speak
3. Person Singular	govor-e	they speak

Infinitive	im-ati	to have
1. Person Singular	im-am	I have
2. Person Singular	im-aš	you have
3. Person Singular	im-a	he has
1. Person Plural	im-amo	we have
2. Person Singular	im-ate	you have
3. Person Singular	im-aju	they have

Infinitive	tres-ti	to shake
1. Person Singular	tres-em	I shake
2. Person Singular	tres-eš	you shake
3. Person Singular	tres-e	he shake
1. Person Plural	tres-emo	we shake
2. Person Singular	tres-ete	you shake
3. Person Singular	tres-u	they shake

Table 3.9: Examples of Serbo-Croatian Verb Conjugations.

3.1.5 Vocabulary Growth

As shown in the previous section, Serbo-Croatian is a representative for a highly inflected language. Due to its abundance of different word forms found both as noun declensions and verb conjugations, the language shows a rapid vocabulary growth. Figure 3.1 plots this development for a text corpus consisting of 12 million words of Serbo-Croatian newspaper articles and texts from news agencies, and television or radio stations retrieved on the world-wide-web[1]. The more newspaper articles are collected, meaning the bigger the size of the text considered to determine the number of unique words within this corpus, the larger gets the resulting vocabulary size. The diagram in figure 3.1

[1] A more detailed description of this text corpus can be found in chapter 4, section 4.1.

also illustrates that even for a text size of almost 12 million words from the available Serbo-Croatian newspaper and web text corpus, no saturation in the number of distinct words is to be expected.

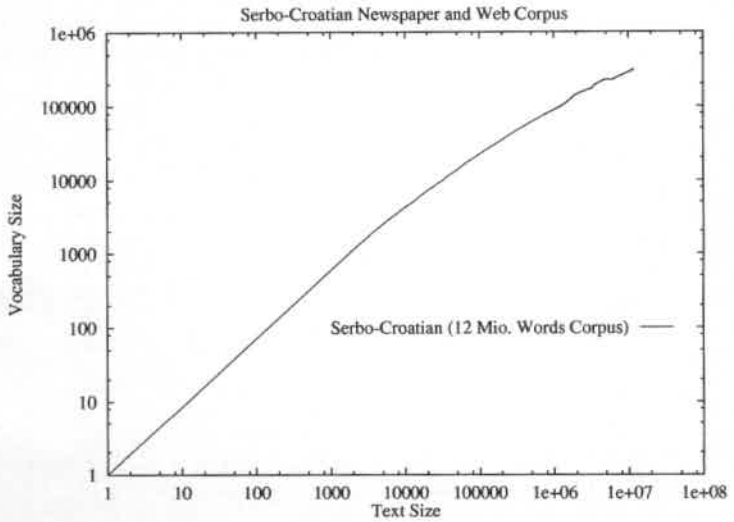

Figure 3.1: Serbo-Croatian Vocabulary Growth (Newspaper and Web Text Corpus).

Beside its vocabulary growth, another interesting feature of a language is the closely related term *coverage*. Coverage denotes the percentage of unique vocabulary words from a text corpus that is contained in or covered by a certain vocabulary. To obtain the vocabulary the coverage is calculated on, often vocabulary lists derived from very large text corpora are used. Two different types of coverage can be distinguished: *self coverage* means calculating the number of distinct words of a text corpus that are already covered by considering only parts of the text itself to generate the vocabulary the coverage is measured on. Starting from an empty vocabulary list, constantly getting larger subsets of various sizes of the available text are considered. In the beginning, with an empty vocabulary list, all words of the respective text will be unknown and the coverage is 0%, until, when finally the whole text is taken into consideration, 100% of all distinct words of the text are covered by itself and the resulting vocabulary list respectively. When determining the *cross coverage* of a text corpus or the vocabulary list that can be generated from it, a

separate test text corpus is used. Cross coverage means calculating the number of distinct vocabulary words of this separate test text that are already included in or covered by the vocabulary list generated from the original background text.

The same corpus as briefly described above is used for figure 3.2 where the self coverage of the corpus and the cross coverage on a distinct test text are indicated.

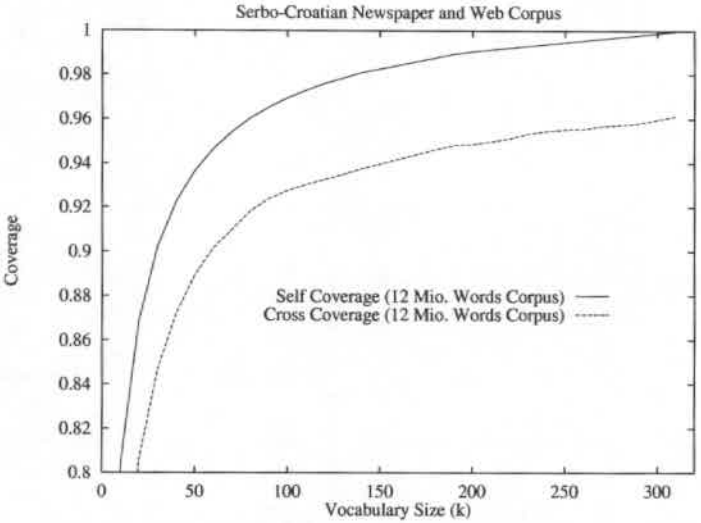

Figure 3.2: Serbo-Croatian Self and Cross Coverage (Newspaper and Web Text Corpus).

3.2 German

3.2.1 Dialects

The German language is a member of the West Germanic group within the Germanic branch of Indo-European languages. German is spoken by about 94 million speakers within Europe. Basically there exist three main dialects: Low German in the north, Central German, and Upper German mainly spread in the south. The term High German is now used to subsume Central and Upper German as opposed to Low German.

As for German phonology, the phonological norms were finally set in the nineteenth century. Speakers from the north tended to accentuate a close relation between phonemes and graphemes. With minor modifications this North German pronunciation of the originally High German became the norm for today's standard German pronunciation. There are still different German dialects spread across Europe that can be found in Germany, Austria and Switzerland. But as far as the written language is concerned, there is now a widespread consensus among the German speaking countries. Taking a closer look at the German syntax, at some point also the characteristic verb position of Modern German was fixed: final position in subordinate clauses, second and first position in main clauses.

3.2.2 Phonology

The phonological standard for the German language that has been set in the nineteenth century is still in use today. The segmental phonemes of Modern Standard German comprise 21 consonant phonemes and 19 separate vowels of which three are diphthongs.

	bilabial	labio-dental	dental	alveolar	palato-alveolar	palatal	velar	glottal
Plain Stop								
voiceless	p		t				k	
voiced	b		d				g	
Fricative								
voiceless		f		s	ʃ	ç	x	h
voiced		v		z	ʒ			
Nasal	m		n				ŋ	
Lateral			l	r				
Semi-Vowel						j		

Table 3.10: Articulation of German Consonants.

3.2.2.1 Consonants

The German language normally distinguishes 21 consonant phonemes. Most of them can take word-initial, word-medial and word-final position. Table 3.11 from [Hawkins 1987] gives examples of the 21 phoneme consonants. If a table entry is left blank, the respective consonant does not occur in the mentioned position within a word.

Phoneme	Word-Initial	Word-Medial	Word-Final
p	passe	Raupen	rieb
b	Baß	rauben	
t	Tasse	baten	riet
d	das	baden	
k	Kasse	Haken	Sieg
g	Gasse	Hagen	
f	fasse	Höfe	reif
v	was	Löwe	
s	Satin	reißen	Reis
z	Satz	reisen	
š	Schatz	rauschen	Rausch
ž	Genie	Rage	
ç	China	reichen	reich
x		rauchen	Rauch
h	hasse		
m	Masse	hemmen	Damm
n	nasse	Hennen	Mann
η		hängen	Rang
l	lasse	Kohle	will
r	Rasse	bohre	wirr
j	Jacke	Koje	

Table 3.11: German Consonants.

3.2.2.2 Vowels

Tongue	front	center	back
high	i: ü: i ü		u: u
middle	e: ö: e õ ɛ:	ə	o: o
low		a a:	

Table 3.12: Articulation of German Vowels.

There are 19 separate vowel phonemes in German, including the three diphtongs "ai", "oi" and "au". Table 3.12 looks into the articulation of these vow-

els by describing the position of the tongue when articulating the respective phoneme.

Pronunciation examples of German vowels are illustrated in table 3.13, also mainly taken from [Hawkins 1987].

Phoneme	Examples	
i:	bieten	Stiel
i	bitten	Stille
ü:	Güte	fühle
ü	Mütter	fülle
u:	Rute	Ruhm
u	Kutte	Bulle
e:	beten	stehle
e	Betten	Stelle
ö:	Goethe	Höhle
ö	Götter	Hölle
o:	rote	Sohle
o	solle	Bonn
ɛ:	bäte	stähle
ə	gesagt	bitte
a:	rate	Bahn
a	Ratte	falle
ai	leite	Feile
oi	Leute	heule
au	Laute	faule

Table 3.13: Pronunciation Examples of German Vowels.

3.2.3 Morphology

The inflectional morphology of Modern German is very rich compared to other modern Germanic languages and preserves major features of the Old High German system. Compared to the old system, the biggest changes involve the inflectional paradigm for nouns. The system of classification according to the phonology of the stem, which is for example still evident in Russian, was destroyed and new paradigms evolved.

3.2.3.1 Nouns

Nouns are now classified according to their inherent gender, masculine, feminine or neuter, and according to their plural forms. The major plural allomorphs are the suffixes

- "-e" ("Tier(e)" "animal(s)")
- "-er" ("Kind(er)" "child(ren")
- "-en" ("Frau(en)" "woman/(women)")
- "-s" ("Kino(s)" "cinema(s)")
- stem vowel mutation plus "-e" ("Stadt" → "Städte" "city/(cities)").

Also, a number of irregular nouns exist where the major plural allomorphs are concatenated to a word stem that underwent a vowel mutation.

- stem vowel mutation plus "-er" ("Mann" → "Männer" "man/(men)")
- stem vowel mutation alone ("Mutter" → "Mütter" "mother(s)").

A summarization of plural endings can be found in table 3.14 where the regular and irregular nouns already introduced above are summarized.

Singular	Plural	Translation
Tier	Tier-e	animal(s)
Kind	Kind-er	child(ren)
Frau	Frau-en	woman/(women)
Kino	Kino-s	cinema(s)
Stadt	Städt-e	city/(cities)
Mann	Männ-er	man/(men)
Mutter	Mütt-er	mother(s)

Table 3.14: German Plural Endings.

Feminine		Noun	Translation
Singular	Nominative	die Frau	the woman
	Accusative	die Frau	
	Genitive	der Frau	
	Dative	der Frau	
Plural	Nominative	die Frau-en	the women
	Accusative	die Frau-en	
	Genitive	der Frau-en	
	Dative	den Frau-en	

Masculine		Noun	Translation
Singular	Nominative	der Mann	the man
	Accusative	den Mann	
	Genitive	des Mann-es	
	Dative	dem Mann-(e)	
Plural	Nominative	die Männ-er	the men
	Accusative	die Männ-er	
	Genitive	der Männ-er	
	Dative	den Männ-ern	

Neuter		Noun	Translation
Singular	Nominative	das Haus	the house
	Accusative	das Haus	
	Genitive	des Haus-es	
	Dative	dem Haus-(e)	
Plural	Nominative	die Häus-er	the houses
	Accusative	die Häus-er	
	Genitive	der Häus-er	
	Dative	den Häus-ern	

Table 3.15: Examples of German Noun Declensions.

Noun phrases as a whole distinguish separate case inflections for nominative, accusative, genitive and dative in both singular and plural. Sometimes these distinctions are only residually marked on the noun itself and are primarily carried by the preceding determiners and adjectives. Since gender distinctions are inherent in the noun and since plurality is richly marked on the noun itself, the most important function of the determiner is to mark case. The definite article ("the") can therefore assume six forms: "der", "den", "des", "dem", "das" and "die". However, some inflection endings are still exhibited by the noun itself. The genitive singular of most masculine and neuter nouns

shows the suffix "-(e)s", the dative singular of many masculine and neuter nouns has an optional "-e" suffix, and the dative plural of nouns ends with a "-(e)n" suffix. The full set of morphological distinctions carried by the German noun phrase, i.e. gender, number and case, are given in the previous table.

3.2.3.2 Adjectives

Adjectives that follow the definite article show case inflections ending in "-e" or "-en" according to the weak paradigm.

		Noun	**Translation**
Singular	Nominative	der gut-e Mann	the good man
	Accusative	den gut-en Mann	
	Genitive	des gut-en Mann-es	
	Dative	dem gut-en Mann-(e)	
Plural	Nominative	die gut-en Männ-er	the good men
	Accusative	die gut-en Männ-er	
	Genitive	der gut-en Männ-er	
	Dative	den gut-en Männ-ern	

Table 3.16: Examples of German Adjective Inflections (weak).

Strong adjective inflections ("-er", "-en", "-es" "-em", "-e") are practically identical in form and distribution to the forms of the definite article.

		Noun	**Translation**
Singular	Nominative	gut-er Wein	good wine
	Accusative	gut-en Wein	
	Genitive	gut-en Wein-es	
	Dative	gut-em Wein	
Plural	Nominative	gut-e Wein-e	good wines
	Accusative	gut-e Wein-e	
	Genitive	gut-er Wein-e	
	Dative	gut-en Wein-en	

Table 3.17: Examples of German Adjective Inflections (strong).

3.2.3.3 Verbs

Concerning the major morphological distinctions carried out by the verb, German, as all other Germanic languages, distinguishes two basic classes of verbs: weak and strong. Table 3.18 conjugates the verb "sagen" ("to say") as an

example for a weak verb, and the verb "tragen" ("to carry") as an example of a strong one. The strong class undergoes vowel alternations in the stem in addition to taking inflectional affixes for person and number agreement.

	Decomposition	Translation
Infinitive	sag-en	to say
1. Person Singular	ich sag-e	I say
2. Person Singular	Du sag-st	you say
3. Person Singular	er sag-t	he says
1. Person Plural	wir sag-en	we say
2. Person Singular	ihr sag-t	you say
3. Person Singular	sie sag-en	they say

Infinitive	trag-en	to carry
1. Person Singular	ich trag-e	I carry
2. Person Singular	Du träg-st	you carry
3. Person Singular	er träg-t	he carries
1. Person Plural	wir trag-en	we carry
2. Person Singular	ihr trag-t	you carry
3. Person Singular	sie trag-en	they carry

Table 3.18: Examples of German Verb Conjugations.

Although the number of strong verbs is historically on the decline, some of the most common verbs in Modern German are members of a large class of strong verbs: "geben" ("to give"), "essen" ("to eat"), "liegen" ("to lie"), "sehen" ("to see"), "riechen" ("to smell"), "sprechen" ("to speak"), "fahren" ("to travel") and many others. The weak class does not undergo such vowel alternation and takes partially different inflectional affixes for person and number agreement.

	Decomposition	Translation
Infinitive	sag-en	to say
Present Participle	sag-end	
Past Participle	ge-sag-t ge-trag-en	
Imperative (Singular)	sag-(e) !	
Imperative (Plural)	sag-t !	
Imperative (polite)	sag-en Sie !	

Table 3.19: Examples of German Verb Inflections.

The German infinitive marker is an "-en" suffix attached to the stem. The present participle is formed by adding the suffix "-end". The past participle exhibits a "-t" suffix for weak verbs and an "-en" for strong ones. A "ge-" prefix is also added in both cases, if the first syllable of the stem is stressed. It is omitted if the first syllable is not stressed. Also, there are three imperative forms with identical morphologies for strong and weak verbs. Table 3.19 summarizes the inflection endings for infinitive, participles and the imperative.

3.2.4 Vocabulary Growth

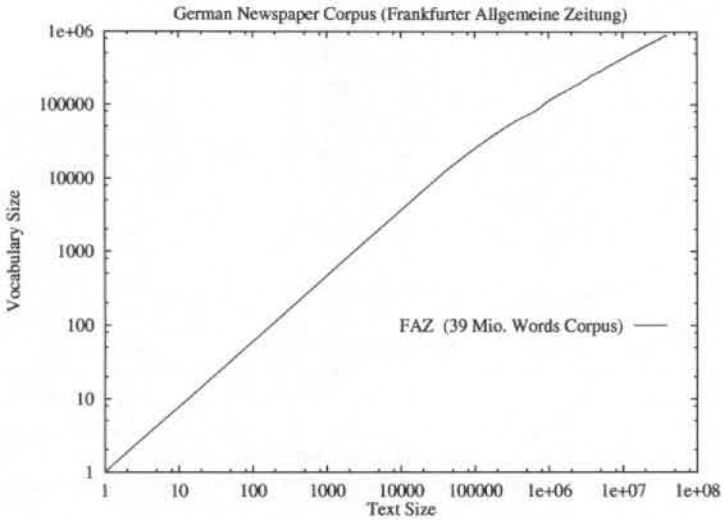

Figure 3.3: German Vocabulary Growth on 39 Mio. Words FAZ (Frankfurter Allgemeine Zeitung).

Just like for the Serbo-Croatian language, the vocabulary growth of German is very fast. This is also due the morphologically rich structure of the language that shows a large number of inflection endings both for nouns as well as for verbs. Unlike Serbo-Croatian, the German language additionally consists of an almost indefinite number of compound words or composita, like for example the always quoted term "Donaudampfschiffahrtsgesellschaftskapitänspatent" that makes one word in German and would constitute six words in English. Figure 3.3 shows the increase of vocabulary entries on a corpus of 39 million

words. This corpus consists of newspaper articles of the German newspaper Frankfurter Allgemeine Zeitung (FAZ). An impressive observation is the fact that the 39 million corpus tokens result in a vocabulary of more than 900,000 words.

Self and cross coverage of this corpus when taking differently sized subsets of the corpus are plotted in figure 3.4. Even with a vocabulary of 300,000 words, the self coverage of this text corpus containing 39 million words is only 98%, meaning that 2% of all words in the text corpus are unknown.

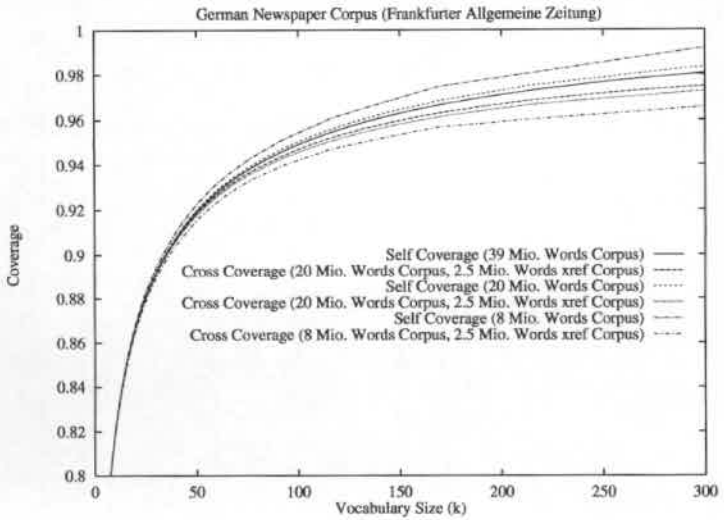

Figure 3.4: German Self and Cross Coverage on 39 Mio. Words FAZ (Frankfurter Allgemeine Zeitung).

Figure 3.5 illustrates the vocabulary growth on an even bigger corpus. This corpus was used for language modeling purposes of the German recognizer described in chapter 4 section 4.2. The already available newspaper texts were enhanced by texts retrieved from the world-wide-web resulting in an overall corpus of 46 million tokens. Also, the cross coverage of a test set used for the automatic transcription of German broadcast news on this corpus is given. For a vocabulary of 61,000 words, as it was used for the German recognizer, the out-of-vocabulary rate is more than 4%. The diagram also shows that for a vocabulary of 500,000 words the number of out-of-vocabulary words would still exceed 1.5%.

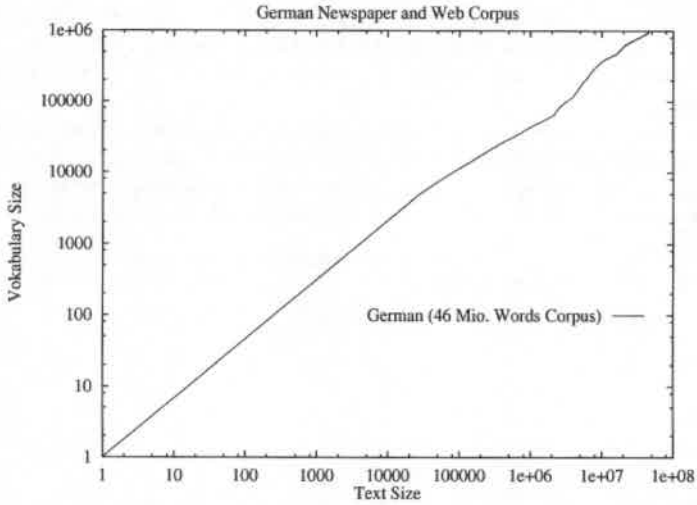

Figure 3.5: German Vocabulary Growth (Newspaper and Web Text Corpus).

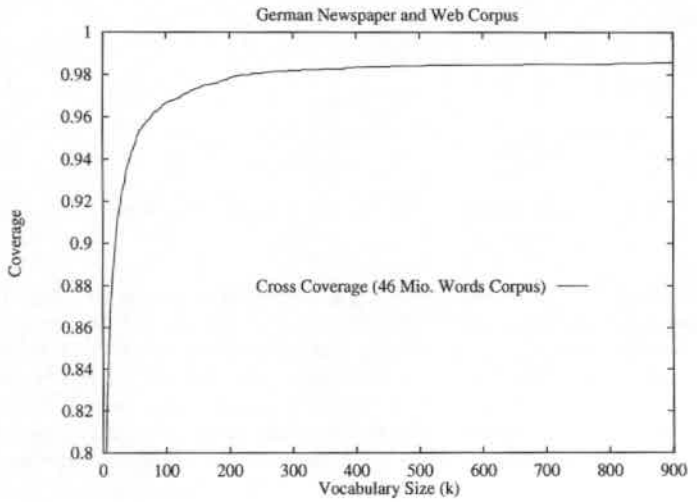

Figure 3.6: German Cross Coverage (Newspaper and Web Text Corpus).

3.3 English

3.3.1 English as a Universal Language

The English language derives from the West Germanic branch of the Indo-European family of languages. It is most closely related to Low German dialects in northern Germany, to Dutch and Frisian. The two standard varieties of English are British and American English, where the existing differences are largely matters of pronunciation and lexicon. In contrast to speech, the standards of writing are very strong and permit surprisingly little variation in grammar, lexicon and orthography.

Whereas Chinese e.g. is spoken by a greater number of people as mother tongue, English is spoken around the globe and has a wider dispersion than any other language. One of the reasons sometimes suggested for the extension of English is the spread of technology. With the expansion of American technologies also the English vocabulary and technical terms disseminated. Certainly its widespread use is due to the simplicity of English inflections and the cosmopolitan character of its vocabulary. Having been introduced to the rich and complicated inflectional structure of Serbo-Croatian and German, the exceptional simplicity of English quickly becomes clear. English inflections are tidy and relatively easy to learn compared to the latter two heavily inflected languages and others that feature complex morphological variations, like e.g. also Russian or Turkish. Another reason often cited for its acceptance around the globe is the breadth of vocabulary of the English language. English is indeed lexically rich, proving this by showing a very large number of synonyms or near synonyms for many words, each suggesting some variation on the semantic core of the word. The fact that its most common words are of such a simple structure certainly also helps explaining the spread of English. 88 of the 100 most frequently written words are monosyllables from "the" ranking first to "down" ranking 100th. Among the next most frequent 1,000 are another 68 monosyllabic. For those among the most frequent 200 words that are not monosyllables, all but five are disyllabic.

It might be clear that languages generally abbreviate words of frequent use in accordance with Zipf's law [Pierce 1961]. English, however, has the additional historical impetus that most disyllabic words ending in an unstressed syllable, became monosyllabic in early Modern English. Also, English is a so-called Subject Verb Object (SVO) language, meaning that subjects precede verbs and verbs precede objects. The claim is that SVO languages are perceptually simpler than languages whose basic orders are SOV or VSO. The perceptual advantage of SVO languages is an easy identification of subjects and objects which are separated by verbs. English tends to have topics in

sentence-initial position, so subject and topic will often coincide.

All things considered, these observations are obvious reasons for the establishment of English as a universal language.

3.3.2 The English Lexicon

Modern English morphology and syntax were established in their current form by about 1400. Looking at the lexical features of the English language, it is very interesting to see that the most frequently occurring words of English are grammatical, i.e. function words like "of" or "the", not lexical or content words like for example "man", "time", "go" or "take". This observation is confirmed by examining the Brown corpus, a standard corpus of present-day English [Kučera & Francis 1967]. The Brown corpus comprises slightly more than a million words and contains about 62,000 different unique word forms, belonging to about 38,000 lemmas. A *Lemma* is defined as a set of word forms, all of which are inflectional or spelling variants of the same base form. Extrapolating these figures to an infinite sample would yield about 170,000 lemmas in English. Remarkably about 2,000 lemmas only, comprising about 2,800 word forms, constitute already 80% of the corpus tokens. But since content words are the least predictable textual elements, knowing these 2,000 lemmas still falls far short of leading to 80% comprehension of an English text.

3.3.3 Phonology

3.3.3.1 Consonants

English phonology consists of 24 consonant phonemes whose word-initial, word-medial and word-final occurrences are illustrated in table 3.20 taken from [Finegan 1987].

Phoneme	Word-Initial	Word-Medial	Word-Final
p	pat	caper	tap
b	bat	labour	tab
t	tap	button	hat
d	dad	ladder	pad
k	cad	sicker	talk
g	gab	dagger	gag
f	file	beefy	thief
v	vile	saving	crave
θ	thin	author	breath
ð	then	weather	breathe
s	sin	mason	kiss

Phoneme	Word-Initial	Word-Medial	Word-Final
z	zebra	posit	pose
š	shame	lashes	push
ž		measure	rouge
č	chin	kitchen	pitch
j	jury	bludgeon	fudge
m	moon	dummy	room
n	noon	sunny	spoon
η		singer	sing
h	hen	ahoy	
y	year	beyond	
r	red	berry	deer
l	lot	silly	mill
w	wind	away	

Table 3.20: English Consonants.

3.3.3.2 Vowels

Concerning vowels, today there exist between 14 and 16 phonemic vowels in different regional varieties of standard English, including the three diphtongs /ay/, /aw/ and /oy/. For a more detailed discussion on English vowels see [Finegan 1987].

3.3.4 Morphology

Old English morphology was considerably more complex than that of Modern English. As a consequence of the extensive phonological reductions and mergers, syncretism of the Old English distinctive inflections occurred and inflectional morphology of Modern English has only eight inflections surviving.

3.3.4.1 Nouns

Compared to Old English where nouns were inflected generally for four cases in the singular and three cases in the plural, English nouns nowadays generally show only two variations: a marked variant for possessive singular (the genitive singular "-s") and all plurals, and an unmarked one for all other functions. Aside from very few exceptions that exploit a functional vowel alternation like "tooth → teeth" or "mouse → mice", or few uninflected plurals like "deer" and "sheep", plurals are formed by adding the common suffix "-s" to the singular.

As for possessives, the rules are identical to those of the plural, except that there are no exceptions.

	Decomposition
Singular	car
Plural	car-s
Genitive Singular	car-'s

Table 3.21: Examples of English Noun Declensions.

3.3.4.2 Adjectives

Furthermore, English exhibits no variation of adjectives regardless of the number, gender or case of the modified noun, like e.g. "the old man", "the old men", "the old woman", or "the old woman's hair". There exists only one form for positive, comparative ("old-er") and superlative ("old-est") degrees, the latter two alternating under specified circumstances with the equivalent analytical forms with "more" and "most" ("more beautiful" and "most beautiful"). The Modern English definite article is formally simple and has only one single orthographic shape: "the" with two standard phonological variants, before vowels and elsewhere.

3.3.4.3 Verbs

Verbs are also only minimally inflected with suffixes for the third person singular concord ("-s"), for the present participle ("-ing"), past tense ("-ed") and past participle (frequently "-en").

	Decomposition
Infinitive	look
3. Person Singular	(he) look-s
Present Participle	look-ing
Past Tense	look-ed
Past Participle	look-ed beat-en

Table 3.22: Examples of English Verb Inflections.

In summary, there are eight productive inflectional suffixes in present-day English: two for nouns, two for adjectives, four on verbs, and no inflectional prefixes or infixes.

3.3.5 Vocabulary Growth

The inflectional simplicity of English compared to highly inflected languages also shows in the speed of vocabulary growth for this language. Figure 3.7 depicts a curve of the vocabulary growth on English broadcast news data. The impressive difference to highly inflected languages becomes even clearer in the direct comparison to Turkish, Serbo-Croatian and German in the next section.

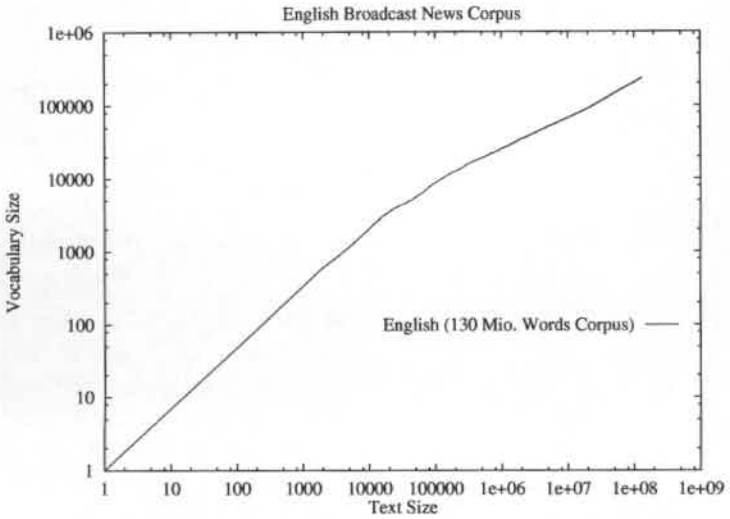

Figure 3.7: English Vocabulary Growth (Broadcast News Corpus).

Self and cross coverage of the English broadcast news data are plotted in figure 3.8. The English broadcast news corpus consists of 130 million tokens and exhibits a vocabulary of almost 250,000 words. For the Serbo-Croatian language a text corpus of 12 million words already results in a vocabulary of more than 300,000 words, in German nearly 950,000 distinct vocabulary entries are found when considering a text consisting of only 46 million tokens.

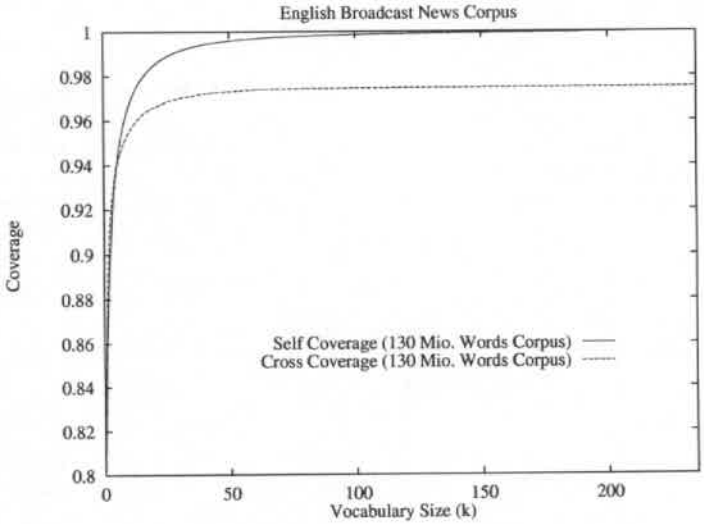

Figure 3.8: English Self and Cross Coverage (Broadcast News Corpus).

3.4 Comparison of Different Languages

A direct comparison of the vocabulary growth for the languages Turkish, Serbo-Croatian, German and English is shown in figure 3.9. The Serbo-Croatian and German data used for this plot is taken from the respective broadcast news databases. The text databases of both tasks contained very few transcripts of actual news broadcasts. A major proportion of the collected text data comprises newspaper texts and texts from newspaper agencies, television and radio stations (for a more detailed description on the used data, see chapter 4). The English plot is based on broadcast news data only. For English a large amount of transcripts for broadcast news shows is available on CD-ROM. All data for the Turkish curve comes from newspaper articles that were collected on the internet.

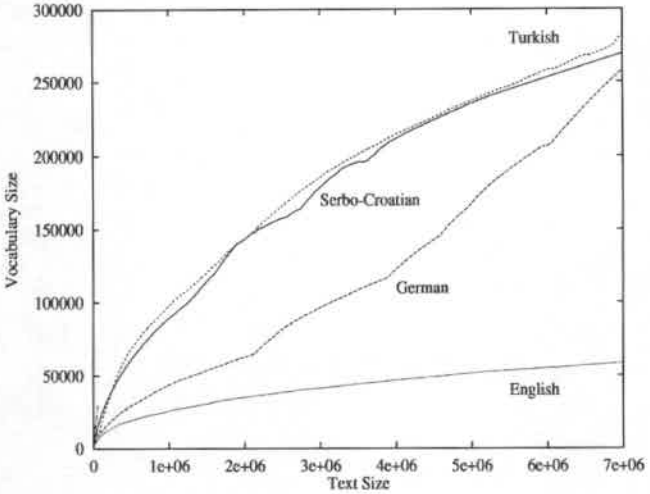

Figure 3.9: Vocabulary Growth on Various Languages.

Chapter 4

Experimental Setup

The speech recognition experiments performed within this work have been conducted either on Serbo-Croatian or German speech data. The following chapter outlines the design, development and training of two speech recognition systems for transcribing Serbo-Croatian and German broadcast news. For both languages the speech and text databases, the training of the speech recognition systems and first baseline results are presented.

4.1 Automatic Transcription of Serbo-Croatian Broadcast News

The following section describes the development of a speech recognition system for transcribing Serbo-Croatian broadcast news [Scheytt et al. 1998]. A Serbo-Croatian large vocabulary system is presented that, despite an only limited amount of acoustic training data, achieves a 30% word error rate when applied to transcribing broadcast news shows. The data collection procedure is introduced, and an overview over the resulting speech and text databases used for building a recognizer is given. Segmentation and labeling of the data according to different acoustic conditions are described, and the results of a first baseline dictation system are presented. This baseline system was used to bootstrap a broadcast news recognition system. Results of this final system on dictation and broadcast news data as well as the influence of language model interpolation on system performance conclude the section.

4.1.1 Introduction

The speech recognition system introduced in this chapter has been realized as part of the Multilingual Informedia Project [Hauptmann & Witbrock 1997a]

at the Carnegie Mellon University (CMU) of Pittsburgh [Scheytt 1997]. It has been designed and developed both for the transcription of broadcast news shows as well as for information retrieval purposes. To this end two different speech recognition engines have been trained: a large vocabulary system that has to be able to handle an almost unlimited vocabulary while transcribing daily news broadcasts, and a much smaller speech recognizer that is able to process spontaneous input queries to an information retrieval system in real time. The queries addressed to the second system are recognized, converted to a textual request and then passed on to the underlying multimedia database of the Multilingual Informedia Project. As a result of the information retrieval process the system then supplies the desired information to the user in form of digital video, audio or textual documents.

However, for the research conducted throughout this work, the large vocabulary system used for transcribing the broadcast news shows before incorporating them into the multimedia database is of much more concern. A reliable speech recognizer is the necessary basis for extracting, structuring and indexing the news data, so that the included information can be easily retrieved again. A major difficulty when building a well-performing large vocabulary system is the fact that Serbo-Croatian is a highly inflected language. It has a very fast growing vocabulary which makes it very difficult to recognize new unknown words with the static vocabularies speech recognition systems normally use. As a result the percentage of out-of-vocabulary words varies from almost 9% up to 22% depending on the size of the used dictionary. This is an additional impediment to the already well-known difficulties of transcribing broadcast news arising from processing spontaneous and ungrammatical speech in noisy environments and unlimited domains.

4.1.2 Serbo-Croatian Databases

Development and training of a large vocabulary conversational speech recognizer always depends on the necessary collection of sufficient databases. For Serbo-Croatian no broadcast news data was available at all, as it is the case for English where a large number of Cable News Network (CNN) broadcasts has already been collected and is available on CD-ROM. In addition, not even other task-unrelated speech or text data was available for the training of a Serbo-Croatian speech recognition system. Therefore sufficient in- and out-of-domain data had to be accumulated for the training of robust acoustic models as well as for the training of reliable language models.

In the course of different projects, two Serbo-Croatian speech databases have been collected by the Interactive Systems Laboratory at the University of Karlsruhe: an 18 hour database consisting of read newspaper speech,

and a total of 18 hours of recorded and transcribed broadcast news shows. The dictation data has been gathered as part of the GlobalPhone project [Schultz et al. 1997]. The broadcast news shows have been collected as training and test material for building a Serbo-Croatian speech recognition system within the Multilingual Informedia Project. For the training of the language model component of the Serbo-Croatian system, the transcription of the collected newspaper articles and news broadcasts were used. In addition, text data from 20 different sources has been collected on the world-wide-web, adding up to a text database of almost 12 million words altogether.

4.1.2.1 Speech Data

The audio data for the first database, the dictation material, was collected in Croatia and Bosnia-Hercegovina. Native speakers were asked to read 20 minutes of news texts extracted from the HRT web site (Croatian Radio and Television) and Obzor Nacional, a Croatian newspaper. The speech was digitally recorded using a portable DAT-recorder at a sampling rate of 48 kHz in stereo quality and further sampled down to 16 kHz with 16 bit resolution in mono quality. The read utterances were checked against the original text to eliminate major errors and mark spontaneous speech effects, like breathing noises, laughter etc. 131 articles were read by 85 speakers resulting in a database of 89,000 text tokens (see table 4.1).

# Speakers	# Articles	Recording Length	# Words
85	131	18 h	89,000

Table 4.1: Serbo-Croatian Dictation System Database.

Since the recorded dictation data is very different to recordings of broadcast news – less or almost no spontaneous effects at all, few or no background noises, uniform recording quality – a second database with domain-specific broadcast news data had to be recorded. This database was also collected at the University of Karlsruhe in Germany. A satellite dish and a dedicated personal computer, equipped with an MPEG encoder board, were installed to record the HRT evening news show. The evening news are transmitted from Croatia via the Eutelsat satellite. The television signal was digitally recorded in MPEG format with a target bit rate of 1.008 Mbit/s, an audio bit rate of 0.192Mbit/s, and a sampling rate of 44.1 kHz. For doing speech recognition the audio signal was uncompressed and sampled down to 16 kHz with 16 bit resolution. As no closed caption was available, transcription of the

news broadcasts was done by native speakers. Beside the television broadcasts, some radio news from the Radio Free Europe / Radio Liberty (RFE/RL) web site in Realaudio (RA) format were downloaded and converted to 16 kHz, 16 bit Wave format for speech processing. In lack of better acoustic data being available this material was initially also used for training but removed from the training set as more suitable within-domain acoustic data was collected. In summary, a total of 34 news broadcasts were made available, comprising 18 hours of speech and 125,000 words (see table 4.2).

Source	Broadcasts	Recording Length	# Words
HRT (MPEG)	27	18 h	118,000
RFE/RL (RA)	7	0.5 h	7,000
Total	34	18.5 h	125,000

Table 4.2: Serbo-Croatian Broadcast News System Database.

Similar to the HUB4 corpus for English broadcast news data, the Serbo-Croatian recordings were divided into segments. Within these segments the acoustic conditions remained constant and each segment was tagged regarding channel quality, background noises and speaker. The various tags used in these three categories are shown in table 4.3, where "Non-Serbo-Croatian" identifies segments of speech in a language other than Serbo-Croatian, most often English. In addition to these acoustic tags only the most frequent and clearly audible spontaneous effects were transcribed: hesitation, breathing, and some other human and non-human noises.

Speaker	Channel	Noise
Male	Clean	Music
Female	Telephone	Second Speaker
Non-Serbo-Croatian	Distorted	Conference
Unknown	Unknown	Street
None		Static Noise
		Other
		None

Table 4.3: Acoustic Segment Tags.

The diacritical letters in Serbo-Croatian were transcribed according to the rules in table 4.4.

Diacritic	Ć ć	Č č	Đ đ	Š š	Ž ž
ASCII	C1 c1	C5 c5	D1 d1	S5 s5	Z5 z5

Table 4.4: Serbo-Croatian Diacritic Rules.

It took about 13 to 18 hours to transcribe a news broadcast of approximately 40 minutes because of:

- no closed caption being available
- high speaking rate in some of the segments
- very noisy segments
- acoustic labeling of the segments.

4.1.2.2 Text Data

In addition to the transcripts of the 27 news shows from the Croatian Radio and Television station HRT, other sources of text data had to be collected to build up a sufficiently large corpus for language modeling purposes. Searching the internet, text data from 20 different sources including television and radio stations, newspapers and news agencies was retrieved. During the postprocessing of these texts one major problem was encountered: many web sites simply map diacritics onto their corresponding non-diacritical letter, e.g. ć and č both become c. In order to build language models based on the web portion of the collected text corpus this mapping had to be automatically inverted.

A statistical approach was used to convert the web texts to usable language model training texts with diacritics. In order to get a reliable conversion, as many Serbo-Croatian texts with diacritics as possible were collected. From these texts a list L_c of correct words was generated. This list served as reference to convert a second list L_f. L_f was extracted from the texts without special characters and contained both correct and false word forms. The used conversion algorithm [Scheytt 1997] works as follows:

- First, all words in L_f that do not contain the letters c, d, s and z are marked as correct.
- In a second step all words which occur in L_f and L_c are labeled as being correct. For some words this might be wrong in certain situations depending on the context. A word trigram model is used to improve the conversion accuracy in such cases.

- In the next step all remaining words in L_f are assigned to their nearest neighbours in L_c. When the Levenshtein editing distance [Manber 1989] does not exceed a certain threshold, the resulting word pair is considered valid and the necessary conversions are performed. When applying this operation to a separate test text only 2% of the words were not converted correctly.

- As last step of the text conversion algorithm a letter trigram model is generated. This model is used to score the likelihood of the different possible character sequences for the remaining words in L_f. By picking the sequence with the highest score the potential diacritic candidates c, d, s and z are switched. In a test text 25% of the words were converted incorrectly using this mechanism. This allows a better conversion than just leaving the words as they are which produces an error rate of 70%.

Thus, finally the combined conversion error rate of the whole algorithm on the test text was 5%. By applying this conversion procedure to all collected web texts without diacritic symbols more than twice the amount of text material than had been available before could be made usable for language model training (see table 4.5).

Character Set	Web Sites	# Words
Diacritics	7	5.5 Mio.
No Diacritics	13	6.5 Mio.
Total	20	12 Mio.

Table 4.5: Serbo-Croatian Internet Text Databases.

4.1.3 Serbo-Croatian Speech Recognition Systems

4.1.3.1 Dictation System

With the objective of building a Serbo-Croatian broadcast news recognizer, first a dictation system for the Serbo-Croatian language was developed. For both applications and domains [Scheytt et al. 1997] the Janus Recognition Toolkit (JRTk) [Finke et al. 1997a] was used. Phone set and a pronunciation dictionary were generated almost automatically since Serbo-Croatian orthography closely matches its pronunciation. As a consequence the phone set corresponds almost exactly to the alphabet and consists of 30 phones, 4 noise and 1 silence models. For a more detailed description of the used phoneme models and the categorization into acoustic and articulatory phoneme classes refer to [Scheytt et al. 1997]. The pronunciation dictionary was created by an

automatic grapheme-to-phoneme tool. Some manual adjustments were necessary for numbers, abbreviations, foreign words and names.

Each allophone is modeled by a left-to-right HMM with mixtures of 16 diagonal Gaussians per state. The preprocessing of the system consists of extracting Mel-frequency cepstral coefficients (MFCCs) every 10 milliseconds. The final feature vector is computed by a truncated Linear Discriminant Analysis (LDA) transformation of a concatenation of MFCCs, and their first and second order derivatives. Vocal Tract Length Normalization (VTLN) and cepstral mean subtraction are used to extenuate speaker and channel differences.

A first context-independent Serbo-Croatian dictation system was trained using the labels generated by a speaker-adapted German recognizer, also called label boosting [Finke & Zeppenfeld 1996]. The Serbo-Croatian phones were initialized by their closest German equivalents. A backoff trigram language model built on the very few available training transcriptions was used. The labels rewritten by the first Serbo-Croatian trained recognizer turned out to be more accurate and were used to train a context-dependent dictation system.

With a vocabulary size of 18,000 words, speaker-dependent VTLN, MLLR adaptation during testing and the use of interpolated language models from different corpora, the initial system performance further improved. Using the dictation engine (D1), a 28.2% word error rate was achieved on a test set of read newspaper articles.

	Vocabulary Size	OOV-Rate[1]	Word Error
Read Data	18,000	8.5%	28.2%
Broadcast News	18,000	22.2%	73.6%

Table 4.6: Initial Results for Dictation and Broadcast News Data using a Dictation System (D1).

4.1.3.2 Broadcast News System

Performance of the broadcast news domain data was first tested on the dictation system already introduced. The test set used for evaluating a speech recognizer on transcribing broadcast news consists of acoustic segments from two news broadcasts. Results reported below correspond to the English partitioned evaluation (PE) test set from the HUB4 evaluation in December 1996 in which the segments and their constant acoustic properties were given for

[1] Out-Of-Vocabulary Rate

training and testing. A first test run on the baseline system described in the section above that had yielded a 28.2% word error rate on dictation data, resulted in 73.6% word error rate on the broadcast news test set (see tables 4.6 and 4.7). This was mainly due to the noisy conditions even in the clean segments of this test set.

The baseline dictation system was then used to label the broadcast news data and train a first recognizer (B0) on only 10 hours of transcribed recordings. This context-dependent system was set up with 2,000 codebooks over 24 input features. The vocabulary size was 29,000, the out-of-vocabulary rate 14.0%.

System	Vocabulary Size	OOV-Rate	Word Error
D1	18,000	22.2%	73.6%
B0	29,000	14.0%	43.6%
B4	31,000	13.6%	36.0%
B5	49,000	8.7%	29.5%

Table 4.7: Recognition Results on Serbo-Croatian Broadcast News.

As the interpolation of different language models had resulted in performance improvements for system D1 on the dictation task (see also [Scheytt et al. 1997] for further details), the same was expected for the training of a broadcast news recognizer. Text corpora from about 20 different sites had been collected and three criteria were applied to divide this text data into different sets: geographical origin (Serbia vs. Croatia), content source (television and radio stations vs. newspaper and news agencies) and language model perplexity. An interpolation of the three resulting text corpora yielded the best recognition results. Compared to using a single language model an absolute improvement of 1.6% word error rate was achieved.

Further improvements were made by a weighted combination of the training data consisting of dictation and broadcast news data, resulting in another Serbo-Croatian speech recognition engine (B4). The size of the dictionary of this system was augmented to 31,000 words which slightly reduced the out-of-vocabulary rate to 13.6%. The performance of the recognizer trained on this data with about twice as many parameters (mixture of Gaussians) as the first broadcast news system B0 was measured to be 36.0% word error rate (see table 4.7).

The final Serbo-Croatian broadcast news recognizer (B5) was trained on 12.5 hours dictation data and 18 hours of transcribed news shows. The context-dependent system is based on 4,000 quinphone models. The preprocessing of

the system consists of extracting an MFCC based feature vector every 10 milliseconds with a window size of 20 milliseconds. The final 32-dimensional feature vector is computed by a truncated LDA transformation of a concatenation of 13 MFCCs, the energy value, their first and second order derivatives, plus zero crossing. Compared to the previous B4 system, two major changes were made:

1. The text material used for language model training was normalized.
2. The vocabulary size of the recognition dictionary was increased from 31,000 to 49,000 entries.

Serbian	Croatian	Translation
reka	rjeka	river
	rijeka	

Table 4.8: Serbian and Croatian Dialectic Variants.

Text normalization, as also reported in [Adda et al. 1997] for French, was performed for the reason already pointed out in chapter 3: in Serbo-Croatian up to three different dialectic variations of one word can be found. The English word "river" for example translates to the Serbian word "reka", but there also exist the Croatian variants "rjeka" and "rijeka" (see table 4.8). When normalizing all available text material the latter two variants are replaced by the first one in all texts (language model corpora, training and test data) and added as pronunciation variants into the dictionary.

Normalization	OOV-Rate
no	10.1%
yes	8.7%

Table 4.9: OOV-Rates with 49,000 Words Vocabulary.

Increasing the vocabulary size from 31,000 to 49,000 words leads to an out-of-vocabulary rate of 10.1% instead of 13.6% on the unnormalized data. As not only the large number of distinct inflection endings of the Serbo-Croatian language but also its numerous dialectic variations of some words are reasons for a rapid vocabulary growth and thus the resulting high out-of-vocabulary rate in Serbo-Croatian, text normalization further reduces the number of out-of-vocabulary words from 10.1% to 8.7% on a vocabulary of 49,000 words. Table

4.10 compares different out-of-vocabulary rates of the Serbo-Croatian broadcast news test set before normalization resulting from an increasing vocabulary size.

Vocabulary Size	OOV-Rate
31,000	13.6%
49,000	10.1%
64,000	9.5%
300,000	4.0%

Table 4.10: OOV-Rates for different Vocabulary Sizes.

As can be seen in table 4.7 the B5 system using a vocabulary of 49,000 words with an out-of-vocabulary rate of 8.7% after text normalization yielded a system performance of 29.5% word error.

4.2 Automatic Transcription of German Broadcast News

This section describes the training and development of a German speech recognition system for transcribing broadcast news [Kemp et al. 1998a]. The resulting large vocabulary system yields a performance of almost 25% word error. After addressing the problems of transcribing German broadcast news that are very similar to the problems in Serbo-Croatian, the data collection process and the resulting German speech and text data material are presented. Finally, a short overview over the used training procedures is given and the performance of the resulting system is introduced.

4.2.1 Introduction

Within a project that aims at the automatic generation of a searchable multilingual database for broadcast news shows [Kemp et al. 1998b], a speech recognition system for transcribing German broadcast news has been developed. When transcribing German news broadcasts similar problems are encountered as those that have already been described for Serbo-Croatian: beside the usual problems of broadcast news recognition, few task-relevant data is available for acoustic and language model training. Especially the number of transcribed broadcast news shows is very limited as this is a tedious and very time-consuming task. Moreover, the German language also shows an excessive

vocabulary growth as every verb can form many different inflection endings and nouns usually appear in at least two different declination forms. Also, new nouns can easily be created by joining one or more independent words together to a new composite one. As a result out-of-vocabulary rates ranging between 4% and more than 9% are observed, depending on the size of the used recognition dictionary.

4.2.2 German Databases

Unlike for Serbo-Croatian, in German plenty of out-of-domain data was available to start with. This included a large quantity of speech data, such as human-to-human conversations, as well as a number of large text databases.

Concerning speech data for the acoustic training of a speech recognition system the available speech material had already been used to train a German recognizer for conversational speech. This well-trained state-of-the-art German large vocabulary conversational speech recognition system [Finke et al. 1997b] provided robust acoustic models and served as baseline system for the development of a German broadcast news transcription system. As a consequence no out-of-domain speech data had to be collected, like it had been done in Serbo-Croatian by training a first baseline system on read newspaper texts. To have task-specific speech data for test and training purposes in addition, news broadcasts of a German television station were collected.

Concerning text data, out-of-domain material is not as interesting for language model training as it is for the acoustic component of a speech recognizer. Some data closely related to the broadcast news task, several volumes of a German newspaper, was available on CD-ROM. Following the Serbo-Croatian model, the necessary additional task- and domain-specific data material could easily be retrieved on the world-wide-web or from the archives of German newspapers. In the end a text database of 46 million words with nearly 950,000 distinct vocabulary entries, consisting of newspaper articles and the transcripts of several broadcast news shows, was available for the training of the language model component.

4.2.2.1 Speech Data

To be able to further train or at least to adapt the existing conversational system, 12 news broadcasts covering three hours of speech data were collected within the same scenario as the Serbo-Croatian data (see table 4.11). At the University of Karlsruhe a satellite dish and a dedicated personal computer, equipped with an MPEG encoder board, were installed to record the "Tagesschau". This evening news show is transmitted by one of the public television

stations via satellite. The television signal was digitally recorded in MPEG format with a target bit rate of 1.008 Mbit/s, an audio bit rate of 0.192Mbit/s, and a sampling rate of 44.1 kHz. For doing speech recognition the audio signal was uncompressed and sampled down to 16 kHz with 16 bit resolution. Closed caption was available only for the parts of the news broadcasts where the anchor speaker is involved. It was used as basis for manual corrections of the transcripts of the anchor speaker. Trancriptions of the foreign correspondent reports within the news broadcasts were done by native speakers. Segmentation of the speech data was manually done according to the acoustic conditions of the audio signal: clean speech from anchor speaker, speech with all kinds of background noise (street noise, war etc.), other speakers in background, telephone speech etc. Unlike for Serbo-Croatian and the F-conditions used in DARPA evaluations for transcribing English broadcast news [Woodland et al. 1998], only two different labels for the segments were used: "clean" and "distorted". The label "clean" corresponds to the F0-condition (= anchor speaker) whereas "distorted" marks everything else.

Source	Broadcasts	Recording Length	# Words
Tagesschau	12	3 h	118,000

Table 4.11: German Broadcast News System Database.

4.2.2.2 Text Data

Four out of the 12 recorded broadcast news shows were used as development and evaluation test set. Thus, only the transcripts of the remaining eight news broadcasts could be used for language modeling purposes. Similar to the development of the Serbo-Croatian speech recognition system, for German also other sources of text material had to be found in order to train a language model with reliable probability estimates. As some volumes of the "Frankfurter Allgemeine Zeitung" (FAZ), a German newspaper, were available on CD-ROM, in the beginning only this data was used to interpolate with the language model built on broadcast news. Then, following the example of the Serbo-Croatian data collection, the internet was searched and text data from several sites (television stations, news agencies and radio stations) was collected. "Germany Live" is an internet newspaper that offers daily news reports. Bayrischer Rundfunk 5 (BR5) is a German radio station in southern Germany that distributes transcriptions of its daily news broadcasts over the internet. Finally, the "Tagesschau" part of the corpus contains transcriptions

of the anchor speaker part of the daily broadcasted "Tagesschau", and also closed caption from another news television show, the "Tagesthemen".

Source	# Articles
Germany Live	56,883
BR5	8,488
Tagesschau	5,683
Total	71,054

Table 4.12: German Internet Text Databases.

A total of 71,000 articles could be retrieved in the beginning. Data collection continued and together with the 39 million words that were already available from the "Frankfurter Allgemeine Zeitung", finally the total database for language model training consisted of 46 million words containing almost 950,000 unique vocabulary entries.

4.2.3 German Speech Recognition Systems

4.2.3.1 Conversational Speech Recognition System

A speech recognition system applied to the recognition of conversational speech was used as baseline system for the development of a speech recognizer to transcribe German broadcast news shows. The Janus Recognition Toolkit (JRTk) based recognizer was trained on a collection of spontaneous human-to-human dialogues. A more detailed description of the system can be found in [Finke et al. 1997b].

4.2.3.2 Broadcast News System

The speech recognition system developed for transcribing broadcast news shows is also based on JRTk. Eight news shows were used to train the context-dependent acoustic models of the broadcast news recognizer. Fully continuous mixtures of Gaussian densities were clustered by growing a decision tree employing linguistic questions with respect to the immediate phonetic context, i.e. no pentaphones as in the Serbo-Croatian system but triphones instead. All mixtures were chosen to have 30 Gaussians, Gaussians are modeled with diagonal covariances.

In the preprocessing stage 13 cepstral parameters per 16 milliseconds frame are computed from 30 melscale filter bank coefficients. The frame shift is set to 10 milliseconds. A simple energy based speech detection is performed

and cepstral mean subtraction and variance normalization is applied to the speech segments only. The 13-dimensional cepstral vector is concatenated to its delta and delta-delta coefficients (first and second order derivatives) to form a 39-dimensional intermediate feature vector. The intermediate vector is then transformed by linear discriminant analysis (LDA) into one 16-dimensional feature vector. All experiments reported here are based on a vocal tract length normalized system where the power spectrum is warped to a reference vocal tract length before the computation of the melscale filter bank coefficients takes place.

First tests (see also table 4.13) were conducted on a baseline system (I14) using a recognition dictionary of 17,000 entries, yielding an out-of-vocabulary rate of 9.3%. For the subsequent systems the size of the dictionary was increased to up to almost 61,000 entries thereby also decreasing the resulting out-of-vocabulary rate on our test set.

System	Vocabulary Size	OOV-Rate	Word Error
I14	17,000	9.3%	32.2%
I23	61,000	4.4%	24.7%

Table 4.13: Recognition Results on German Broadcast News.

The recognition dictionary of our latest I23 system consists of the 60,784 most frequent words derived from the language modeling corpus described above. The out-of-vocabulary rate is 4.4%. For all but the 2,000 most frequent words pronunciation variants were discarded. Thus, the resulting recognition dictionary has 61,685 entries in total. The phonetic alphabet to transcribe these entries consists of 44 base phones which were derived from the SAMPA phone set. In order to account for spontaneous effects such as hesitations and noises as observed in broadcast news data, five special noise models were added to the set of phones: two hesitation models, one model for breathing noise, one model for various other human noises such as glottal noises, and one model accounting for all other noises. All these noise models were treated as context-independent models. A standard Kneser Ney backoff trigram language model based on 46 million words of newspaper texts and radio broadcast transcriptions as described above is used.

All recognition results reported are based on the following test procedure: a first three-pass decoding run yields a set of transcriptions using the unadapted acoustic models. These transcriptions are used to compute three MLLR matrices to adapt the mean vectors of the acoustic models. The final transcriptions are generated by running the three-pass decoder using the adapted acoustic

models. Two broadcast news shows were used as test material yielding the performances presented in table 4.13. For a vocabulary size of 61,000 words and an out-of-vocabulary rate of 4.4.%, a word error rate of 24.7% could be achieved.

Chapter 5

Morphology-Based Speech Recognition

When doing speech recognition on highly inflected languages, the usage of smaller base units than words for the recognition process is one possibility to counteract the fast vocabulary growth and thereby also decrease the number of out-of-vocabulary words. This chapter reports on several decomposition methods applied to the German and Serbo-Croatian language. All methods are either based on linguistic knowledge about the morphology of the respective language or employ clustering techniques that are based on similarity measures. Examples of the different decompositions are given and the recognition results that were achieved when conducting speech recognition experiments on these new base units are presented.

5.1 Introduction

Looking at the lexical properties of highly inflected languages as described in chapter 3, and introducing the resulting vocabulary growth graphs and out-of-vocabulary rates has already shown how difficult it is to guarantee unrestricted natural language processing. Huge dictionaries that increase search space and result in performance degradations are required. Moreover, even when using these huge dictionaries there still remains a non-negligible number of out-of-vocabulary words that automatically lead to recognition errors and worsen performance.

To overcome this problem of new and unknown words, especially in languages like German and Serbo-Croatian, an obvious solution seems to be the usage of smaller base units than the notion of words. Smaller units reduce the number of distinct dictionary entries, thereby simultaneously decreasing the number of out-of-vocabulary words. The reduction in out-of-vocabulary

rate results from being able to compose new unseen words out of several parts already known to the dictionary.

The question of what the optimal base unit looks like is difficult to answer. A first attempt is to use units that are clearly defined on a linguistic level [Pelz 1996]. Looking at the German language as an example, the next smaller unit after using words would be the application of morphemes. Within this work a morpheme is defined according to the morphological concepts and properties of languages presented in chapter 3. Following this definition of morphology a word consists of a root form or word stem, and possible inflection endings. The word "Frauen" ("women") would be decomposed into its word stem "Frau" and the declination ending "-en". The same applies to verbs where the word "kommen" ("to come") would be split into the morphological stem "komm-" and the conjugation ending "-en". A next step would be to break morphemes further down to syllable or even phoneme level. The smaller the size of the base units recognition is conducted on, the more new and unseen words can of course be formed from an already existing vocabulary of a fixed size. On the other hand, the smaller the recognition units are, the more difficult it gets for a speech recognition system to recognize them correctly. Very small base units are more challenging both for the acoustic as well as the language model component of a recognizer. Small units tempt the acoustic component to erroneously insert these short segments frequently. As for the language model, to take the same amount of context into consideration as it would be the case with a word-based model, for smaller units a much longer-range language model than the traditional trigram has to be used.

Both German and Serbo-Croatian are used as examples for highly inflected languages throughout this chapter. For the decomposition experiments described in the following two sections linguistically based morphemes and morpheme-like units automatically determined through clustering techniques have been chosen as base units for the recognition process. This size of the recognition units was chosen as a compromise between the conventional model of words and even smaller units, such as syllables or phonemes. The recognition experiments performed in both languages were conducted on two different tasks and databases: a speech recognition system able to handle conversational speech between two human beings scheduling a meeting in German, and the transcription of broadcast news for Serbo-Croatian.

5.2 Morphology-Based Approaches for the German Language

The motivation for using a different unit of speech than the conventional word-based model within the recognition process for German is caused by the unusually fast vocabulary growth in this language. This is due to both a high number of conjugation and declination endings for verbs and nouns, as well as the feature of compound words that is special to the German language. By decomposing words into smaller units and using these units throughout the recognition process a further growth should be contained.

5.2.1 The Spontaneous Scheduling Task Database

All data that has been used for the German experiments on morphology-based speech recognition is taken from dialogues of the German Spontaneous Scheduling Task (GSST). For comparisons with a language showing a much slower vocabulary growth English data has been examined. This data is also part of the corresponding English Spontaneous Scheduling Task (ESST). The Spontaneous Scheduling Task consists of human-to-human dialogues recorded at different sites in Germany and the United States of America with various recording scenarios. Goal of every conversation is to arrange a meeting of two people within their given schedules.

	Training	**Testing**
# Dialogues	225	25
# Utterances	5,629	378
# Words	117,489	7,803
Vocabulary Size	3,821	735

Table 5.1: Training and Test Material for the German Spontaneous Scheduling Task (GSST).

To be able to assess the influence of inflection endings and composita on the vocabulary growth of the two languages, dialogues from both tasks, English and German, were used for comparison. In English a total of 160 dialogues had been collected. They were divided into a training set of 146 dialogues consisting of 1,395 utterances, and 14 test conversations. For German a total of 250 dialogues from 4 different sites was available for training and testing. 225 of them were used for acoustic training and the training of an overall language model, the rest of 25 was used for testing and evaluation purposes.

Recognition experiments were performed with the speech recognition engine of the JANUS-2 system [Suhm et al. 1995]. Table 5.1 gives a more detailed description of the German training and test material.

5.2.2 Comparison of the English and German Language

The idea of using morpheme-based recognition units both for the acoustic as well as the training of the language model is triggered by the observation that the German language differs from other languages by an outstanding number of distinct inflection endings. When for example looking at the word "kommen" in German ("to come" in English) the difference of both languages becomes clear: whereas in English the conjugation of this regular verb consists of simply 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 5.2 illustrates the differences in German and English for this example.

German		English	
ich	komm-e	I	come
Du	komm-st	you	come
er/sie/es	komm-t	he/she/it	come-s
wir	komm-en	we	come
ihr	komm-t	you	come
sie	komm-en	they	come

Table 5.2: Examples of Inflection Endings for German Verbs.

Besides, the German language has an uncountable number of compound words. Formation of compounds is not only possible for nouns but also for verbs. Several prefixes can be attached to every verb, each time creating a new word.

German	English
hinein-gehen	to go in
aus-gehen	to go out
weg-gehen	to go away
.....	.....

Table 5.3: Examples of German Verb Prefixes.

The same applies to noun composita. Nouns can be concatenated to long noun chains, every chain creating a word with a new meaning, e.g.:

German	English
Sprach-erkennung-s-modul	speech recognition module
Sprach-erkennung-s-genauigkeit	speech recognition accuracy

Table 5.4: Examples of German Compound Words.

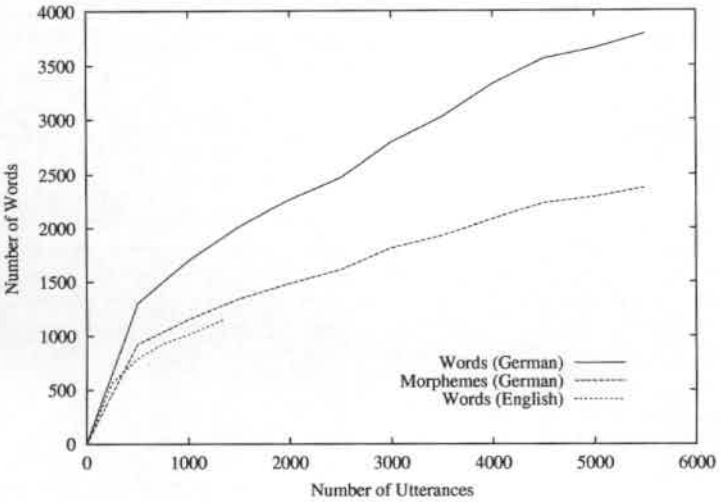

Figure 5.1: Vocabulary Growth of Words and Morphemes in German and English.

Naturally these three characteristics of German lead to a much faster vocabulary growth when the amount of training data increases than it is the case in English. Even for the relatively small and, compared to the newspaper and broadcast news databases already introduced in chapter 3, limited Spontaneous Scheduling Task an increasing number of training dialogues still results in a steady vocabulary growth with no saturation to be expected. An impressive demonstration of this fact is given in figure 5.1 which shows the increase of the English vocabulary compared to the growth of vocabulary words

on the German database. In English 1,395 utterances result in a vocabulary of 1,169 words. After the same number of utterances in German the number of distinct vocabulary entries even increases to 1,971 words which amounts to 168% of the English vocabulary.

But even with a steadily growing dictionary not all out-of-vocabulary words that might appear in the recognition process can be foreseen. Tables 5.5 and 5.6 both show the vocabulary coverage of the German and the English test text measured on different amounts of training utterances and their resulting vocabulary sizes. The smaller English vocabulary already covers 92% of English vocabulary words in the test dialogues whereas the fourfold amount of training data in German only covers 88%. As a logical consequence it is desirable to work on smaller base recognition units than words to be able to compose new unseen words out of several parts already known to the dictionary and thereby decrease the percentage of words unknown to the recognizer.

#Utterances	Vocabulary Size (Words)	Coverage	Vocabulary Size (Morphemes)	Coverage
500	1,301	65%	925	72%
1,000	1,696	70%	1,151	76%
1,500	2,015	75%	1,344	82%
2,000	2,271	78%	1,485	84%
2,500	2,468	79%	1,612	85%
3,000	2,793	81%	1,814	87%
3,500	3,032	83%	1,930	88%
4,000	3,331	85%	2,087	89%
4,500	3,563	86%	2,236	90%
5,000	3,658	87%	2,293	90%
5,500	3,791	88%	2,376	91%
5,629	3,821	88%	2,391	91%

Table 5.5: Vocabulary Coverage (German).

#Utterances	Vocabulary Size (Words)	Coverage
500	791	87%
1,000	1,013	91%
1,395	1,169	92%

Table 5.6: Vocabulary Coverage (English).

5.2.3 Several Decomposition Methods

When using morphemes instead of the traditional word units in speech recognition the following advantages arise: for highly inflected languages the large vocabulary growth with increasing training material is limited, thereby automatically decreasing the rate of out-of-vocabulary words. Also, by using morpheme-based n-gram language models instead of the traditional word-based n-grams more robust probability estimates for small training databases are possible. Considering the lexical properties special for the German language, morpheme decomposition can be performed on three different levels:

1. Decomposition can be done strictly morpheme-based, meaning that the rules that are used for decomposing are based on linguistic knowledge and the decomposition strictly follows these linguistic rules. For example the word "weggehen" ("to go away") is decomposed into the prefix "weg-", the word stem "geh-" and the inflection ending "-en":

   - weggehen → weg-geh-en (*to go away*)
   - Dialoge → Dialog-e (*dialogues*)
   - Spracherkennung → Sprach-er-kenn-ung (*speech recognition*)

2. Decomposition can be done by reducing words to just their root forms, meaning that for both verbs and nouns the inflection endings are deleted:

   - weggehen → weggeh (*to go away*)
   - Dialoge → Dialog (*dialogues*)

3. Decomposition can be done by combining strictly morpheme-based decomposition with the reduction to root forms:

   - weggehen → weg-geh (*to go away*)
   - Dialoge → Dialog (*dialogues*)

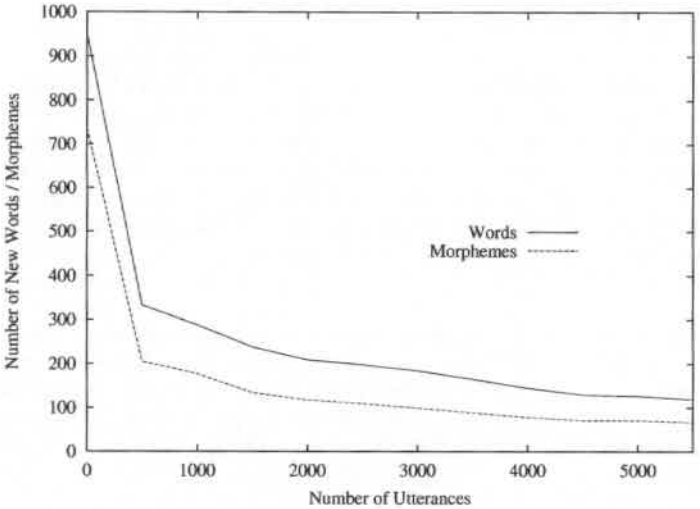

Figure 5.2: Reduction of New Words in German.

For the German Spontaneous Scheduling Task (GSST) the decomposition of training texts into strictly linguistically based morphemes results in a reduction of the vocabulary size by 37% (see figure 5.1). Whereas the word dictionary of the used database contains 3,821 words, the corresponding morpheme dictionary consists of only 2,391 entries (see table 5.7). This reduction would certainly have been even larger if more data had been available. In addition, the number of new unknown words or rather morphemes within the vocabulary of the test set decreases much faster when using morphemes instead of words (see figure 5.2). As it can be seen in table 5.5 the test set coverage of strictly linguistically decomposed morphemes almost approaches the 92% coverage found in the English language on word level. It exceeds the German word coverage of 88% by 3% based on the same training data. Thus, the original rate of 12% new vocabulary entries on word basis is decreased to a share of 9% when using morpheme-based recognition units.

	Words	Morphemes
#Tokens	117,489	146,990
Vocabulary Size	3,821	2,391

Table 5.7: Comparing Word and Morpheme Vocabulary.

Comparing the number of tokens for the respective word and morpheme databases in table 5.7, it can be seen that on the average one word becomes 1.25 tokens within the morpheme-based framework. As the immense reduction in vocabulary size would have anticipated a much higher average decomposition rate, this figure seems very low. It also suggests that a much higher number of words than expected consists of a unique word stem only, like for example names and places, and is not decomposed at all.

5.2.4 Morphology-Based Recognition

Recognition performance was tested on the conventional word-based speech recognizer as well as three recognition systems applying the decomposition methods described above. The experiments were conducted with the JANUS-2 speech recognition system. All available 225 training dialogues were used for building several overall language models: one was trained on word level, three other models were built on the respective morphemes or root forms resulting from the different decomposition methods presented in the previous section. Smoothing was done by absolute discounting in all cases [Ney & Essen 1993]. The acoustic training of the speech recognizer was performed on less dialogues than the training for the language models, resulting in an acoustic dictionary of 3,062 words (see also table 5.8). Both for the experiments on a word-based recognizer as well as for the morphology-based tests the same acoustic models trained on word units were used.

The baseline recognition result is 35.3% word error rate achieved on word basis with a conventional word bigram model. In this experiment the test set contained 9% new vocabulary entries regarding the training vocabulary, meaning that the vocabulary coverage of the test text was 91%. In comparison the percentage of words of running text within the test set that was not included in the training vocabulary, the out-of-vocabulary rate, amounted to 12%. If experiments are performed on speech data that contains words unknown to the dictionary of the used speech recognition system, this is called an *open-vocabulary* scenario. The *closed-vocabulary* case, a more or less artificial creation for tasks and languages with fast vocabulary growth, is met when every word in the test data is already included in the recognition dictionary and

thus the out-of-vocabulary rate is 0%. As a desirable baseline, word accuracy was also tested on a closed-vocabulary scenario yielding a word error rate of 33.1% for a dictionary size of 3,085 words.

5.2.4.1 Morpheme-based Decomposition

Even though the restriction of dictionary growth is highest when using a strictly linguistic-based decomposition of words (see experiment MORPH1 in table 5.8), recognition results are degrading compared to the word-based performance. Whereas the language model profits from a very small unit decomposition by more reliable probability estimates, it also loses important knowledge about the context of an utterance by only being able to consider a much smaller range of speech. To take into account the same context as a word bigram model, language models trained on smaller base units would have to consider a much wider range than bigrams do. Also, the acoustic part of the JANUS-2 speech recognizer suffers as expected from these small components, as they tend to be erroneously inserted frequently. Hence a more balanced way of decomposition has to be found which limits the vocabulary growth but also guarantees improved recognition performance.

	Vocabulary Size	**Word Error**	**Speed Acceleration**
Word Bigram Model (closed-vocabulary)	3,085	33.1%	–
Word Bigram Model (open-vocabulary)	3,062	35.3%	–
Morpheme Bigram Model (MORPH1) (open-vocabulary)	1,946	35.5%	$\approx$ 30%
Morpheme Bigram Model (MORPH2) (open-vocabulary)	2,204	34.6%	$\approx$ 30%
Morpheme Bigram Model (MORPH2) (trigram rescoring)	2,204	34.2%	–

Table 5.8: German Recognition Results for Morpheme-Based Decomposition.

The results of the experiment above suggested to create a second, not strictly linguistically oriented decomposition (MORPH2). In this experiment

basically almost only compound words are decomposed into their components. Very short inflection endings, especially endings that consist of only one letter or phoneme are not cut off. By only considering morphemes that have a certain minimum length the resulting vocabulary reduction is of course smaller than experienced before for the first decomposition. Nevertheless the achieved recognition results are slightly better than for word-based recognition assuming an open-vocabulary scenario. Pure morpheme-based recognition measured on word basis slightly outperforms the result achieved with word bigram models by 0.7%. When applying trigram rescoring to the so far best performing morpheme-based speech recognition system, thereby allowing the language model to use a larger context, the word error rate is reduced even further. Table 5.8 shows that an improvement of 1.1% absolute is achieved, resulting in an overall performance of 34.2% word error rate. As the vocabulary size of the acoustic dictionary used within the recognition process is much smaller than on word basis, in addition the recognition speed is accelerated by more than one third.

5.2.4.2 Root Form Decomposition

Another way of reducing the number of different vocabulary words and building stronger language models is the decomposition of words into their root forms (ROOT). In this experiment all words of the same word stem but with different suffixes are reduced to their root form resulting in a vocabulary of 3,205 words instead of the original 3,821 words[1]. This means a 16% reduction in the size of the vocabulary used as basis for language modeling. However, root forms cannot be used as units of the acoustic dictionary since the suffixes of all inflections also have to be recognized acoustically. Therefore the acoustic dictionary used as search dictionary for the recognizer has to consist of all 3,062 full forms. Consequently recognition speed is the same as for the baseline system using words as recognition units. During the recognition process the full forms are mapped to their root forms, and this information is passed on to the language model module, as it is shown in figure 5.3.

[1]Note that of course this only applies to the language model vocabulary, the acoustic dictionary still has to contain full form words.

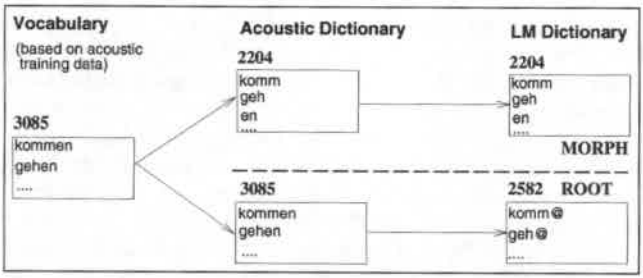

Figure 5.3: Mapping of Acoustic and Language Modeling Dictionaries during Recognition Process.

When determining the achieved performance of this experiment one has to be careful as it is not possible to measure real word accuracy but only root form accuracy. Accuracy results achieved through decomposition in root forms cannot directly be compared to word accuracy results achieved so far. The recognition result of this method for the open-vocabulary scenario, measured in root form error instead of word error rate, is given below in table 5.9.

	Vocabulary Size	Root Form Error
Simulation of Root Form Decomposition (Words) (closed-vocabulary)	3,085	31.9%
Simulation of Root Form Decomposition (Words) (open-vocabulary)	3,062	33.8%
Root Form Bigram Model (ROOT) (open-vocabulary)	3,062	36.5%
Combined Bigram Model (COMB) (open-vocabulary)	2,998	34.9%

Table 5.9: Recognition Results (Root Forms).

However, this experiment is not meant to be judged on word basis but on the basis of root form accuracy. Of course word accuracy could also be measured but would always stay below comparable experiments where the

language model does not generalize over root forms by ignoring different suffixes. A reason for this inferior behaviour is that the differentiation between the recognition of distinct word suffixes then would have to done by the acoustic component alone when determining word accuracy through this method. However, when using a speech recognition engine as input component for a speech-to-speech translation system [Geutner et al. 1996], recognition of full word forms is not a necessary prerequisite for successful end-to-end performance. As the recognizer output is supposed to be input into a semantic-based parser, good recognition of root forms would be sufficient for the following parsing process [Ward 1991] leaving good translation accuracy untouched. For comparison of root form and word-based recognition, the root form error rate of the open-vocabulary word bigram model case has been measured: 33.8%. The corresponding experiment for root form decomposition results in an error rate of 36.5%. Obviously a full form word-based language model better supports the recognition process than a root form based one.

5.2.4.3 Combination

For the final experiment both already described methods MORPH2 and ROOT were combined (COMB). The achieved performance outperforms the root form only language model, but with 34.9% root form error rate still stays below our assumed baseline of 33.8%.

5.3 Morphology-Based Approaches for the Serbo-Croatian Language

As the Serbo-Croatian language shows a vocabulary growth even worse than it is found in German, a first approach here is also to limit this expansion and thus the number of possible out-of-vocabulary words by using morpheme-based recognition. Unlike German this fast vocabulary growth is only due to a large number of distinct inflection endings as the notion of real composita does not exist in Serbo-Croatian. All experiments described here were conducted on the Serbo-Croatian speech recognition system and the broadcast news database that has previously been described in chapter 4.

5.3.1 Several Decomposition Methods

When switching from word-based speech recognition to the recognition on smaller base units for the Serbo-Croatian language, the same reasons concerning inflection endings apply as for German. As a consequence the application of

decomposition methods that allow to conduct recognition on morpheme-based units offers the same advantages and disadvantages as it does in German. For Serbo-Croatian two very different approaches were used to decompose words into their morphological components: the first one follows the strictly linguistically based example already introduced in the previous section. The second one applies clustering techniques to identify typical inflection endings that are then used to decompose words into stems and suffixes:

1. Decomposition is done by using morphological knowledge about grammatical suffixes of the Serbo-Croatian language. This approach is strictly linguistically based like the MORPH1 approach for German used in the previous section:
   - baviti → bavi-ti
     (*to deal with*)
   - razgovorima → razgovor-ima
     (*(through the) conversations*)
2. Decomposition into word stems and suffixes is done by determining typical inflection endings through the clustering of all words into similarity classes:
   - baviti → bavit-i
     (*to deal with*)
   - razgovorima → razgov-orima
     (*(through the) conversations*)

5.3.1.1 Linguistically Based Morpheme Decomposition

The Serbo-Croatian language does not have the notion of composita. Thus the immense vocabulary growth is determined by the various conjugations of verbs and declensions of nouns. Splitting up the Serbo-Croatian vocabulary into word stems and suffix endings is therefore done by using the grammatical knowledge and rules that are used to build inflection endings of regular verbs and nouns. By using this information a list of the 60 most common grammatical suffixes for noun, adjective and verb endings was collected. All words of the vocabulary are reduced to their word stems by chopping off these endings. This algorithm is applied to all text material available for the Serbo-Croatian language. Decomposing the corpus that consists of almost 12 million words and a vocabulary of 300,000 distinct tokens, the resulting number of word stems is 220,000. Both, the resulting word stems and the previously determined suffixes are then used as new base units of the vocabulary for the recognition process. Table 5.10 gives an example of the used suffixes for the Serbo-Croatian language.

Suffix Length	Suffixes
6	itijim ijskoga ijskome stvenu stvima ...
5	enima evima ijama nogim ovske ...
4	able ajte enom itom nove novo stvi stvo ...
3	aju ale eve evi hom hov **ima** imo iti ljo lju oga oma sko ...
2	aj an ga ha im iš ji ju me mo om oh oj **ti** ...
1	a e g h i j m u ...

Table 5.10: Examples for Serbo-Croatian Suffixes based on Linguistic Knowledge.

5.3.1.2 Similarity-Based Morpheme Decomposition

When using a similarity-based criterion for determining possible suffix endings, all words of the available text databases are clustered into similarity classes. This is done by calculating the Levenshtein editing distance between them. For each word or cluster the closest neighbour is computed and the classes are expanded recursively by using bottom-up clustering. The algorithm terminates when a certain distance threshold is exceeded. After termination the longest common prefix for each of the resulting classes is determined. This prefix corresponds to the notion of a word stem when using linguistic knowledge for the decomposition process. Through this procedure a list of one prefix and several suffixes is gained. This global prefix and suffix list is then used to form a new morpheme dictionary for the subsequent recognition experiment. Table 5.11 shows the suffixes that were gathered through similarity clustering.

Suffix Length	Suffixes
6	jači jbolja ...
5	čević jedan kinom **orima** rlane tniji tnike ...
4	aoko cija djan ijek isto ...
3	ako eva ića ima jem mog nim sto tri uša vić ...
2	al an ar ić ih ma oj om ri su ...
1	a e g h **i** j m u ...

Table 5.11: Examples for Serbo-Croatian Suffixes based on Similarity Measures.

5.3.2 Morphology-Based Recognition

The morpheme-based recognition experiments conducted on the Serbo-Croatian database all were performed within the JRTk speech recognition framework. The acoustic models that had originally been trained on the basis of words were not retrained but used as is. For the recognition experiments trigram language models were used. In total three different models were trained: one was built on words as base units, the other two on units according to the decomposition methods presented in the previous section. The baseline recognition result for a system with a vocabulary size of 31,000 words yielded a word error rate of 44.9%. The out-of-vocabulary rate measured on word level was 13.6% for this experiment.

	Vocabulary Size	OOV-Rate (Word-Based)	Morpheme Error	Word Error
Baseline (Word-Based)	31,000	13.6%	34.9%	44.9%
Morpheme-Based (MORPH)	17,000	7.5%	39.5%	53.3%
Morpheme-Based (MORPH)	31,000	5.5%	36.6%	51.4%
Morpheme-Based (SIMILARITY)	17,000	3.7%	51.5%	59.0%

Table 5.12: Serbo-Croatian Recognition Results for Morpheme-Based Decomposition.

For using linguistic knowledge to decompose the vocabulary words into morphemes two different dictionary sizes were investigated: 17,000 and 31,000, the same size of vocabulary that was used before for the word-based recognition experiment. As can be seen in table 5.12 the linguistically based approach works better than the similarity-based one. While the reduction in out-of-vocabulary rate is much higher for the latter, the grammatically derived decomposition rules seem to suggest a more natural way of splitting words into morpheme components. Recognition results of the linguistically based method are significantly better than the performance of the similarity-based decomposition. This implies that both the acoustic as well as the language model component of the recognition engine are able to deal much better with this kind of decomposition. However, the word-based approach, even with the much higher out-of-vocabulary rate, outperforms both approaches clearly. Main reasons are certainly the missing retraining on morpheme units instead using the acoustic models trained on word basis as well as the loss of context that is inherent in the use of smaller base units for language modeling.

5.4 Conclusions

Experiments on both languages have proven that the decomposition of words into morpheme-based units is able to limit the vocabulary growth and also decrease the number of words unknown to the recognizer. However, recognition performance in general worsens compared to a word-based approach. This is due to a large number of hypothesized morpheme concatenations that do not always map to legal words. A very small improvement in word accuracy is possible only with a very specialized and restricted decomposition method on the German data where a certain minimum length is established for the parts a word is decomposed into. The generally higher error rate of the performed recognition experiments is certainly due to the missing retraining of acoustic models on morpheme basis. Context-dependent polyphone models would become more accurate after retraining, particularly when they contain word boundary tags. The second reason is the loss of context information inherent in the usage of morpheme language models. Whereas the word-based recognizer is able to utilize a group of three words when applying a trigram model throughout the recognition process, considering the same context in a morpheme-based system might take a sequence of only two words into account. This especially applies to the Serbo-Croatian language where, unlike in German, most words that can be decomposed are split into only two parts: word stem and inflection ending. For German the loss of context for a strictly morpheme-based decomposition might be even worse as a lot of words can be decomposed into more than two separate parts. When applying a trigram language model on morpheme basis, the context in between words cannot be considered by the language model anymore. The integration of 4-gram or even 5-gram language models could compensate this drawback, but those long-range language models are often difficult to integrate into the search engine of a recognizer. Also, these models usually do not provide very robust probability estimates as the frequencies of seen word sequences of length four or five are very low.

Chapter 6

Hypothesis Driven Lexical Adaptation (HDLA)

The idea of the previous chapter was to reduce high out-of-vocabulary rates by switching the units of recognition from words to smaller base units, namely morphemes. Instead of a dictionary of words the underlying recognition lexicon consists of subword units. The coverage of such a dictionary in terms of word units by these subword units or concatenations thereof, is significantly better than the coverage of a dictionary of words of the same size. However, the recognizers built on top of these units suffer a severe degradation in performance because of those hypothesized morpheme concatenations that do not map to legal words. The decrease in performance is also due to the effect that the context in terms of the language model cannot be handled appropriately using a statistical n-gram model.

In this chapter an alternative approach is taken: words are still considered as the semantic content bearing units of recognition. But instead of having a static dictionary of those words, the concept of a dynamic dictionary is introduced which has the same fixed size as the static dictionary but is tailored on the fly to each specific utterance to be recognized. Thus, with each utterance having its own customized dictionary the size of the recognition dictionary is virtually unlimited. The following sections will introduce the motivation behind this idea, describe the developed **H**ypothesis **D**riven **L**exical **A**daptation (HDLA) algorithm in detail, and give an overview over the various selection criteria that can be used within the framework of HDLA.

6.1 Motivation

The main reason for the large number of out-of-vocabulary words in Serbo-Croatian and German is the high variety of inflection endings that can be

found in both languages. In addition, the German language allows to form composite words from several independent nouns which also accelerates the speed of vocabulary growth compared to languages like, for example, English. Moreover, independent of the language the task of transcribing broadcast news shows an unproportionally high number of new words that denote names and places, also called named entities. Consequently the three main reasons for the high out-of-vocabulary rates when transcribing Serbo-Croatian and German broadcast news are:

- a large variety of inflection endings
- composita
- named entities

- REF: gegen *** **SOZIALE** ungerechtigkeiten haben sich heute die beiden großen christlichen kirchen in IHREN OSTER botschaften gewandt der ratsvorsitzende der evangelischen kirche in deutschland engelhardt erklärte *** OHNE den blick für das leid der menschen sei die soziale marktwirtschaft nicht tragfähig EIGENNUTZ SO engelhardt sei zwar EINE ******* TRIEBKRAFT der marktwirtschaft er dürfe aber nicht *** ********** IN ZERSTÖRERISCHEN EGOISMUS AUSARTEN das oberhaupt der deutschen katholiken bischof lehmann verlangte aufrichtigkeit und wahrheit UM rücksichtslosigkeit und menschenverachtung zu ÜBERWINDEN

  HYP: gegen MIT **SOZIALER** ungerechtigkeiten haben sich heute die beiden großen christlichen kirchen in IHREM GRUSSWORT botschaften gewandt der ratsvorsitzende der evangelischen kirche in deutschland engelhardt erklärte UND IN den blick für das leid der menschen sei die soziale marktwirtschaft nicht tragfähig EINEN RÜCKZIEHER engelhardt sei zwar EIN BETRIEB KRAFT der marktwirtschaft er dürfe aber nicht INS HERSTELLER ICH NIE BUDDHISMUS AUSMACHTEN das oberhaupt der deutschen katholiken bischof lehmann verlangte aufrichtigkeit und wahrheit UND rücksichtslosigkeit und menschenverachtung zu ÜBERDENKEN

- REF: das FLÜCHTLINGSDRAMA ***** in der adria hat ein juristisches nachspiel wegen der **SCHIFFSKOLLISION** am karfreitag abend ermittelt jetzt die staatsanwaltschaft in *** **BRINDISI** *** der kapitän des albanischen ***** ***** **MOTORSCHIFFES** DAS mit einer unbekannten anzahl von menschen gesunken war wurde festgenommen der vorwurf **** FLUCHTHILFE AUS GEWINNSUCHT außerdem WAR DAS SCHIFF offenbar für EINE ÜBERFAHRT nach italien gar nicht geeignet GEGEN DEN kommandanten der italienischen KORVETTE wird wegen des verdachts der fahrlässigen tötung ermittelt DAS **KRIEGSSCHIFF** **** wurde vorerst beschlagnahmt

  HYP: das FLÜCHTLINGS DRAMA in der adria hat ein juristisches nachspiel wegen der **SCHIEDSKOMMISSION** am karfreitag abend ermittelt jetzt die staatsanwaltschaft in **DEM DIE SIE** der kapitän des albanischen **MOTOR STEHT FEST** DASS mit einer unbekannten anzahl von menschen gesunken war wurde festgenommen der vorwurf FÜR SCHÄFER AUSGEGEBEN SUCHT außerdem *** BEIDES CHEF offenbar für EINEN ÜBERFALL nach italien gar nicht geeignet DIENEN DEM kommandanten der italienischen KORREKTE wird wegen des verdachts der fahrlässigen tötung ermittelt DASS **KRIEGS CHEF** wurde vorerst beschlagnahmt

Figure 6.1: Example Alignments of German Utterances.

This observation is confirmed by performing an error analysis both of the Serbo-Croatian and German recognition results presented in chapter 4. Figure 6.1 presents a typical alignment of hypothesized speech output and reference text for the German broadcast news recognition system. "soziale" ("social") is an example for a new word of which the word stem with a different ending, "sozialer" ("social"), was already contained in the dictionary and therefore was recognized instead. "Schiffskollision" ("ship collision") and "Motorschiffes" ("motor vessel") are occurrences of out-of-vocabulary words where either one acoustically similar word, "Schiedskommission" ("arbitration committee"), or a sequence of phonetically similar words, "Motor steht fest" ("motor stands still"), was recognized instead. "Brindisi" denotes a new place that is not in the recognition dictionary and is hypothesized by a sequence of phonetically similar words, "dem die sie" ("him the she"), that do not make any sense at all. The hypothesized words "Kriegs Chef" ("war chief") are an excellent

example where a new composite word, "Kriegsschiff" ("warship"), could not be recognized as a whole because its individual parts only were contained in the recognition dictionary.

When looking at an error analysis of the Serbo-Croatian recognition results, even more examples of misrecognitions concerning only the morphological suffix of a word are found. Figure 6.2 shows the alignment of a Serbo-Croatian broadcast news utterance and the recognized hypothesis as example. As can be seen almost half of the misrecognized words are new word forms of words that are already included in the dictionary with a different inflection ending: "hrvatskim" vs. "hrvatski", "potpisale" vs. "potpisali", "namera" vs. "namerava", "hrvatskom" vs. "hrvatskome" etc. For these errors only the inflection ending is wrong, but the word stem of the respective word was recognized correctly.

- REF: američke kompanije koje su SA **HRVATSKIM** partnerima **POTPISALE** pisma **NAMERA** O gradnji prometnica i termoelektrana smatraju kako hrvatska strana **** PRESPORO privodi kraju pregovore o tim zajedničkim projektima premijer mateša moraće ukloniti sumnje kako se na **HRVATSKOM** tržištu SUŽAVA prostor za američku robu iz vošingtona za **SREDIŠNJI DNEVNIK** željko PORPAR

  HYP: američke kompanije koje su ZA **HRVATSKI** partnerima **POTPISALI** pisma **NAMERAVA** * gradnji prometnica i termoelektrana smatraju kako hrvatska strana PRED SPORO privodi kraju pregovore o tim zajedničkim projektima premijer mateša moraće ukloniti sumnje kako se na **HRVATSKOME** tržištu SLUŽILA prostor za američku robu iz washingtona za **SREDIŠNJEM DNEVNIKU** željko KORPAR

Figure 6.2: Example Alignment of a Serbo-Croatian Utterance.

All alignments show that the output of the recognizer is not random at places where the word to be recognized is not in the dictionary, thus constituting an out-of-vocabulary word. Instead, together with the knowledge of the special lexical characteristics of both languages three different types of misrecognition errors due to out-of-vocabulary words can be identified:

- Errors where only the inflection ending of the misrecognized word is wrong, but the word stem is recognized correctly. This is due to an

out-of-vocabulary word where this new word constitutes an unseen word form of a word already included in the recognition dictionary.

Example: "soziale" → "sozialer"

- Errors where a sequence of words or word stems is hypothesized instead of a correct recognition of the actually uttered compound word itself. This kind of error occurs when the individual parts of a composite word are included in the recognition dictionary whereas the compound itself is not.

  Example: "Kriegsschiff" → "Kriegs Schiff"

- Errors where a phonetically similar word or several phonetically similar words that are already included in the dictionary are recognized instead of the uttered out-of-vocabulary word or fragments thereof. Misrecognitions like these usually involve new unseen content words and, to a very high percentage, named entities. Especially new places and names not included in the recognition dictionary are replaced by hypothesizing one or more phonetically similar words.

  Example: "Kriegsschiff" → "Kriegs Chef"
  "Brindisi" → "dem die sie"

Concerning the problem of morphology and composita, chapter 5 has shown that recognition performed on smaller base units than words is able to counteract the rapid vocabulary growth of highly inflected languages but cannot improve recognition performance. The combination of arbitrary morphemic affixes by way of a morphemic language model leads to an overgeneration of illegally inflected word hypotheses and thus increases error rates.

When using conventional word-based recognition units instead, the size of a static dictionary will always be too small to incorporate all words needed for the recognition procedure. Consequently the speech segment to be recognized will always contain some words unknown to the recognizer. This is the case even for very large sizes of a recognition dictionary, especially when dealing with highly inflected languages. It also applies to tasks like the recognition of broadcast news that will always come up with new vocabulary entries denoting so far unknown names and places. Looking at the application of transcribing Serbo-Croatian broadcast news, even in case of using a lexicon including 300,000 entries, the percentage of unknown words in the test set would still be 4%.

As this is a dictionary size that is not practical for the speech recognition toolkit used throughout this work, an out-of-vocabulary rate of 8.7% while using a recognition vocabulary of size 49,000 words is a more realistic assumption. When dealing with tasks like the transcription of broadcast news, this

number represents an order of magnitude a speech recognizer has to be able to deal with and still keep up a decent recognition performance.

Thus, two obvious solutions to solve the problem of large quantities of out-of-vocabulary words are not practical at all:

1. The usage of other, smaller base units than words throughout the recognition process.

2. An interminable increase of the size of the recognition dictionary.

As a consequence the idea of a dynamic expansion of the search dictionary during the recognition process has been investigated. Following this idea the size of the vocabulary used within the recognition process is still considered finite but virtually allows for a larger number of words to be recognized. This is done by not using a fixed dictionary for all speech utterances but exchanging the vocabulary entries of the used recognition dictionary 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 rather hypothesized words. The second performs another recognition run on the adapted vocabulary that has a lower out-of-vocabulary rate, and will result in a better recognition performance. The dictionary used for both recognition runs stays the same fixed size, only the 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 the limitation of the dictionary to a certain size N is overcome. Therefore techniques and algorithms have to be found that are able to dynamically adapt the dictionary, thus decreasing the number of words unknown to the recognizer. The following sections describe the algorithm developed within this thesis that attains this goal by taking advantage of the three major types of misrecognitions due to out-of-vocabulary words presented above. The knowledge about morphological and phonetical affinity of actually uttered and hypothesized words is incorporated into an adaptation procedure that allows speech recognition on a virtually unlimited vocabulary. 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, thereby also improving recognition performance. Especially when transcribing broadcast news, this should keep the out-of-vocabulary rate limited and thus improve the word error rate.

6.2 The HDLA-Algorithm

The observations described above are the foundation for the development of an adaptation procedure for the recognition dictionary. They suggest that the speech recognition engine itself might be employed to compile a dictionary that is most utterance-specific and thus has fewer out-of-vocabulary words. A significant reduction of out-of-vocabulary words is the main idea of the following approach.

A hypothesis as generated by a recognizer does, even though errorful due to out-of-vocabulary words, represent valuable information about the missing words. In case of missing inflections, typically words with the same stem are substituted in the hypothesis. Also, phonetically similar words are favoured for new words by the recognizer. Considering these two facts, the idea is to start off from a first recognition hypothesis and use this for building a new dictionary on the fly. This customized dictionary contains all words that showed up in the hypothesis augmented by those words from a much larger background lexicon that are most "similar" to the hypothesized words. The lexicon used as background or *fallback lexicon* potentially contains an order of magnitude more words than the recognition lexicon. Decoding the very same utterance a second time using this new utterance-specific dictionary hopefully yields a better accuracy due to a potentially lower out-of-vocabulary rate. The underlying assumption is that by keeping the words of the hypothesis as elements of the new and adapted utterance-specific dictionary a second recognition run should be able to recover the original accuracy at least.

To achieve this goal, a first baseline recognition run on the general broadcast news task-specific dictionary is performed. A domain-specific dictionary of size N normally consists of the N most relevant words for this task. Usually "relevance" is defined by calculating the frequency estimates for all words found in the largest available text database that contains data relevant for the task in question. Hypothesized speech output of this first baseline recognition pass can be generated in form of a hypothesis representing the most likely uttered sentence. A second, much richer representation of the search space as pruned by the recognizer is provided by word graphs or *word lattices*. When using a lattice as output of the JRTk speech recognition system, all possibly uttered word sequences for a certain speech input are stored in this data structure. The lattice begins with a symbol for "beginning of utterance" and terminates in a common end point also called "end of utterance". In between these two fixed points the recognizer does not output a single word for each part of the spoken input. Quite the reverse, at most places there will be branchings as the recognizer is not totally certain what word actually has been uttered. At these branchings several words will occur that, with a certain probability estimate,

are all likely to be uttered at this point of time. Figure 6.3 shows an example of a typical lattice output of the JRTk speech recognizer. Using a word lattice as output of the speech recognition process, N-Best lists of the N most likely uttered sentences can easily be created. Both lattices as well as these lists might then be used for rescoring purposes or as input into a postprocessing component of a speech recognizer, e.g. parsing the recognition output.

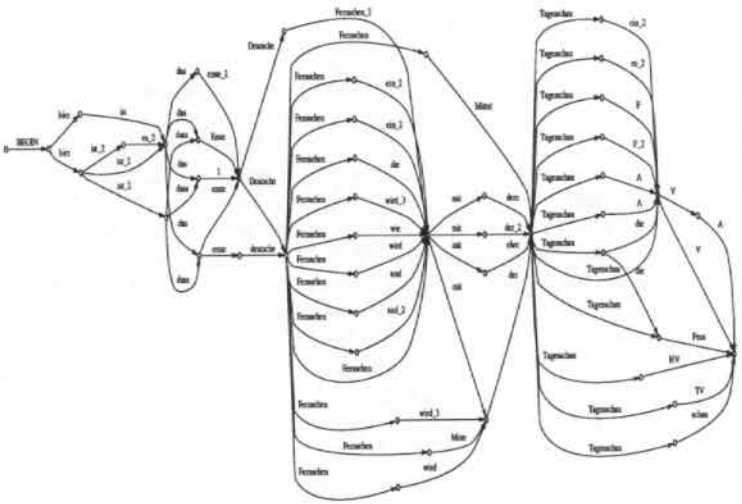

Figure 6.3: Word Lattice of the JRTk Speech Recognizer.

Thus for our purposes, instead of considering the first-best hypothesis only as starting point to the vocabulary selection procedure, the lattice output of the first recognition pass is a much better candidate. The lattice created in this first recognition run with the general task-specific dictionary is then used to determine which words have actually been uttered most likely in the input sentence, namely all words represented in the lattice. For each utterance to be recognized this lattice leads to an utterance-specific vocabulary. When counting the number of different words within a lattice the size of this vocabulary is very small compared to the size of the recognition dictionary used for creating it. In the experiments conducted within this work a lattice for one of the broadcast news speech segments contained at most 10% of the original vocabulary, in most cases the size was even smaller than that.

Based on this vocabulary list a new customized vocabulary for the speech segment in question has to be found. Looking at the misrecognition errors that were made by the recognizer, it seems appropriate to look for words that are "similar" to the ones hypothesized in the lattice, but were not included in the recognition dictionary of the first recognition run. Similarity in a very general sense is, based on the observations of the previous section, defined through the following definition. Two words are similar, if

1. they vary only in their inflection endings from each other, but begin with the same stem (*morphological similarity*), or
2. they are phonetically similar to each other (*phonological* or *acoustic similarity*).

To find similar words that were not contained in the dictionary of the first recognition run a much bigger lexicon is used for comparison purposes. Therefore all vocabulary entries that were found in the largest available in- and out-of-domain database used for the respective task are collected in a large lexicon, the fallback lexicon. Usually the language model training text is used for this purpose. Not only are all words seen in the text database included in this lexicon, but it is also annotated with the frequencies of the words within this corpus. Based on this fallback lexicon and the generated word list the dynamically adapted vocabulary for the second recognition run is determined. The different criteria for selecting new vocabulary entries from the fallback lexicon that are similar to already recognized words are summarized in the next section. A more detailed description of each of the methods as well as experimental results for all selection criteria applied to real recognition tasks can be found in the following four chapters.

After a new adapted vocabulary has been created, the recognition dictionary for the second recognition pass and a new language model are automatically created. A second recognition run with much lower out-of-vocabulary rates is then performed on this adapted system, resulting in an improved recognition performance.

The algorithm below describes in detail all steps of the whole **Hypothesis Driven Lexical Adaptation** (HDLA) process [Geutner et al. 1997a] [Geutner et al. 1998c]:

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 used to look up all similar words in the fallback lexicon consisting of all words that were observed in the largest available text corpus.

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 remains N.

4. In an automatic procedure a new dictionary and language model are created to perform a second recognition run.

Figure 6.4 illustrates the two-pass lexical adaptation procedure of the HDLA framework applied to the recognition of a German broadcast news show, the "Tagesschau". Applied to Serbo-Croatian and German broadcast news data the HDLA algorithm yields significant improvements both in out-of-vocabulary as well as in word error rate.

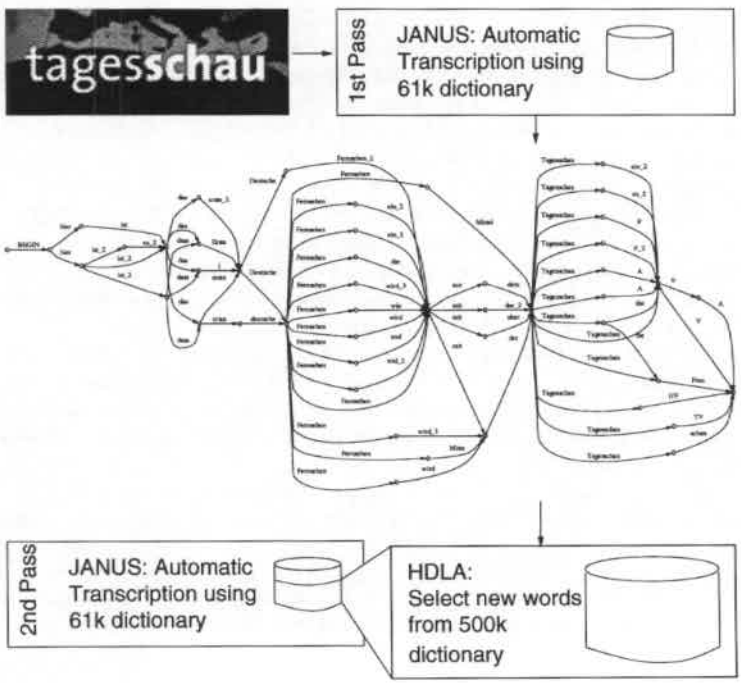

Figure 6.4: Two-Pass Recognition and Lexical Adaptation based on Lattices.

6.3 Different Selection Criteria

As already introduced various criteria for selecting the adapted vocabulary can be applied [Geutner et al. 1999]. Figure 6.5 summarizes the ideas and methods that have been used as selection criteria within the Hypothesis Driven Lexical Adaptation procedure:

1. Linguistic knowledge about morphology and inflection endings.

   The morphology-based approach developed for lexical adaptation is described in chapter 7, where two words are considered similar if they share the same word stem and only differ in the inflection ending.

2. Distance-based measures on grapheme or phoneme level.

   Triggered by the idea of phonetic similarity, chapter 8 introduces various distance measures that are either based on the letter sequence of words (grapheme-based) or their phoneme sequence (phone-based). For the phone-based approach three different methods of calculating the phonetic distance between two words will be presented. Also the notion of composita is taken into account when determining word distances.

3. A combination of a phonetic distance measure with the usage of an artificially created fallback lexicon.

   If no large database is available for a certain language, language-specific morphological rules for generating inflection endings can be defined to artificially create a fallback lexicon. Chapter 9 describes the approach where this man-made lexicon is used to determine the phonetic distance of word pairs.

4. World-wide-web-based retrieval.

   The usage of information retrieval techniques for creating a customized dictionary is presented in chapter 10. Two approaches are experimented with: one employs a search engine to retrieve texts similar to the hypothesized output of the first recognition run, the other uses the topicality of a news show to retrieve similar texts.

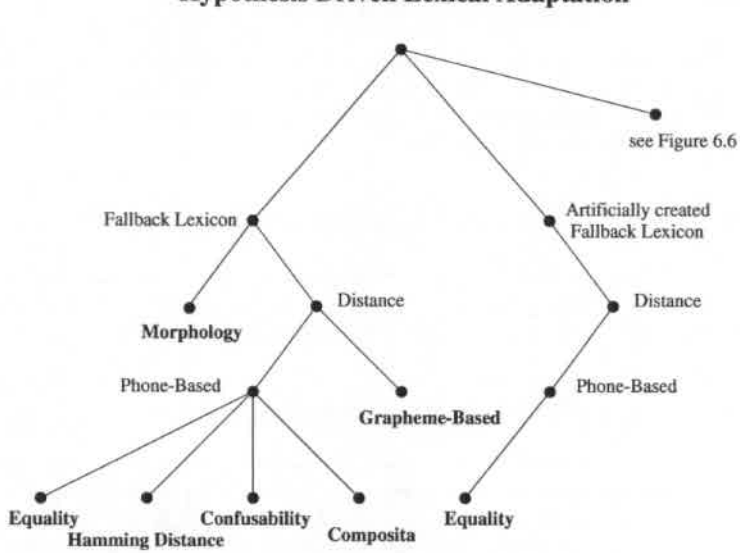

Figure 6.5: Overview over all HDLA Methods.

Whereas the first two methods rely on the availability of a large fallback lexicon obtained through a huge text database for the language in question, the usage of an artificially created fallback lexicon only depends on linguistic knowledge about the respective language. Depending on the special characteristics of the respective language HDLA is performed on, different procedures lead to an optimal performance, as the summarization of all results in chapter 11 will illustrate.

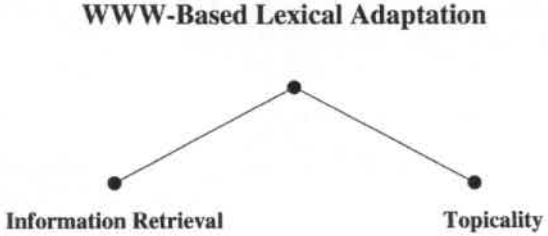

Figure 6.6: WWW-Based Lexical Adaptation.

Based on the ideas presented above algorithms to select the customized vocabulary for the second recognition run have been implemented. More detailed descriptions of the four approaches will be presented in the following chapters.

6.4 Experimental Setup

Baseline systems used for all results that will be reported in the following four chapters are the Serbo-Croatian [Scheytt et al. 1998] and German broadcast news recognizers introduced in chapter 4.

Chapter 7

Morphology-Based Lexical Adaptation

Within the framework of the HDLA algorithm introduced in the previous chapter, linguistic knowledge about morphology and inflection endings can be used to dynamically adapt the recognition dictionary to the utterance to be recognized. Morphological similarity is then used as selection criterion for the adaptation process. This chapter describes a morphology-based approach for lexical adaptation and presents recognition results on Serbo-Croatian and German broadcast news data.

7.1 Motivation

The idea to use linguistic information about word stems and inflection endings as similarity criterion based on morphological knowledge for the HDLA procedure was triggered by a thorough error analysis performed on the baseline results of our Serbo-Croatian and German broadcast news recognizers. As an exemplary occurrence of misrecognitions where only the inflection ending is wrong but the word stem of the respective word was recognized correctly, a typical alignment of a Serbo-Croatian broadcast news utterance and the hypothesized speech output is shown in Figure 7.1. About half of the misrecognized words are new word forms of words that are already included in the dictionary with a different inflection ending: so "hrvatski" was recognized instead of "hrvatskim", "potpisali" instead of "potpisale", and "namerava" was erroneously hypothesized for the out-of-vocabulary word "namera". Taking this observation into account of course suggests the usage of linguistic knowledge about morphology and inflection endings to adapt the recognition dictionary to the utterance to be recognized.

- REF: americ5ke kompanije koje su SA **HRVATSKIM** partnerima **POTPISALE** pisma **NAMERA** O gradnji prometnica i termoelektrana smatraju kako hrvatska strana **** PRESPORO privodi kraju pregovore o tim zajednic5kim projektima premijer mates5a moracle ukloniti sumnje kako se na **HRVATSKOM** trz5is5tu SUZ5AVA prostor za americ5ku robu iz vos5ingtona za **SREDIS5NJI DNEVNIK** z5eljko PORPAR

  HYP: americ5ke kompanije koje su ZA **HRVATSKI** partnerima **POTPISALI** pisma **NAMERAVA** * gradnji prometnica i termoelektrana smatraju kako hrvatska strana PRED SPORO privodi kraju pregovore o tim zajednic5kim projektima premijer mates5a moracle ukloniti sumnje kako se na **HRVATSKOME** trz5is5tu SLUZ5ILA prostor za americ5ku robu iz washingtona za **SREDIS5NJEM DNEVNIKU** z5eljko KORPAR

Figure 7.1: Alignment of a Serbo-Croatian Utterance.

As a consequence, knowledge about the inflectional structure and the grammatical rules that apply to noun declensions and verb conjugations were used to create a list of frequently occurring suffixes. This was done for the Serbo-Croatian language as well as for German.

Word	Decomposition	Translation	
žena	žen -a	woman	
ženo	žen -o	woman!	(vocative)
ženu	žen -u	(the) woman	(accusative)
žene	žen -e	woman's	(genitive)
ženi	žen -i	(to the) woman	(dative)
ženom	žen -om	(with the) woman	
govoriti	govor -iti	to speak	
govorim	govor -im	I speak	
govoriš	govor -iš	you speak	(singular)
govori	govor -i	he speaks	
govorimo	govor -imo	we speak	
govorite	govor -ite	you speak	(plural)
govore	govor -e	they speak	

Table 7.1: Examples of Serbo-Croatian Morphology.

Table 7.1 presents typical inflections of the female noun "žena" ("woman") and the verb "govoriti" ("to speak") as examples for Serbo-Croatian morphology. Typical German inflection endings of various nouns and the verb "gehen" ("to go") are found in table 7.2.

Word	Decomposition	Translation
Wahrheit	Wahr -heit	truth
Schwierigkeit	Schwierig -keit	difficulty
Kinder	Kind -er	children
Kindern	Kind -ern	children
gehen	geh -en	to go
(ich) gehe	geh -e	(I) go
(Du) gehst	geh -st	(you) go
(er) geht	geh -t	(he) goes

Table 7.2: Examples of German Morphology.

A subset of frequently occurring Serbo-Croatian and German suffixes are summarized in table 7.3 for the Serbo-Croatian language. In table 7.4 the same is done for German. Both lists of suffixes were applied in the subsequent experiments on morphology-based lexical adaptation for the two languages. They were used to split words into their word stems and inflection endings, in order to decide if two words consist of the same word stem and only differ in their suffix.

Suffix Length	Suffixes
6	itijim ijskoga ijskome stvenu stvima .....
5	enima evima ijama nogim ovske .....
4	able ajte enom itom nove novo stvi stvo ...
3	aju ale eve evi hom hov ima imo iti ljo lju oga oma sko ...
2	aj an ga ha im iš ji ju me mo om oh oj ti ....
1	a e g h i j m u ...

Table 7.3: Examples of Serbo-Croatian Suffixes.

Suffix Length	Suffixes
4	keit lich
3	ern und ung
2	en er es in st
1	e n r s t

Table 7.4: Examples of German Suffixes.

7.2 Usage as Selection Criterion

For the lexical adaptation of the recognition vocabulary based on morphological knowledge word stem equivalence is used as similarity criterion. The basic algorithm of chapter 6 can be applied exactly as introduced, using morphological similarity as selection criterion for the adaptation of the vocabulary.

1. A first recognition run is performed on the general baseline recognition dictionary and word lattices are generated for each speech utterance to be recognized. Each of these word lattices is then converted into an utterance-specific vocabulary list.

Step 2 of the algorithm, the determination of words similar to words contained in this list, in more detail looks as follows for the morphology-based approach:

2. **a)** The vocabulary list derived from the word lattice of the first recognition run is split into word stems and suffixes.

   (For the experiments performed on Serbo-Croatian and German data different combinations of word stem and suffix lengths were tested. Note that the word stem length had to be at least two letters long.)

2. **b)** The resulting word stem list is then used to look up all words with the same stem in the fallback lexicon consisting of all words that were observed in the largest available text corpus.

The following steps of the algorithm are then performed independent of the used selection criterion as described in chapter 6:

3. All words that have the same word stem as one of the words of the utterance-specific vocabulary list are incorporated into the original baseline recognition vocabulary by replacing the least relevant words that were not hypothesized in the first recognition pass. Through this procedure a new adapted lexicon for the second recognition run of the same size as the original lexicon is made available.

4. Finally, phonetic transcriptions for the new dictionary are created and a new language model is generated. Both are used in a second recognition run where the out-of-vocabulary rate for each speech segment is reduced.

As already mentioned above, different combinations of word stem and suffix lengths were tested. The optimal minimum and maximum lengths for Serbo-Croatian and German are described in the following sections. For both languages the minimum word stem length was bound to two letters, as one letter can hardly be considered as a valid word stem.

7.3 Results on Serbo-Croatian Data

The suffix list from table 7.3 was used for experiments on our Serbo-Croatian broadcast news system. The first recognition run was done on the domain-specific baseline dictionary using the N most frequent words from the 12 million words text corpus. Experiments were performed both on the B4 and B5 system of chapter 4, meaning that for system B4 N equalled 31,000 words, and a size of 49,000 words for system B5.

Different minimum lengths were used for the word stem and also suffix lengths from one to four were experimented with. For Serbo-Croatian the best performance could be achieved for a word stem length of five and the consideration of suffixes up to the length of four.

	Minimum Word Stem Length				
Suffix Length	2	3	4	5	6
1	9.7%	9.0%	8.7%	8.4%	9.0%
1+2	–	8.9%	8.2%	8.2%	8.6%
1+2+3	–	–	8.1%	8.0%	8.4%
1+2+3+4	–	–	8.2%	**7.9%**	8.3%

Table 7.5: Serbo-Croatian OOV-Rates with different Splitting Methods. The baseline OOV-Rate is 13.6% (System B4).

Applying the vocabulary adaptation procedure to our Serbo-Croatian broadcast news data and the broadcast news recognition system B4 yields a significant improvement both in terms of the out-of-vocabulary rate as in terms of word accuracy. When comparing the baseline dictionary with the newly adapted one, the out-of-vocabulary rate of the speech recognition system is reduced by more than 40% from 13.6% to 7.9% (see table 7.5). Concerning

the system performance of B4 the original word error rate is reduced by 5.8% absolute (see table 7.6). This means that the reduction in out-of-vocabulary rate of 5.7% absolute is reflected in an identical reduction in word error rate from 36.0% to 30.2%.

B4	**Vocabulary Size**	**OOV-Rate**	**Word Error**
Baseline	31,000	13.6%	36.0%
Morphology-Based HDLA	31,000	7.9%	30.2%

Table 7.6: Serbo-Croatian Recognition Results of System B4 based on Adapted Vocabulary.

Compared to system B4 the size of the baseline dictionary for system B5 is 49,000 words where the *cutoff*, the number of times a word has to appear in the text database in order to get included into the recognition dictionary, is 14. Following a first recognition run with this baseline dictionary, a second one with a dynamically adapted dictionary of the same fixed size $N = 49{,}000$ but with a much lower out-of-vocabulary rate is performed. Especially when automatically transcribing broadcast news shows, the time factor is not a critical issue. Performing the second recognition run after a first raw and fast baseline recognition is a practicable approach and can just as well be done overnight.

The same experiments as described above for system B4 were also performed on our latest B5 system. Starting off with a baseline performance of 29.5% word error and an out-of-vocabulary rate of 8.7%, through morphology-based HDLA the number of out-of-vocabulary words can almost be divided in half from 8.7% to 4.8%. The 3.9% improvement in out-of-vocabulary rate is again reflected in a 3.5% improvement in word error rate yielding a performance of 26.0% word error.

	Minimum Word Stem Length					
Suffix Length	2	3	4	5	6	7
1+2+3+4	–	5.2%	5.0%	**4.8%**	5.4%	5.9%

Table 7.7: Serbo-Croatian OOV-Rates with different Splitting Methods. The baseline OOV-Rate is 8.7% (System B5).

B5	Vocabulary Size	OOV-Rate	Word Error
Baseline	49,000	8.7%	29.5%
Morphology-Based HDLA	49,000	4.8%	26.0%

Table 7.8: Serbo-Croatian Recognition Results of System B5 based on Adapted Vocabulary.

7.4 Results on German Data

The following experiments show that the same results hold for German news data where also a significant reduction of the out-of-vocabulary rate is observed. For the German language the fixed list of suffixes in table 7.4 is used to create the word stems. Experiments were done both on the I14 and the I23 system described in chapter 4.

The first baseline recognition run was done with the domain-specific dictionary of system I14 that consists of 17,000 word entries. With a fixed list of inflection endings and the experience on Serbo-Croatian data showing that the optimal word stem length was five or more letters long, the HDLA procedure was applied to the minimum lengths four, five and six. The original out-of-vocabulary rate of 9.3% could hereby be reduced to 6.0% (see table 7.9).

	Minimum Word Stem Length				
Suffix Length	2	3	4	5	6
fixed	–	–	7.7%	**6.0%**	6.5%

Table 7.9: German OOV-Rates with different Splitting Methods. The baseline OOV-Rate is 9.3% (System I14).

Improvements on our latest German broadcast news recognizer are in the same order of magnitude. For system I23 the lexicon size is 61,000 words which results in a baseline out-of-vocabulary rate of 4.4%. Our morphology-based HDLA procedure reduces this out-of-vocabulary rate by one third to 2.9%.

	Minimum Word Stem Length				
Suffix Length	2	3	4	5	6
fixed	–	–	3.2%	**2.9%**	3.1%

Table 7.10: German OOV-Rates with different Splitting Methods. The baseline OOV-Rate is 4.4% (System I23).

7.5 Conclusions

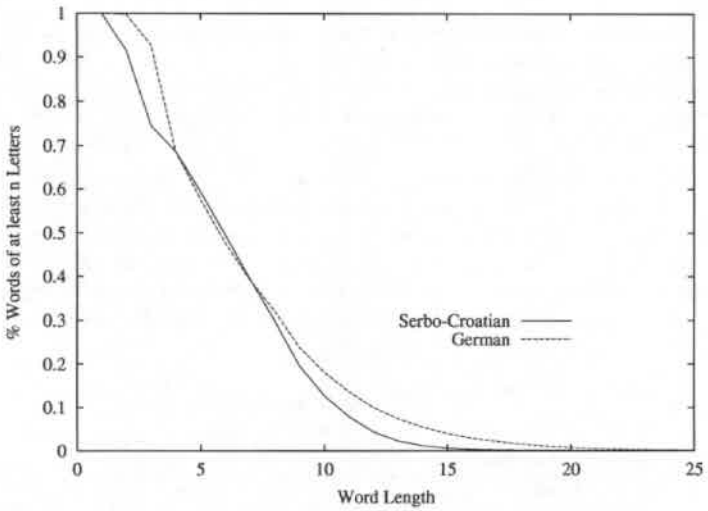

Figure 7.2: Cumulative Distribution of Word Length for the German and Serbo-Croatian language.

The approach described in this chapter uses morphological knowledge about word stems and inflection endings. Based on suffix lists derived from grammatical rules for a specific language, vocabulary adaptation is performed. Whereas in German a fixed suffix list is used, for Serbo-Croatian the optimum is to use suffixes up to length four. In both languages it turned out to be a good choice to fix the stem length to five which is correlated with the distribution of word

lengths: 50% of all words in Serbo-Croatian and German are longer than five letters. Figure 7.2 shows the distribution of different word lengths on our Serbo-Croatian and German databases.

Both on Serbo-Croatian as well as German broadcast news data the morphology-based HDLA procedure presented here achieves a reduction of out-of-vocabulary rates by 30% to 45%. These reductions also yield significant improvements in word error rate as presented for our Serbo-Croatian broadcast news system. However, to apply morphology-based lexical adaptation to a specific language, this language has to show certain regularities in its morphological structure. The ability to formulate grammatical rules is required, and the knowledge of a language expert is needed to define these rules. For some languages such an expert might be difficult to find and the definition of rules might be costly and time-consuming. Also, morphology-based lexical adaptation is only able to correct misrecognitions due to out-of-vocabulary words when the actually uttered word differs from the erroneously hypothesized word exclusively in its word ending. Both words have to have the same word stem in order to be considered "similar". This is a strong limitation to errors due to different inflection endings only, suggesting that other ways have to be found to also help eliminate misrecognitions of phonetically similar words that might differ in one or more phonemes in the beginning or middle of the word.

Nevertheless, if a large number of misrecognitions of a speech recognition system is due to falsely recognized word endings only, and if there are possibilities to formulate morphological rules for the language in question that can be used as selection criterion within the HDLA framework, the morphology-based approach described here is able to reduce high out-of-vocabulary and word error rates significantly.

Chapter 8

Distance-Based Lexical Adaptation

Beside the application of morphological similarity derived from grammatical rules and linguistic knowledge about a language, also similarity based on acoustic or grapheme distance measures can be used as selection criterion for the dynamic adaptation of a recognition dictionary. The following chapter on distance-based selection criteria for HDLA is divided into three sections: the first part describes a grapheme-based distance measure that is used to choose similar words for the adapted vocabulary. The second part introduces distance measures based on phonetic distances where three different methods of calculating these distances are investigated. Finally, a third approach, also based on phonetic distances, is presented where special consideration is given to the phenomenon of composita.

8.1 Motivation

The previous chapter has presented results of vocabulary adaptation where the HDLA procedure is only able to consider words that differ in their endings but have the same word stem. However, as the performed error analysis on Serbo-Croatian and German data shows, also a lot of errors due to unknown or out-of-vocabulary words result in the recognition of phonetically similar words.

If two words are phonetically very similar and differ in one or two phonemes only, this difference does not necessarily appear in the end of the word. The distinguishing phoneme might occur right at the beginning, e.g. "cat" and "fat", or somewhere in the middle, like in "resolution" versus "revolution". Differences like this cannot be captured through linguistic knowledge or grammatical rules but through a distance measure. Depending on the language in question this distance measure can either be based on grapheme or phoneme

level.

Also, dependency on knowledge about a specific language, especially linguistic knowledge about morphology is not desirable. Speech recognition systems might be built by a person that is not an expert on the language to be recognized. In this case the definition of classes of inflection endings would be impossible for the developer him- or herself. A language expert would have to supply the necessary classification scheme for the developer. However, even if the knowledge is available, coming up with suitable grammatical rules and classification schemes is a tedious and extremely time-consuming job. A better alternative would be to use knowledge inherent in the data itself (see the grapheme-based approach in section 8.2) or knowledge that can be acquired through tools that are included in the recognition system anyway, e.g. a grapheme-to-phoneme tool that is needed to build a first baseline dictionary (see section 8.3).

8.2 Grapheme Distance-Based Lexical Adaptation

An obvious solution when calculating the distance between words is the usage of the minimal editing distance between word pairs. Particularly for languages where the orthography closely matches its pronunciation the use of literary language is an adequate basis to calculate word distances.

Of course phonetic similarity is of more importance for our application within a speech recognition system than grapheme-based similarity. But for languages, like e.g. Serbo-Croatian, the grapheme-to-phoneme conversion follows easy-to-formulate rules and the mapping of graphemes to phonemes in most cases is one-to-one. Thus, the effort of converting all words to their phoneme representations can be spared and words can be compared on a letter basis. This especially applies to a large fallback lexicon of up to 500,000 words as it has been used for our German broadcast news data. Conversion of all lexicon entries to their phoneme representations would require some more work than really needed for the recognition process since the recognition dictionary requires the mapping of the N most frequent words only.

Using literary language respectively letter sequences instead, the distances between words of the utterance-specific vocabulary list derived from the baseline recognition run and the words of the fallback lexicon are calculated on grapheme level. The vocabulary list for the second recognition run is then chosen based on this distance, where word pairs are considered similar if their distance stays below a certain threshold.

8.2.1 Usage as Selection Criterion

The algorithm of chapter 6 again only has to be modified in step 2, where now instead of word stem equivalence the distance between two words on grapheme level is calculated:

2. The vocabulary list derived from the word lattice of the first recognition run is compared with all words of the fallback lexicon based on grapheme distances.

To determine the distance of two letters l and m the following formula is used:

$$distance(l, m) = (1 - \delta(l, m)) \tag{8.1}$$

where

$$\delta(l, m) = \begin{cases} 1 & : \text{ if } l = m \\ 0 & : \text{ if } l \neq m. \end{cases}$$

Thus, two letters l and m have distance 0 if they are identical, 1 otherwise. To calculate the distance between two words or letter sequences, the Levenshtein distance has been used as described in [Manber 1989]. The minimum number of editing steps that are required to convert one word or letter sequence into another is determined by using the technique of *dynamic programming*. Dynamic programming is an effective approach for problems that depend on the solutions of several slightly smaller subproblems. The instance of dynamic programming that has been used for our purposes is based on the computation of a large matrix. The essence of our approach is the creation of large tables where these tables are constructed iteratively based on all known previous results. Each entry is computed from a combination of other entries above it or to the left of it in the matrix. Main problem is to organize the construction of the matrix in the most efficient way.

For the problem of calculating the minimum number of editing steps to convert one word into another, the following solution has been used: Let $A(n)$ and $B(m)$ denote the substrings $a_1a_2...a_n$ respectively $b_1b_2...b_m$. The problem to be solved is to retrieve the best way to convert $A(n)$ into $B(m)$, thus also determining the distance between the two words. This distance, or minimum cost of converting $A(n)$ to $B(m)$, is denoted by $C(n, m)$. To this end a matrix $C[1..n, 1..m]$ is constructed. Each entry $C[i, j]$ of the matrix holds the value of $C(i, j)$. To compute $C[i, j]$ the values of $C[i-1, j]$, $C[i, j-1]$ and $C[i-1, j-1]$ are needed. Input to the algorithm is the string A of size n and the string B of length m. Output is the distance between the word pair A and B. The detailed algorithm used for calculating the minimum editing distance is given below.

```
Begin
    for i := 0 to n do C[i,0] := i;
    for j := 1 to m do C[0,j] := j;
    for i := 1 to n do
      for j := 1 to m do
        x := C[i-1,j] + 1;
        y := C[i,j-1] + 1;
        if a_i = b_j then
          z := C[i-1,j-1]
        else
          z := C[i-1,j-1] + 1;
        C[i,j] := min(x,y,z);
end
```

Figure 8.1: Algorithm for determining the distance between word pairs.

Based on this distance as selection criterion the words for the second recognition run are then chosen according to the HDLA algorithm and a second recognition pass is performed.

8.2.2 Results on Serbo-Croatian Data

Applying grapheme similarity as selection criterion to the HDLA framework, vocabulary adaptation experiments have been conducted on the Serbo-Croatian broadcast news data. As already mentioned, no grapheme-to-phoneme conversion had to be done. Distances were calculated on letter sequences only. A necessary prerequisite for those calculations was the existence of a large fallback lexicon retrieved from web texts. Nevertheless, for considering two words as being "similar", a certain minimum word stem length and a maximum distance between them had to be chosen. This is necessary as otherwise too many words would fulfill the claim of being similar to each other. Defining a fixed minimum word stem length ensures that for a one-letter word, such as "a", not all three-letter words have the distance two, hereby claiming that "a" and "all" are equally similar to each other as "want" and "wanted".

Experiments on a development test set resulted in an optimal minimum word length of six and a maximum distance of four. When applying the HDLA framework using grapheme similarity to our Serbo-Croatian broadcast news data, significant reductions in the percentage of new words could be achieved. As can be seen in table 8.1 the out-of-vocabulary rate of 8.74% is divided in half to 4.04% for this combination of parameters.

Minimum Word Stem Length	Maximum distance				
	1	2	3	4	5
3	8.64%	5.54%	4.49%	4.39%	–
4	8.64%	5.54%	4.42%	4.37%	–
5	8.64%	5.68%	4.32%	4.25%	–
6	8.64%	5.99%	4.35%	**4.04%**	4.05%
7	8.64%	6.46%	4.86%	4.23%	4.19%
8	8.64%	6.88%	5.43%	4.49%	4.34%
9	8.67%	7.38%	6.30%	5.54%	4.88%

Table 8.1: Serbo-Croatian OOV-Rates with various minimum word lengths based on grapheme distance measures. The baseline OOV-Rate is 8.74%.

8.3 Phonetical Distance-Based Lexical Adaptation

Unfortunately not all languages have such a close relationship of their orthography to their phonology, so that grapheme distances might not always be a suitable selection criterion for the HDLA procedure. One result of the error analysis of recognition results when transcribing broadcast news data was the observation that in case of misrecognitions due to out-of-vocabulary words often phonetically similar words are hypothesized instead. This suggests to use some kind of acoustically based distance measure as selection criterion to determine word similarity for the adaptation algorithm. As some kind of grapheme-to-phoneme conversion is always necessary to generate the phonetic representations of vocabulary words for a speech recognition system, usually a grapheme-to-phoneme tool is available to get a first baseline dictionary. In most cases this dictionary is then hand-corrected by human experts on pronunciations.

For the application described here, the available grapheme-to-phoneme tool has been used to generate phoneme representations not only for the baseline dictionary of the system but also for all words of the fallback lexicon retrieved from web texts. The up to 500,000 entries of this lexicon were not hand-corrected but used as is. Costly manual corrections were not necessary as the phonetically based approach requires the ability to generate phonetical representations for arbitrary vocabulary words automatically on the fly. The phonetic distance between words obtained from the word lattice of the first recognition pass and the words of the fallback lexicon was then used as sim-

ilarity criterion [Geutner et al. 1998d] to decide if a word was added to the dictionary for the second recognition run. Examples for the phoneme representation of Serbo-Croatian words as they were used in our Serbo-Croatian speech recognition engine can be found in table 8.2.

Word	**Phoneme Representation**
SIL	SIL WB
#eh#	+hGH WB
#ehm#	+hGH WB
#fragment#	+hGH WB
#hale#	+hBR WB
#human#	+hGH WB
#nonhuman#	+QK WB
abc	A WB B E C E WB
abdic1a	A WB B D I C1 A WB
abdic1	A WB B D I C1 WB
abdulah	A WB B D U L A H WB
administrativni	A WB D M I N I S T R A T I V N I WB
administrativnog	A WB D M I N I S T R A T I V N O G WB
...	...

Table 8.2: Examples of Serbo-Croatian Phoneme Representations of Words.

8.3.1 Usage as Selection Criterion

Modifying the already presented adaptation procedure to the needs of applying phonetic distances as similarity measure results in the following step 2 of the HDLA algorithm:

2. The vocabulary list derived from the word lattice of the first recognition run is compared with all words of the fallback lexicon based on phonetic distances.

For actually calculating the distance between two phonemes three different methods have been investigated:

1. Equality
2. Hamming distance with respect to a binary vector of phonetic features
3. Acoustic confusability

The following three sections will describe in more detail how the phonetic distance between two phonemes is determined in each of the three cases. For all methods the phonetical distance between two words is calculated according to the notion of Levenshtein distance, as described in the previous section.

8.3.2 Equality

Using the equality criterion the distance of two phonemes p and q is either 0, if the phonemes are identical, or 1 otherwise. To determine the distance between two phonemes the same formula as already applied for calculating letter distances in the grapheme-based approach is used:

$$distance(p,q) = (1 - \delta(p,q)) \tag{8.2}$$

where

$$\delta(p,q) = \begin{cases} 1 & : \text{ if } p = q \\ 0 & : \text{ if } p \neq q. \end{cases}$$

The distance between the phoneme representations of two words is computed through the algorithm described in figure 8.1. The resulting out-of-vocabulary rates when applying the equality criterion as selection criterion to the HDLA framework are presented in tables 8.3 and 8.5 for Serbo-Croatian and German broadcast news data.

8.3.2.1 Results on Serbo-Croatian Data

Again, a minimum length for a word had to be chosen and also a limit for a maximum distance had to be defined. This was necessary to prevent HDLA from creating word lists being too large because almost every word was considered "similar" to the other. Different parameter combinations were tried (see table 8.3) but the optimum was found for a minimum length of six and a maximum distance of four. For this combination the out-of-vocabulary rate could be decreased by 55% from 8.74% to 3.99%, a result similar to the grapheme-based approach. As the Serbo-Croatian language shows a one-to-one mapping from graphemes to their phonemes in most cases, this outcome was to be expected when applying phonetic-based instead of grapheme-based similarity measures.

Minimum Word Stem Length	Maximum Distance				
	1	2	3	4	5
3	8.74%	5.64%	4.49%	4.35%	–
4	8.74%	5.66%	4.46%	4.37%	–
5	8.74%	5.76%	4.30%	4.20%	–
6	8.74%	6.11%	4.32%	**3.99%**	**3.99%**
7	8.74%	6.65%	4.89%	4.13%	4.16%
8	8.74%	7.05%	5.54%	4.63%	4.25%
9	8.74%	7.56%	6.32%	5.71%	5.04%

Table 8.3: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures according to equation 8.2. The baseline OOV-Rate is 8.74%.

Recognition experiments at the minimum out-of-vocabulary rate of 4.0% were also performed on our baseline system B5 with 29.5% word error rate. The reduction in out-of-vocabulary rate is reflected in a 4.1% improvement of the word error rate to 25.4% (see table 8.4).

	Vocabulary Size	OOV-Rate	Word Error
Baseline	49,000	8.7%	29.5%
Phone-Based HDLA	49,000	4.0%	25.4%

Table 8.4: Serbo-Croatian Recognition Results based on Adapted Vocabulary using Phonetic Distances.

8.3.2.2 Results on German Data

The HDLA procedure was then applied to the German broadcast news recognizer, also using phonetical equality as selection criterion to choose the words for the adapted vocabulary of the second recognition run. Starting off with a baseline out-of-vocabulary rate of 4.43%, the best improvements could be achieved for a minimum word stem length of seven or eight respectively, and a maximum distance of four (see table 8.5). Using this combination of minimum word stem length and maximum distance, a 30% reduction in the out-of-vocabulary rate from 4.43% to 3.05% could be attained.

Minimum Word Stem Length	Maximum Distance				
	1	2	3	4	5
3	3.82%	3.54%	3.60%	3.66%	–
4	3.81%	3.51%	3.55%	3.60%	–
5	3.81%	3.51%	3.19%	3.33%	–
6	3.81%	3.59%	3.09%	3.44%	3.50%
7	3.81%	3.62%	3.16%	**3.05%**	3.62%
8	3.81%	3.65%	3.25%	**3.05%**	3.51%

Table 8.5: German OOV-Rates with various minimum word lengths based on phonetic distance measures according to equation 8.2. The baseline OOV-Rate is 4.43%.

8.3.3 Hamming Distance

For the experiments described in the previous section the distance of two phonemes was either 0, if the phonemes were identical, or 1 otherwise. In a second experiment we considered a distance measure based on phonemes that also takes various grades of similarity between different phonemes into account. To this end the Hamming distance with respect to a binary vector of phonetic features $f_i(p)$ for each pair of phonemes p and q is computed through the following formula:

$$distance(p,q) = \frac{1}{n}\sum_{i=1}^{n}(1 - \delta(f_i(p), f_i(q))) \tag{8.3}$$

where

$$\delta(f_i(p), f_i(q)) = \begin{cases} 1 & : \text{ if } f_i(p) = f_i(q) \\ 0 & : \text{ if } f_i(p) \neq f_i(q), \end{cases}$$

and $f_i(p)$ denotes the phonetic feature representation of phoneme p for the feature i, and n represents the number of available phonetic features.

If, for example, two phonemes share exactly the same phonetic features, their distance is defined to be 0. If they have no features in common at all, their distance corresponds to the number of used features, or in our case phoneme classes. After normalization their distance would be equal to 1.

In the experiments performed here, the features used to determine phoneme similarity by calculating the Hamming distance were the different acoustic categories a phoneme can be classified into. Thus, the Hamming distance used in our experiments was 0 if both phonemes were categorized into exactly the same phoneme classes. If they did not share the same characteristics concerning all

available features or phoneme classes, their distance corresponded to the number of classes one of them was in and the other was not. As a consequence, in the most extreme case the distance of two phonemes having no similar characteristics at all and not sharing any common phoneme classes equalled the number of acoustic categories that were used for the recognition process. Examples of the phoneme classes used in our Serbo-Croatian recognizer are given in table 8.6.

Class	**Phonemes**
NOISES	+QK +hGH +hBR +nGN
CONSONANT	B C C1 C5 D D1 DZ5 F G H J ...
VOWEL	A E I O U
VOICED	B D D1 DZ5 G J L LJ M N NJ ...
UNVOICED	C C1 C5 F K P S S5 T H
COMPACT	C1 D1 S5 Z5 K G H J
DIFFUSE	P B F M V
...	...

Table 8.6: Examples of Serbo-Croatian Phoneme Classes.

The phoneme "B" for example is a consonant and belongs to the classes "Voiced" and "Diffuse". The phoneme "F" is also a consonant and belongs to the class "Diffuse" but not to the class "Voiced". Instead it is classified into the complementary category "Unvoiced". Assuming that the seven classes introduced in table 8.6 would already constitute the whole phoneme set of our Serbo-Croatian speech recognition engine, the distance between the phonemes "B" and "F" would be calculated by using the classification given below:

Class	**B $\in$ Class**	**F $\in$ Class**
NOISES	0	0
CONSONANT	1	1
VOWEL	0	0
VOICED	1	0
UNVOICED	0	1
COMPACT	0	0
DIFFUSE	1	1

Table 8.7: Classification of the Phonemes "B" and "F".

Thus, the phonetic feature representation of "B" would be the binary vector

$(0,1,0,1,0,0,1)$ and $f_1 = 0$, $f_2 = 1$, $f_3 = 0$, $f_4 = 1$, $f_5 = 0$, $f_6 = 0$ and $f_7 = 1$. The binary vector for "F" would look like $(0,1,0,0,1,0,1)$ and $f_1 = 0$, $f_2 = 1$, $f_3 = 0$, $f_4 = 0$ and so forth. According to formula 8.3 the distance between the phoneme "B" and the phoneme "F" would be equal to 2 / 7, as they share the same characteristics for all but the two phoneme classes "Voiced" and "Unvoiced".

The overall distance of two words represented by sequences of phonemes are again determined by calculating their Levenshtein distance. When conducting experiments based on the Hamming distance as similarity criterion with Serbo-Croatian broadcast news data, distances were normalized to 1.0 and the best parameter combination turned out to be a minimum word length of four and a maximum distance of 0.7 or 0.8, where the baseline out-of-vocabulary rate could be improved from 8.74% to 5.43%.

	Maximum Distance					
Minimum Word Stem Length	0.4	0.5	0.6	0.7	0.8	0.9
3	6.11%	5.73%	5.52%	5.45%	5.45%	5.50%
4	6.13%	5.66%	5.50%	**5.43%**	**5.43%**	5.45%
5	6.23%	5.97%	5.68%	5.64%	5.66%	5.68%
6	6.46%	6.15%	5.78%	5.64%	5.66%	5.50%
7	7.02%	6.74%	6.27%	6.13%	6.04%	6.11%

Table 8.8: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures using phone-wise Hamming distances. The baseline OOV-Rate is 8.74%.

8.3.4 Acoustic confusability

The last method applied to determine the phonetical distance between two phonemes is their acoustic confusability. Confusability here means the confusion of phonemes when conducting recognition experiments. In order to obtain a reliable estimate for this distance measure, a recognition experiment on several broadcast news shows of the Serbo-Croatian training material was performed. The alignment of the hypothesized recognizer output and the words that had actually been uttered was taken as the basis to ascertain all emerging confusion pairs. The distance between an actually uttered phoneme and the output of the recognizer was then calculated through the various approaches that are described in the following four sections.

8.3.4.1 Baseline Approach

In a first approach simply the number of times the hypothesized phoneme p_h was mistaken for the uttered phoneme p_u are counted. Normally the probability of confusing a recognized phoneme p_h with a phoneme p_u is calculated by dividing the number of confusions of the pair p_u,p_h by the number of all confusions of the phoneme p_u found in the training set.

$$distance_1(p_u, p_h) = \frac{\#conf(p_u, p_h)}{\sum_{i=1}^{n} \#conf(p_u, p_i)} \tag{8.4}$$

where i equals the number of phonemes that are used within the recognition process of the Serbo-Croatian speech recognizer.

As the distance of uttered and recognized phoneme has to be 0 in case of a correct recognition, the obtained counts are first divided by the maximum number of counts a phoneme p_u was confused with any other phoneme instead. This maximum equals the number of counts a phoneme p_u was correctly recognized as the identical phoneme p_u.

$$distance_2(p_u, p_h) = \frac{\#conf(p_u, p_h)}{\max_{i=1}^{n} \#conf(p_u, p_i)} = \frac{\#conf(p_u, p_h)}{\#conf(p_u, p_u)} \tag{8.5}$$

Then the logarithm of the resulting outcome is taken to ensure a distance of 0 for a phoneme to itself. As phoneme pairs that have been confused more often are considered to be more similar than phonemes that were confused very rarely, the negative logarithm is applied to formula 8.5. This ensures smaller distances for frequently confused phoneme pairs.

$$distance_3(p_u, p_h) = -\log\left(\frac{\#conf(p_u, p_h)}{\#conf(p_u, p_u)}\right) \tag{8.6}$$

The distance of phoneme pairs that have never been confused in our recognition experiment, so that their number of confusions is 0, is set to ∞. Overall the following formula is used to determine the distance between an uttered phoneme and its hypothesized counterpart:

$$distance(p_u, p_h) = \begin{cases} \infty & : \#conf(p_u, p_h) = 0 \\ -\log\left(\frac{\#conf(p_u, p_h)}{\#conf(p_u, p_u)}\right) & : \#conf(p_u, p_h) \gg 0. \end{cases} \tag{8.7}$$

Again the Levenshtein distance is employed to calculate the overall distance between a word from the word lattice generated in the first recognition pass and a word included in the large fallback lexicon.

The results of this method are significantly worse than the results that were achieved by using the equality criterion. Whereas the equality criterion was able to decrease the out-of-vocabulary rate for the Serbo-Croatian data from 8.74% to 3.99%, the usage of acoustic confusability as selection criterion leads to a reduction to only 5.76% (see table 8.9). Primary reason for this poor performance is the fact that confusions of phoneme pairs that were never observed in the training experiment are not taken into account at all by setting their distance to ∞. As a consequence these pairs are considered to be the least similar ones where in reality it just means that they were never mistaken for each other, but the phonemes might still be more similar than other phoneme pairs. It also means that two words that contain such a pair will not be considered at all during the selection process for the new and adapted vocabulary of the second recognition run, as their distance is infinite. This behaviour is a quite familiar phenomenon in the field of language model training when dealing with the problem of new unseen events that never occurred in the training database. The next section will present a solution to this problem where, similar to the approaches taken in language modeling, a backoff scheme is applied to assign a non-zero count to confusion pairs that have never been observed in the recognition experiments on our training data.

	Maximum Distance			
Minimum Word Stem Length	1	2	3	4
3	7.70%	6.72%	5.97%	5.99%
4	7.73%	6.74%	6.04%	**5.76%**
5	7.92%	7.02%	6.34%	5.92%
6	8.03%	7.40%	6.93%	6.55%
7	8.20%	7.78%	7.49%	7.31%

Table 8.9: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures using speech recognizer acoustic confusability. The baseline OOV-Rate is 8.74%.

8.3.4.2 Baseline Approach with Backoff Scheme

The main focus of this second and modified baseline approach is to allow a distance $\neq \infty$ to be assigned to confusion pairs unseen in the training data.

Following an idea originally used in the area of language modeling, the baseline approach from the previous section is enhanced through a backoff scheme that uses absolute discounting [Ney & Essen 1993]. The counts of all confusion pairs observed in a test on broadcast news shows of the training set are discounted by a constant backoff factor b. The obtained discounted mass, which equals the number of different confusion pairs p_u, p_h observed in the training data multiplied by the backoff factor b, is then redistributed over all unseen events. For calculating the phonetic distance of two phonemes according to this idea, the following formula is applied:

$$distance(p_u, p_h) = \begin{cases} -\log\left(\frac{n_{p_u} * b}{(n - n_{p_u}) * (\#conf(p_u, p_u) - b)}\right) & : \#conf(p_u, p_h) = 0 \\ -\log\left(\frac{\#conf(p_u, p_h) - b}{\#conf(p_u, p_u) - b}\right) & : \#conf(p_u, p_h) \gg 0, \end{cases} \tag{8.8}$$

where b is the backoff factor, n_{p_u} the number of different confusion pairs p_u, p_h observed on the training data, and n the number of different phonemes in the phoneme set of the used recognition system.

Experiments with different backoff factors (b = 0.1, 0.25 and 0.75) have been performed and are summarized in tables 8.10 to 8.12. The best reduction in out-of-vocabulary rate was achieved when using a backoff factor of 0.75, but the resulting 4.43% out-of-vocabulary words still constitute a higher rate of unknown words than the 3.99% that can be attained by using the equality criterion as phonetic distance measure. However, the resulting out-of-vocabulary rates ranging between 4.4 and 4.6% are much lower than using the baseline approach of the acoustic confusability criterion for the vocabulary adaptation procedure without applying a backoff scheme.

	Maximum Distance										
	3	4	5	6	7	8	9	10	11	12	13
5	–	–	–	–	4.8%	4.8%	4.7%	4.8%	4.8%	4.8%	4.8%
6	–	–	–	–	4.9%	4.8%	4.7%	4.7%	4.7%	**4.6%**	**4.6%**

Table 8.10: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures using speech recognizer acoustic confusability (Baseline Approach with Backoff Factor = 0.1). The baseline OOV-Rate is 8.74%.

	Maximum Distance										
	3	4	5	6	7	8	9	10	11	12	13
5	6.4%	6.1%	5.7%	5.1%	4.8%	4.8%	4.7%	4.8%	4.8%	4.8%	4.8%
6	6.9%	6.6%	6.0%	5.4%	4.9%	4.9%	4.6%	4.6%	4.6%	**4.5%**	4.6%
7	–	–	–	–	–	–	5.4%	5.3%	5.3%	5.2%	5.2%

Table 8.11: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures using speech recognizer acoustic confusability (Baseline Approach with Backoff Factor = 0.25). The baseline OOV-Rate is 8.74%.

	Maximum Distance										
	3	4	5	6	7	8	9	10	11	12	13
5	6.4%	6.0%	5.6%	5.2%	4.9%	4.7%	4.6%	4.7%	4.7%	4.7%	4.7%
6	6.9%	6.6%	6.0%	5.6%	5.0%	4.7%	4.5%	4.5%	**4.4%**	**4.4%**	4.5%
7	–	–	-	–	–	–	–	4.9%	4.8%	4.8%	4.8%

Table 8.12: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distance measures using speech recognizer acoustic confusability (Baseline Approach with Backoff Factor = 0.75). The baseline OOV-Rate is 8.74%.

8.3.4.3 Normalized Baseline Approach with Backoff Scheme

For this experiment the same backoff scheme as in the previous section is used. In addition, the distances of all phoneme pairs are normalized to 1.0, resulting in the formula presented below:

$$distance(p_u, p_h) = \begin{cases} -\log\left(\frac{n_{p_u} * b}{(n - n_{p_u}) * (\#conf(p_u, p_u) - b)} * n\right) & : \#conf(p_u, p_h) = 0 \\ -\log\left(\frac{\#conf(p_u, p_h) - b}{\#conf(p_u, p_u) - b} * n\right) & : \#conf(p_u, p_h) \gg 0, \end{cases} \tag{8.9}$$

where

$$n = \frac{1}{\max_{i=1}^{n} distance(p_u, p_i)}$$

is the normalization factor and $\max_{i=1}^{n} distance(p_u, p_i)$ is the maximum distance of any confusion pair p_u, p_h before normalization.

Results of this experiment with a backoff factor of 0.75 are shown in table 8.13. The best result achieved, an out-of-vocabulary rate of 4.35%, yields almost the same percentage of unknown words as the unnormalized version of the algorithm.

	Maximum Distance						
Minimum Word Stem Length	1	1.5	1.8	1.9	2	2.1	≥ 2.2
4	–	–	–	–	4.49%	–	–
5	–	–	4.44%	4.44%	**4.35%**	**4.35%**	**4.35%**
6	5.45%	4.75%	4.52%	4.48%	4.39%	–	–
7	–	–	–	–	4.44%	–	–

Table 8.13: Serbo-Croatian OOV-Rates with various minimum word lengths based on phonetic distances using speech recognizer acoustic confusability (Normalized Approach with Backoff). The baseline OOV-Rate is 8.74%.

8.3.4.4 Modified Baseline Approach with Backoff Scheme

As a final experiment, the method introduced in section 8.3.4.2 where a backoff scheme had been applied to the baseline approach, is modified. This is done by dividing the counts of all obtained confusion pairs not by the number of counts a phoneme p_u was correctly recognized as the identical phoneme, but by the number of all confusions of the phoneme p_u found in the training set, as already proposed in equation 8.4. In this experiment no consideration is given to the fact that the distance of an uttered and a recognized phoneme should be 0 in case of a correct recognition. Instead pure probability estimates are used as distances between phoneme pairs:

$$distance(p_u, p_h) = \begin{cases} -\log\left(\dfrac{n_{p_u} * b}{(n - n_{p_u}) * (\sum_{i=1}^{n} \#conf(p_u, p_i) - b)}\right) & : \#conf(p_u, p_h) = 0 \\ -\log\left(\dfrac{\#conf(p_u, p_h) - b}{\sum_{i=1}^{n} \#conf(p_u, p_i) - b}\right) & : \#conf(p_u, p_h) \gg 0, \end{cases} \tag{8.10}$$

Again, a backoff factor of 0.75 was used for all experiments. For a minimum word stem length of six and a maximum distance of 18, the resulting out-of-vocabulary rate of 4.25% is the lowest value achieved by experiments based on acoustic confusability. Still, this number is not able to exceed the best result of 3.99% achieved so far on Serbo-Croatian data when using grapheme or phoneme distances based on the equality criterion.

	Maximum Distance			
Minimum Word Stem Length	17	18	19	20
5	–	4.35%	4.39%	4.39%
6	4.30%	**4.25%**	4.28%	4.30%
7	–	4.37%	4.41%	4.44%

Table 8.14: Serbo-Croatian OOV-Rates with various different minimum word lengths based on phonetic distances using speech recognizer acoustic confusability (Modified Approach). The baseline OOV-Rate is 8.74%.

8.3.4.5 Summarization of Results

All results achieved using acoustic confusability for the selection process of the adapted vocabulary within the HDLA framework are summarized in table 8.15.

	OOV-Rate
Baseline	8.74%
Acoustic Confusability (Baseline Approach)	5.76%
Acoustic Confusability (with Backoff)	4.43%
Acoustic Confusability (Normalized with Backoff)	4.35%
Acoustic Confusability (Modified Approach)	**4.25%**

Table 8.15: Serbo-Croatian OOV-Rates using Acoustic Confusability.

8.4 Phonetical Distance-Based Lexical Adaptation Extended to Composita

- REF: bundesgesundheitsminister seehofer plant eine neuregelung bei der sozialhilfe bei der gewährung von sogenannten ****** **EINMALLEISTUNGEN** WIE der anschaffung von ****** **WINTERKLEIDUNG** oder ****** **MÖBELSTÜCKEN** soll es künftig ****** **BUNDESEINHEITLICHE PAUSCHALBETRÄGE** ******** GEBEN dadurch so seehofer könnten verwaltungskosten von jährlich 250 millionen mark eingespart werden zu einschnitten für die sozialhilfe empfänger werde es dabei in solchen kommunen kommen die bisher überdurchschnittliche LEISTUNGEN GEWÄHRT hatten

  HYP: bundesgesundheitsminister seehofer plant eine neuregelung bei der sozialhilfe bei der gewährung von sogenannten **EINMAL LEISTUNGEN** DIE der anschaffung von **WINTER KLEIDUNG** oder **MÖBEL STÜCKEN** soll es künftig **BUNDES EINHEITLICHE PAUSCHAL BETRÄGE** GEGEN dadurch so seehofer könnten verwaltungskosten von jährlich 250 millionen mark eingespart werden zu einschnitten für die sozialhilfe empfänger werde es dabei in solchen kommunen kommen die bisher überdurchschnittliche LEISTUNG GEWÄHLT hatten

Figure 8.2: Alignment of a German Utterance.

When performing experiments on German broadcast news data, unlike for Serbo-Croatian, the morphology-based HDLA approach from chapter 7 results in a better reduction in the number of new words unknown to the recognizer than the best phonetical distance-based method using the equality criterion in this chapter (see section 8.3.2). Although the difference between both out-of-vocabulary rates is statistically not significant (2.86% vs. 3.05%), this observation suggests that in German less misrecognitions than in Serbo-Croatian are due to out-of-vocabulary words that are phonetically very close to the actually uttered word and differ in up to four or five phonemes only. Instead, for German a higher percentage of erroneously hypothesized words seem to have the same stem as the actually uttered word and differ in the word ending only. However, even when selecting the recognition vocabulary for a specific

utterance based on word stem equivalence, only differences in the word ending of up to a certain maximum length can be considered. In Serbo-Croatian suffixes of up to length four and for German a fixed list of suffixes was used where the longest suffix also had the length four (see also chapter 7). This is sufficient to cover inflection endings of nouns and verbs, but usually fails for composite words. In most cases the independent parts of a compound word consist of components much longer than up to four letters or phonemes, so that the phonetic distance between a compound word and one of its individual subparts is larger than four.

Figure 8.2 shows a German alignment in which out of 14 misrecognitions 10 are due to erroneously hypothesized parts of German compound words. Eight of these errors could have been avoided if the compound words also would have been included in the recognition dictionary of the recognizer, One of the compound words was included in the recognition dictionary, but still not recognized correctly. In all cases the separate independent parts of the composite word were included in the dictionary of the recognizer and, as individual words, were recognized correctly.

	Maximum Distance				
Minimum Word Stem Length	1	2	3	4	5
3	3.57%	2.90%	2.63%	2.64%	–
4	3.55%	2.88%	2.62%	2.60%	–
5	3.55%	2.88%	2.30%	2.34%	–
6	3.57%	2.98%	2.29%	2.42%	2.61%
7	3.58%	3.03%	2.41%	**2.14%**	2.56%
8	3.60%	3.10%	2.53%	2.19%	2.57%

Table 8.16: German OOV-Rates with various minimum word lengths based on phonetic distance measures considering composita. The baseline OOV-Rate is 4.43%.

By being able to find a method that selects an erroneously recognized compound word as part of the recognition dictionary for the second recognition pass of the HDLA procedure, all of these errors could be corrected. To this end not only the phonetic distance of the utterance-specific vocabulary list derived from the word lattice generated by the first recognition run to all words of the fallback lexicon is calculated. Exactly the same is done concerning the distance of all sequences of two nouns that appear next to each other in the word lattice to all entries of the fallback lexicon. As the majority of compound words in German are composed out of two autonomous separate

words, considering sequences of two words from the word lattice is enough. For calculating the phonetic distance the so far best method from section 8.3.2, the equality criterion, is used.

Conducting an experiment on our German broadcast news data, for a minimum word stem length of seven and a maximum distance of four phonemes, the baseline out-of-vocabulary rate of 4.4% can be improved by 53% to 2.1%.

8.5 Conclusions

Similar to the usage of morphological knowledge about the grammatical rules and lexical properties of a specific language, also distance-based methods can be used within the HDLA framework. Whereas the morphological approach requires a certain expertise on the respective language, all distance-based methods introduced in this chapter can be applied easily. Calculating the distance between two words based on grapheme or phoneme level requires comparably less effort than determining the morphological similarity of a word pair. Thus, both the calculation of grapheme as well as phonetic similarity offer a convenient alternative to the morpheme-based approach for the HDLA algorithm. Experiments have been performed on several methods for determining the phonetic distance of a word pair. Acoustic similarity can be calculated by using different criteria, such as the equality criterion, the Hamming distance with respect to a binary vector of phonetic features, or acoustic confusability. The optimal method however may vary from language to language, as the rather language-specific extension of phonetic distances to composita for the German language shows. In summary, determining the adapted vocabulary by means of acoustic or grapheme similarity offers a convenient and easy-to-apply technique to be used as selection criterion for the proposed HDLA approach.

Chapter 9

Artificial Creation of the Fallback Lexicon

So far only techniques for lexical adaptation have been presented that rely on the availability of a large fallback lexicon for a specific language. Both methods, the usage of morphological similarity as well as the application of distance-based measure, require a huge lexicon for the selection of word entries of the adapted vocabulary. Creation of such a lexicon so far always depended on the existence of an enormous text database for a particular language. The following chapter introduces a method that has no need for a large amount of text data. Instead it is based on the formulation of morphological rules for the language in question and the possibility of measuring phonetic distances between words. New words are automatically generated according to the defined language-specific rules and these morphological variations are then incorporated into an artificially created fallback lexicon. Phonetic distances between word pairs are then determined by using this artificially created lexicon.

9.1 Motivation

The previous chapters have presented two different methods for adapting the recognition dictionary to the speech utterance to be recognized. The morphology-based approach depends on expert knowledge about the characteristics of a specific language. The phonetic distance-based approach does not require linguistic knowledge or information about morphology, but needs the ability to convert words from literary language to their phoneme representations. In addition, both procedures have the need of a very large fallback lexicon for the language in question. Words included in this lexicon are the basis from which new words for the customized dictionary of the second recognition run are chosen. Success of both techniques highly depends on the quality

of the used fallback lexicon. To be of real benefit words contained in this lexicon have to be annotated with frequency counts based on their occurrence within a particular domain. Also, the fallback lexicon should exceed a certain minimum size to be able to offer an as large as possible choice of words for the adapted dictionary.

Unfortunately, for certain languages lexica like this might not be available. Sometimes even the necessary large text databases to generate them are difficult and costly to be retrieved, or cannot be found at all. Often simple word lists do exist, but the entries of these lists are not annotated according to frequency counts. Therefore a third method has been developed where the existence of a large text database is not required. Instead, linguistic knowledge about morphology is sufficient, as the fallback lexicon for adapting the dictionary of the second recognition run is generated artificially based on this knowledge.

9.2 Rule-Based Generation of Word Inflections

Similar to the morphology-based approach, the artificial creation of a fallback lexicon requires the formulation of language-specific rules for the generation of morphological inflection endings. Each rule defines the class a word belongs to and describes the morphological variations it can take by being part of this class. Examples of rules for the Serbo-Croatian language that were used in our experiments can be found in table 9.1.

Rule 1	**If** the word consists of word stem *w* and inflection ending "-ama", **then** generate *w*-ina, *w*-ica, *w*-ska ...
Rule 2	**If** the word consists of word stem *w* and inflection ending "-ba", **then** generate *w*-be, *w*-bena, *w*-barska ...
Rule 3	**If** the word consists of word stem *w* and inflection ending "-liki", **then** generate *w*-lika, *w*-likoj, *w*-like ...
Rule 4	...

Table 9.1: Example Rules for artificially creating Inflection Endings.

Based on these rules, the artificially generated fallback lexicon is created as follows: all words included in the baseline recognition dictionary used for the first recognition run are assigned to a morphological category. After this classification the appropriate rule is applied to each word of the utterance-specific word list derived from the word lattice generated in the first recognition

run. According to the rule that is applicable to this word, new words are generated and incorporated into a fallback lexicon.

To this end a large set of rules or a restricted subset can be used. The number of rules considered for the creation process of the fallback lexicon varies depending on the degree of freedom that is used for generating new words. Some rules might not only create lexically correct words. Instead, their application leads to the generation of words that are correct according to the syntax of the rule but have no semantic meaning at all. Also, the interpretation of the rules can be done very strictly or more loosely. Within this work various levels of strictness to interpret rules have been experimented with. Strict interpretation of rules means that a rule is applied exactly in the form as the examples given above. A more relaxed way to interpret them is to consider all inflection endings that appear either in the if-part or the then-part of the rule. All of these morphological word endings are put into one set used as the prerequisite for the application of the rule. The rule is then applicable to a word if its word ending matches one of the endings in the set. In this case rule interpretation is done according to table 9.2. Both the number of used rules as well as the employed interpretation level of strictness influence the size of the resulting fallback lexicon. The more importance is attached to generating legal words only, the smaller is the size of the obtained word list. If more freedom is allowed, the size of the fallback lexicon increases, thereby also including a higher percentage of "illegal" words.

Rule 1	**If** the word consists of word stem w and one of the inflection endings of set s_1 (consisting of -ama, -ina, -ica, -ska ...), **then** also generate all other inflections endings of this set.
Rule 2	**If** the word consists of word stem w and one of the inflection endings of set s_2 (consisting of -ba, -be, -bena, -barska ...), **then** also generate all other inflections endings of this set.
Rule 3	**If** the word consists of word stem w and one of the inflection endings of set s_3 (consisting of -liki, -lika, -likoj, -like ...) **then** also generate all other inflections endings of this set.
Rule 4	...

Table 9.2: Loose Interpretation of Rules.

Four different variations were used to create four distinct fallback lexica on our Serbo-Croatian data. The tested approaches varied both in the number of available rules as well as the used interpretation level when applying these rules. The larger set of rules included 28 definitions for generating word inflections, the smaller set consisted of 18 rules. The interpretation level could be either

strict or loose as illustrated in the previous tables. The four approaches (A1 - A4) experimented with to generate artificial fallback lexica are summarized in table 9.3.

System	# of Rules	Rule Interpretation
A1	28	loose
A2	28	strict
A3	18	loose
A4	18	strict

Table 9.3: Variations on the Artificial Creation of the Fallback Lexicon.

All combinations of the large or small set of rules combined with a loose or strict interpretation of these sets have been applied to the Serbo-Croatian broadcast news data. The resulting artificially created words were then compared with the Serbo-Croatian lexicon containing 300,000 legal words that had been acquired from the texts retrieved on the world-wide-web. This was done to determine the number of legal words generated through this procedure. Table 9.4 shows the number of artificially created words and the resulting size of the fallback lexicon for all the strategies employed. Also the percentage of generated words that were legal are given, as well as the percentage of legal words that were finally included in the recognition dictionary of the second pass of the HDLA procedure.

System	# of Artificially Created Words	Size of Fallback Lexicon	% Legal Words Generated	Legal Words in Lexicon
A1	487,643	491,132	11.6%	12.2%
A2	105,300	126,873	31.9%	19.7%
A3	140,272	121,053	24.2%	33.4%
A4	24,798	54,609	35.4%	66.2%

Table 9.4: Properties of Resulting Fallback Lexica.

9.3 Usage as Selection Criterion

The fallback lexicon resulting from the artificial creation process described in the previous section is then used in the same way as the fallback lexicon derived from a large text database that has been used when determining phonetic and grapheme distances. Again only step 2 of the HDLA algorithm has to be modified:

2. The vocabulary list derived from the word lattice of the first recognition run is compared with all words of the artificially created fallback lexicon based on phonetic distances.

All other steps of the algorithm are performed exactly the same way as if a data-driven fallback lexicon would have been used.

9.4 Results on Serbo-Croatian Data

For all strategies A1 to A4 the resulting out-of-vocabulary rates of the recognition dictionary for the second recognition run were determined by applying the HDLA algorithm. Phonetic distances based on the equality criterion were used to select the similar words integrated in the adapted vocabulary of the second recognition pass. Results of the experiments performed on systems A1 to A4 are given in the following four tables. Again different minimum word stem lengths and maximum distance restrictions were applied. Optimal results in terms of the greatest reduction in the out-of-vocabulary rate were achieved through different parameter combinations for the four systems. In summary, a minimum word stem length of five or six and a maximum distance of three or four depending on the used underlying system yielded the best improvements.

	Maximum distance			
Minimum Word Stem Length	1	2	3	4
3	6.11%	5.94%	5.97%	5.87%
4	6.11%	5.83%	5.94%	5.80%
5	6.11%	5.85%	5.85%	**5.78%**
6	6.11%	5.90%	5.92%	5.92%
7	6.11%	5.94%	5.87%	6.06%
8	6.11%	5.92%	5.80%	6.34%

Table 9.5: Serbo-Croatian OOV-Rates with various minimum word lengths and rule-based generated fallback lexicon based on phonetical distances. A1-System. The baseline OOV-Rate is 8.74%.

	Maximum distance			
Minimum word Stem Length	1	2	3	4
3	6.93%	6.86%	6.81%	6.93%
4	6.93%	6.86%	6.81%	6.93%
5	6.93%	6.86%	6.81%	6.93%
6	6.93%	6.86%	**6.79%**	6.93%
7	6.93%	6.91%	6.84%	6.93%
8	6.93%	6.91%	6.84%	6.81%

Table 9.6: Serbo-Croatian OOV-Rates with various minimum word lengths and rule-based generated fallback lexicon based on phonetical distances. A2-System. The baseline OOV-Rate is 8.74%.

	Maximum distance			
Minimum Word Stem Length	1	2	3	4
3	6.69%	6.51%	**6.48%**	6.65%
4	6.69%	6.51%	**6.48%**	6.65%
5	6.69%	6.53%	**6.48%**	6.67%
6	6.69%	6.53%	**6.48%**	6.60%
7	6.69%	6.58%	6.51%	6.06%
8	6.69%	6.58%	6.51%	6.51%

Table 9.7: Serbo-Croatian OOV-Rates with various minimum word lengths and rule-based generated fallback lexicon based on phonetical distances. A3-System. The baseline OOV-Rate is 8.74%.

	Maximum distance			
Minimum Word Stem Length	1	2	3	4
3	**7.09%**	**7.09%**	**7.09%**	7.12%
4	**7.09%**	**7.09%**	**7.09%**	7.12%
5	**7.09%**	**7.09%**	**7.09%**	7.12%
6	**7.09%**	**7.09%**	**7.09%**	7.12%
7	**7.09%**	**7.09%**	**7.09%**	**7.09%**
8	**7.09%**	**7.09%**	**7.09%**	**7.09%**

Table 9.8: Serbo-Croatian OOV-Rates with various minimum word lengths and rule-based generated fallback lexicon based on phonetical distances. A4-System. The baseline OOV-Rate is 8.74%.

The results of the different strategies for creating an artificial fallback lexicon using various sizes of rule sets and varying interpretation levels are summarized in table 9.9. As a conclusion the best strategy is to be very loose in rule interpretation and allow to generate a high percentage of illegal words. The approach that allowed the most freedom in word generation and used the larger set of rules, A1, also resulted in the best reduction of the out-of-vocabulary rate from 8.7% to 5.8%.

System	OOV-Rate
A1	5.78%
A2	6.79%
A3	6.48%
A4	7.09%

Table 9.9: Resulting OOV-Rates for artificially created Fallback Lexica.

9.5 Conclusions

Artificially creating a fallback lexicon for the HDLA framework based on linguistic and morphological knowledge is an alternative to fallback lexica retrieved from the world-wide-web. If no large databases for a specific language are available, the rule-based generation of such a lexicon is a practical approach. However, improvements in out-of-vocabulary rate are significantly smaller when using an artificially created lexicon than a corpus-based one.

One reason certainly is an overgeneration of illegally inflected words when applying a large number of morphological rules and interpreting them loosely. On the contrary, if a smaller set of rules is used and their application is enforced more strictly, the size of the newly created fallback lexicon is very small. In consequence it does not contain a sufficiently large number of words to offer the HDLA framework the required freedom to find enough words for choosing the vocabulary of the adapted recognition dictionary from. In summary, selecting the adapted vocabulary through the presented technique is a viable approach when no large databases for a specific language are available. In case a language expert is at hand and the language in question allows the formulation of linguistic or morphological rules, it is a valid alternative to other methods while still decreasing the percentage of unknown words by one third.

Chapter 10

World-Wide-Web-Retrieval-Based Lexical Adaptation

Selection of similar words for lexical adaptation within the HDLA framework can be done by applying linguistic knowledge about the morphology of a language or by using grapheme or phoneme distances. Beside these already presented selection criteria for choosing similar words for the adapted vocabulary, also information retrieval techniques can be employed. The underlying idea is to retrieve texts from the world-wide-web related to the utterance to be recognized. By processing texts dealing with the same topic, relevant words for the adaptation procedure can be found that might be neither linguistically nor phonetically close to hypothesized words of the first HDLA recognition pass, and thus cannot be found through the methods presented so far. This chapter describes two information retrieval approaches that can be used within the HDLA framework: the first is based on the Okapi similarity measure, the second uses the topicality of a news show to retrieve similar texts.

10.1 Motivation

So far emphasis has been put on including words into the adapted lexicon of the second recognition run that are similar to words already recognized in the first recognition pass. The similarity measures applied are either based on morphological similarity, or distances on grapheme or phoneme level. However, a great many of out-of-vocabulary words are neither similar in a morphological nor a phonetical definition. They are composed of proper names or named entities, meaning words that mark names and places not included in the dictionary of the speech recognition system. In German 15% of the out-of-vocabulary words

in the used test set are named entities, in Serbo-Croatian the same percentage consists of names and places.

Being able to recognize this kind of new words is extremely important when doing automatic transcription of broadcast news. First of course the conventional line of reasoning applies: the more words are hypothesized correctly by a speech recognizer, the lower is the achieved word error rate. Secondly, when looking at the shift of evaluation measures that is currently taking place especially in the field of information retrieval on multimedia broadcast news databases like the multilingual Informedia database [Hauptmann & Witbrock 1997a], more and more importance gets attached to the perfect recognition of named entities. Names and places are content words or semantic content bearing units and as that are used as keywords both for indexing the content of large multimedia news databases as well as for information retrieval and extraction.

Necessary prerequisite to recognize these relevant keywords correctly both for indexing and retrieval is of course that these words are part of the recognition dictionary. Selection of the vocabulary for a task-specific recognition dictionary is usually based on frequency observations of large text databases. As a consequence, keywords or topics that were covered by broadcast news shows over a small period of time only might not feature the necessary minimum count to be included, thus constituting an out-of-vocabulary word. As the correct recognition of these words is mandatory for a satisfactory retrieval performance of video or audio segments, methods have to be found to identify those words and include them into the lexicon for the respective speech recognition system.

10.2 Lexical Adaptation Based on Information Retrieval Techniques

When applying information retrieval techniques to the adaptation of the recognition lexicon, the same algorithm is used as for all selection criteria introduced so far. The output of a first recognition pass provides the basis to determine a set of similar words. A new vocabulary is created by combining entries from the original dictionary with the words hypothesized in the word lattice of the first recognition run and all similar words. Thus, for the information retrieval-based lexical adaptation procedure all steps of the HDLA algorithm are carried out exactly the same way as for all other techniques.

As an only difference, the decision of whether a word is "similar" to a word of the lattice vocabulary is based on a totally different criterion to the ones used so far. No large fallback dictionary, either derived from a large text

database or artificially created, is needed. No morphological knowledge or distance measures are necessary. As a consequence, the method described here can even be easily applied if few or no knowledge about the respective language is available. Also, no phonetic distances have to be calculated between the baseline recognition dictionary and a large fallback lexicon. Thus, it is not necessary to collect large text databases in order to be able to create a fallback lexicon from. Instead, there is only the need for a search engine that guarantees reasonably well retrieval results. The output of the first recognition run, meaning a segment- or utterance-specific word list, is used as input into the search process.

For the experiments concerning the application of information retrieval techniques as selection criterion for the HDLA procedure, German broadcast news data has been used. Speech recognition experiments were performed using the speech recognizer described in chapter 4. Output of this speech recognizer is supposed to be integrated into a system that follows the idea of the Multilingual Informedia Project. Similar to the multimedia Informedia database the View4You system [Kemp et al. 1998a] allows to index and retrieve transcriptions of German broadcast news shows. The information retrieval process described here (see also [Kemp & Waibel 1998]) can be performed by any search engine that is offered on the world-wide-web, like e.g. Yahoo, Alta Vista or Lycos. However, in order to be able to understand exactly how the used search engine works and also have some influence on how the search is done, a new information retrieval engine was developed at the University of Karlsruhe. This search engine was designed within the View4You system. The exact mode of operation of this engine will be described in the following section.

10.2.1 The View4You Information Retrieval Engine

The search engine developed within the View4You system is based on the Okapi similarity measure [Beaulieu et al. 1997]. This measure has been evaluated thoroughly in the context of the **T**ext **RE**trieval **C**onference (TREC) information retrieval contests co-sponsored by the **N**ational **I**nstitute of **S**tandards and **T**echnology (NIST), and has been found to be very powerful. The Okapi measure can be parameterized to the special requirements of a task. For the information retrieval engine of the View4You system a parameterization that has been found to work very well especially for short queries has been used [Wilkinson et al. 1995]:

$$d(q,d) = \sum_{t \in q \wedge t \in d} \left(\frac{f_{d,t}}{f_{d,t} + \sqrt{\frac{f_d}{E(f_d)}}} \right) \log \left(\frac{N - f_t}{f_t} \right) \tag{10.1}$$

$= \text{Okapi}(k_1 = 1, k_2 = 0, k_3 = 0, b = 1, r = 0, R = 0)$

where N is the number of documents in the collection used, f_t is the number of documents containing term t, $f_{d,t}$ is the frequency of term t in document d, and f_d is the number of terms in document d which is an approximation to the document length. A *term* in this context is the same as a word. However, the 500 most frequent words, like e.g. "I", "other", "a" and so on, are excluded. The database engine computes the distance between a query and each article in the database and returns the articles sorted in decreasing order of similarity to the query. For longer queries of about 50 words typically several 1,000 articles are found.

Search is done on a corpus that consists of texts retrieved from three different world-wide-web sites. "Germany Live" is an internet newspaper that offers daily news reports. "Bayrischer Rundfunk 5" (BR5) is a German radio station in southern Germany that distributes transcriptions of its daily news broadcasts over the internet. Finally, the "Tagesschau" part of the corpus contains transcriptions of the anchor speaker part of the daily broadcasted "Tagesschau", and also closed caption from another news television show, the "Tagesthemen". The View4You database consists of texts that cover a 1-year period from mid-1996 to mid-1997.

Source	# Articles
Germany Live	56,883
BR5	8,488
Tagesschau	5,683
Total	71,054

Table 10.1: The View4You Text Database.

10.2.2 Results on German Data

Recognition experiments were conducted on German broadcast news data [Kemp & Waibel 1998]. A first recognition run was performed on the baseline system described in chapter 4 yielding a performance of 24.7% word error rate. For each speech segment of the test set an utterance-specific word list was created. This word list was then used as input into the information retrieval engine described above.

Depending on the length of the speech segment in question, the word list contained a vocabulary between 100 and 2,000 words. As a consequence, each query supplied a large number of documents. Here in this context a query is defined as one word from the utterance-specific word list. In order to limit

the flood of information that can be retrieved through each utterance, only the 1,000 most relevant documents for each query were used for the vocabulary adaptation process. The set of 1,000 documents for each query was ranked according to the relevance of each document and labeled with this relevance criterion. Then, all documents for one utterance that could be retrieved through the one-word-based queries were put together into one large text corpus. Finally, frequency counts of all words in this new corpus were determined. This was not done the conventional way where just the occurrences of this word within the text database are counted, but each count was multiplied by the relevance criterion attached to the document the word was found in. That means that words originating from more relevant texts were weighted higher than words found in less relevant ones.

Having derived a new word list based on relevance and frequency, the words from this list were integrated into the recognition vocabulary of the second recognition pass. To this end the original dictionary was filled up with words from the frequency-relevance word list to a certain percentage. Experiments on a development test set have shown that an optimal reduction of the out-of-vocabulary rate and thus the best improvement of recognition performance is achieved when including a number of approximately 10% of the original lexicon size into the dictionary for the second recognition run. The new vocabulary entries are incorporated into the recognition lexicon by exchanging them for the least frequent words in the original dictionary.

Using this procedure, the original out-of-vocabulary rate of our test set of 4.4% is reduced by 1.2% absolute to 3.2% (see table 10.2).

10.3 Usage of Topicality for Lexical Adaptation

The application of world-wide-web-based retrieval techniques as described in the previous section is one technique to account for out-of-vocabulary words representing proper names and named entities. Since this approach is able to reduce the number of out-of-vocabulary words by only 23%, a second approach was pursued.

Instead of using the output of the first recognition run for information retrieval the knowledge about the topicality of a news broadcast is exploited. When trying to find named entities and proper names that are totally new or have rarely occurred in the database of the task so far, the usage of current news texts is a very promising source of information.

However, the usage of world-wide-web texts for vocabulary adaptation alone would only help in correcting proper names. Therefore a different ap-

proach is pursued: the technique that had worked best for German broadcast news data, namely the usage of phonetic distances when also considering compositа, was combined with using the topicality of the broadcast news show. This means that the HDLA procedure of section 8.4 was applied which uses the phonetic distances between word pairs as selection criterion and also considers distances between word sequences from the lattice. Based on this already adapted vocabulary, the vocabulary from the web texts for a certain date is added, so that step 2 of the HDLA algorithm looks as follows:

2. a) The vocabulary list derived from the word lattice of the first recognition run is enhanced by generating compositа.

2. b) The resulting word list is compared with all words of the fallback lexicon based on grapheme distances.

2. c) All words found in the word lattice and all similar words (i.e. all words that have a phonetic distance below a certain maximum) are included into the recognition vocabulary of the second run.

2. d) All vocabulary entries that have been found in the text corpus created through the topicality criterion and are not yet included in the dictionary are added.

For this experiment the web site of the radio station BR5 was used, as this site provides transcriptions of the daily broadcasted radio news in a compact form.

Applying the procedure described above to our German broadcast news recognition system, the baseline out-of-vocabulary rate of 4.4% is reduced by 57% to 1.9%. Compared to the phonetical distance-based approach extended to compositа from section 8.4 that had resulted in an out-of-vocabulary rate of 2.1%, this means an improvement of 0.2% absolute or 10% relative. When looking at the percentage of named entities in the set of words unknown to the recognizer, this numbers approximately correspond to each other.

	OOV-Rate
Baseline	4.4%
Information Retrieval-Based HDLA	3.2%
Topicality-Based HDLA	1.9%

Table 10.2: German OOV-Rates based on information retrieval.

A recognition experiment on our latest I23 system was performed using the adapted recognition dictionary with the lowest out-of-vocabulary rate achieved

for our German data. The baseline word error rate of 24.7% could thereby 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 10.3: German Recognition Results based on Adapted Vocabulary.

10.4 Conclusions

The information retrieval techniques described above show the distinct advantage of not having to accumulate linguistic knowledge about a certain language. No huge text databases have to be collected and further processed. Any available search engine can be used for the retrieval process. However, using information retrieval techniques alone for vocabulary adaptation only yields a small reduction in terms of the out-of-vocabulary rate. Better results are achieved when also taking the topicality of a news broadcast into account. Combined with the already improved vocabulary through phonetical distances for words and composita, this approach yields an out-of-vocabulary rate reduction of 57% from 4.4% to 1.9%.

Chapter 11

Conclusions

Speech recognition systems for conversational speech have to be able to handle very large vocabularies as spontaneous human-to-human conversations cannot be restricted to a predefined vocabulary. The same argument applies when using speech recognition systems for the task of transcribing broadcast news shows. Similar to a conversation between two people it is impossible to limit the spontaneous speech input to a fixed vocabulary. As a consequence there will always exist a certain amount of new words that cannot be foreseen and thus are not included in the recognition dictionary. Each of these out-of-vocabulary words will automatically lead to a recognition error and usually also trigger some additional errors in the region around the unknown word. Whereas this fact is not a major issue in languages like English with less than 1% of words unknown to the recognizer when transcribing English broadcast news using a vocabulary of 64,000 words, this number is much higher for languages such as Serbo-Croatian and German.

Reason for the higher percentage of unknown words is a very fast vocabulary growth that can be observed in these two languages. This is due to the lexical and morphological structure of Serbo-Croatian and German: both languages show a very large number of distinct inflection endings both for nouns as well as for verbs and adjectives. In addition, the notion of composita is found in German. Also, due to the task of transcribing broadcast news, both languages show a higher percentage of task-specific proper names than there might occur in other domains.

The significant difference between highly inflected languages and languages that possess a very simple structure becomes clear when comparing the resulting vocabulary size of three text corpora of the respective languages English, Serbo-Croatian and German covering the same number of tokens. Whereas a corpus of one million tokens in English shows a vocabulary of 26,000 words, a corpus of the same size results in 44,000 words in German, and in even 89,000

for Serbo-Croatian. Naturally a speech recognition system for a certain application using a recognition dictionary of a fixed size N will cover a smaller percentage of words in a language with a rapid vocabulary growth than it would cover in English.

For a recognition dictionary containing 64,000 word entries an English broadcast news recognizer will have an out-of-vocabulary rate of 0.7%. The same dictionary size leads to an out-of-vocabulary rate of 4.4% in German and 7.8% in Serbo-Croatian. Extrapolating this number to a 15-minutes broadcast news show consisting of approximately 2,000 words, this means that out of these 2,000 words only 14 words would be misrecognized in English, 88 in German and already 156 in Serbo-Croatian. These misrecognition errors would be due to the fact that the actually uttered word is not included in the recognition dictionary.

As an indefinite expansion of the size of the recognition dictionary is not possible, other ways have to be found to reduce the number of out-of-vocabulary words during the recognition process and thus improve recognition performance. One possibility is the usage of other base units than words during recognition. In this work, decomposition of words according to morphological rules has been performed and the resulting morphemes were used as smaller base units. By being able to compose new unseen words out of parts already included in the vocabulary, the speed of the vocabulary growth for a specific language is alleviated and simultaneously the rate of out-of-vocabulary words for a recognizer of a certain vocabulary size N is decreased. However, experiments have shown that in spite of a significant reduction in the percentage of unknown words, recognition results cannot be improved through this approach. Recognizers built on top of these units suffer a severe degradation in performance due to the overgeneration of illegally inflected word hypotheses leading to a large number of hypothesized morpheme concatenations that do not map to legal words.

As already mentioned, an indefinite increase of the size of the recognition dictionary is impractical for the implementation of most existing speech recognition systems and even very large lexica would not be sufficient to account for all out-of-vocabulary words. Therefore a second approach is the dynamic expansion of the recognition vocabulary. Whereas the size of the dictionary used within the recognition process is still considered to be finite, this idea virtually allows for a much larger number of words to be recognized. This is done through a multipass strategy called **Hypothesis Driven Lexical Adaptation** (HDLA) where two recognition runs are performed: a first run is done on a fixed baseline dictionary consisting of the N most relevant words for the task in question. The hypothesized output of this first pass in form of a word lattice is then used to dynamically adapt the recognition dictionary to

the speech segment to be recognized. Through this approach the number of out-of-vocabulary words is reduced and a second recognition run is carried out on the improved vocabulary.

The techniques used to perform the actual adaptation are based on error analyses of typical alignments for Serbo-Croatian and German recognition results. In both languages misrecognition of an out-of-vocabulary word usually happens in of the following variations:

- The actually uttered word differs only in the inflection ending from the erroneously hypothesized word.
- The actually uttered word is phonetically similar to the hypothesized word or sequence of words.
- The actually uttered word denotes a named entity that could not be recognized because it is unknown to the recognition dictionary (in this case the hypothesized word is often also phonetically similar to the unknown named entity).

Knowledge about these three types of errors can be used for selecting the utterance-specific vocabulary for the second recognition run by exchanging some words of the baseline dictionary through words similar to the word hypotheses of the first recognition pass. Figure 11.1 illustrates different selection criteria that can be used within the HDLA procedure to choose these similar words for the adapted vocabulary of the second recognition run:

- morphological knowledge
- grapheme-based distance measures
- phone-based distance measures (using equality, Hamming distance or acoustic confusability)
- phone-based distance measures also considering composita
- phone-based distance measures using an artificially created fallback lexicon
- world-wide-web-based retrieval (information retrieval techniques or topicality-based)

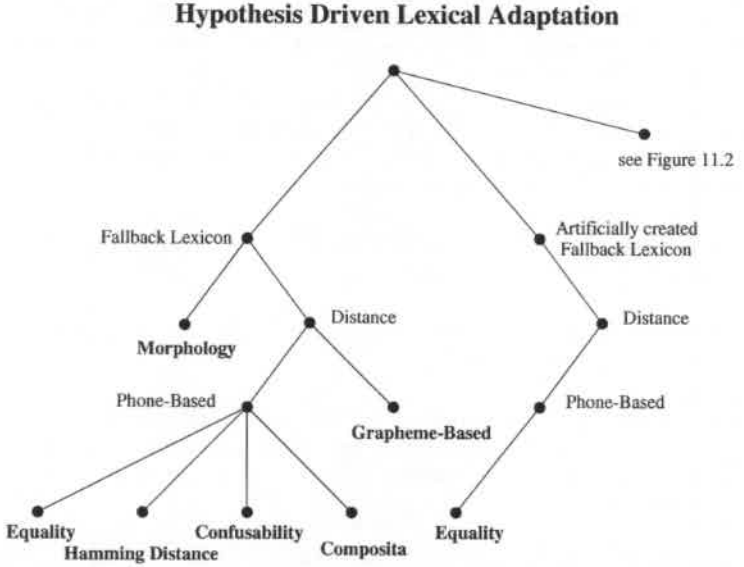

Figure 11.1: Selection Criteria for the HDLA Approach.

However, for the different ways of selecting the new customized vocabulary, different prerequisites have to be fulfilled. Using morphology-based lexical adaptation, linguistic knowledge about morphology and inflection endings has to be provided by a language expert which can be an expensive, tedious and extremely time-consuming task. The same applies when an artificially created fallback lexicon is used to determine phoneme or grapheme distances between words and use them as selection criterion for choosing similar entries. In this case language-specific rules have to be defined that generate new words even if a very small database of the respective language is available as basis only.

In cases where the linguistic knowledge of the language in question is very difficult to get, it might be easier to use distance measures either based on grapheme or phoneme level. For the calculation of grapheme distances the usage of literary language is enough. When determining phoneme distances a grapheme-to-phoneme tool is needed that is necessary for the creation of a dictionary anyway. The usage of these distance measures of course requires the availability of large text databases. A large amount of text material is needed to create a big enough fallback lexicon for determining the distances between word pairs. If databases like this cannot be retrieved from the world-wide-web

or similar sources, the usage of linguistic knowledge might be an alternative. However, the trade-off between linguistic knowledge and the availability of a large database for a specific language remains: in order to perform HDLA either expert knowledge on the morphological structure or a large enough database of the language has to be used. The correlation of both prerequisites for the main branches of the HDLA approach is shown in table 11.1.

HDLA Approach	Requirements:	
	Morphological Knowledge	Large Database
Morphology-Based	yes	yes
Distance-Based	no	yes
Artificially created Fallback Lexicon	yes	no

Table 11.1: Requirements for applying HDLA.

Lexical adaptation based on world-wide-web retrieval is a special case of HDLA where linguistic knowledge is not important, but the availability of text data is imperative. For this approach either information retrieval techniques or the knowledge about the topicality of a news broadcast can be used (see also figure 11.2). When adapting the vocabulary for the second recognition run based on information retrieval techniques a search engine and a large database to perform the retrieval on is needed. If knowledge about the topicality of a broadcast news show is used to customize the recognition dictionary, access to current-affairs news texts has to be guaranteed.

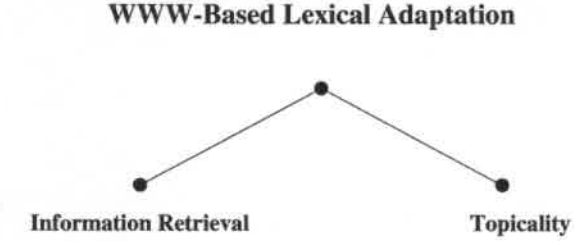

Figure 11.2: WWW-Based Lexical Adaptation.

Experiments with various selection criteria for the HDLA procedure have been performed on Serbo-Croatian and German broadcast news data. Figure 11.3 illustrates the resulting out-of-vocabulary rates that were obtained

on our Serbo-Croatian test set. Table 11.2 shows the improvement in word error rate that was achieved when using morphology-based and phone-based HDLA. Adapting the recognition dictionary for the second recognition run through the usage of morphological knowledge reduces the out-of-vocabulary rate from originally 8.7% to 4.8%, and gives a 26.0% word error rate. An even better result can be achieved when using the optimal method for this specific language: phonetic distance measures based on equality. With a 54% reduction in out-of-vocabulary rate from 8.7% to 4.0% the word error rate is decreased from 29.5% to 25.4%.

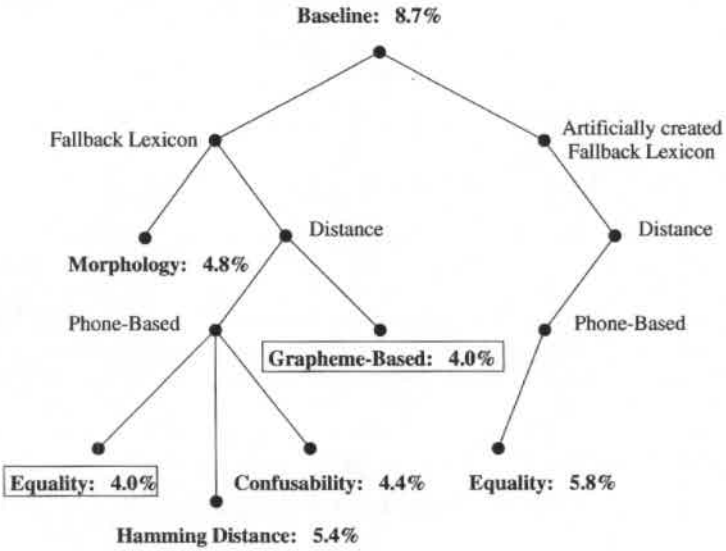

Figure 11.3: OOV-Rates for HDLA performed on Serbo-Croatian Data.

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

Table 11.2: Serbo-Croatian Recognition Results based on Adapted Vocabularies.

Whereas grapheme-based or phonetically based distances based on equality turned out to be the optimal methods for the reduction of the out-of-vocabulary rate in Serbo-Croatian, the picture is different for German. Here, the morphology-based approach leads to a larger reduction in out-of-vocabulary rate than the phone-based approach. This is certainly due to the fact that the usage of knowledge about the morphological structure of the language also considers the existence of compound words in German. Figure 11.4 shows that an even better result is possible when using phonetic distances not only on word basis but also considering composita. A 53% reduction in the number of unknown words from 4.4% to 2.1% can be achieved. This result can even be improved when combining it with the usage of the topicality of the news broadcast. When conducting a recognition experiment on a dictionary giving the best out-of-vocabulary rate of 1.9% by performing topicality-based lexical adaptation, a 57% reduction to the baseline out-of-vocabulary rate of 4.4% is achieved. Also the baseline recognition result of 24.7% word error rate is improved to 23.1% (see also table 11.3).

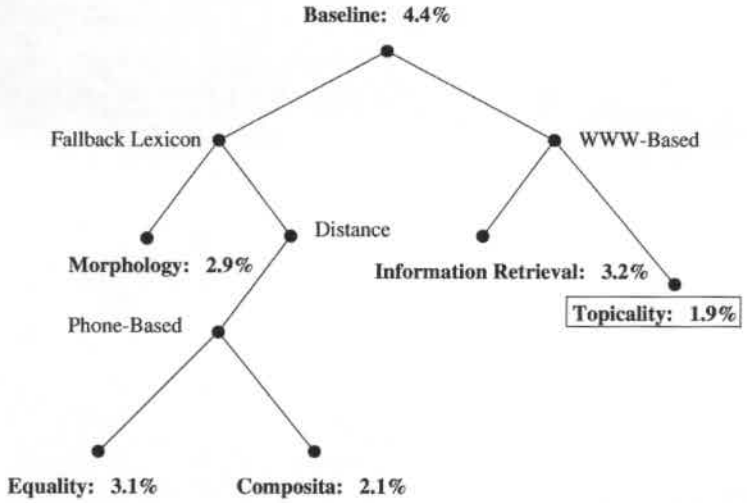

Figure 11.4: OOV-Rates for HDLA performed on German Data.

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

Table 11.3: German Recognition Results based on Adapted Vocabulary.

Table 11.4 summarizes the out-of-vocabulary rates of all experiments conducted within the presented HDLA framework. It is interesting to see that different methods and selection criteria are optimal for the different languages experiments were conducted on. Whereas the distance-based approach on either grapheme or phoneme level turns out to be the optimal procedure to reduce high out-of-vocabulary rates for the Serbo-Croatian language, in German the morphology-based approach outperforms the distance-based selection criterion applied to simple words only. It is not until also taking the notion of compound words into account, which is a very special feature of the German language, that the phone-based distance measure as selection criterion produces better results than using morphological knowledge. The outcome of the performed experiments shows clearly that it is helpful to consider the special characteristics of a language when trying to find useful selection criteria for the lexical adaptation procedure.

	Serbo-Croatian	**German**
	OOV-Rates	
Baseline	8.7%	4.4%
Morphology-Based	4.8%	2.9%
Grapheme-Based	**4.0%**	–
Equality	**4.0%**	3.1%
Hamming Distance	5.4%	–
Acoustic Confusability	4.4%	–
Phone-Based (Composita)	–	2.1%
Artificially created Fallback Lexicon	5.8%	–
Information Retrieval Based	–	3.2%
Topicality-Based	–	**1.9%**

Table 11.4: OOV-Rates for Serbo-Croatian and German Data on all HDLA Approaches.

Within this work improvements in the performance of both a Serbo-Croatian as well as a German broadcast news recognition system have been presented. With respect to the problem of encountering excessive growth of vocabularies in heavily inflected languages the developed Hypothesis Driven Lexical Adaptation procedure turned out to be a very effective means of reducing the rate of out-of-vocabulary words. By applying this two-pass recognition technique to the task of transcribing broadcast news shows in Serbo-Croatian and German a significant reduction of up to 57% in the out-of-vocabulary rate of both languages could be achieved, thereby also improving recognition performance by decreasing the word error rates by up to 14%.

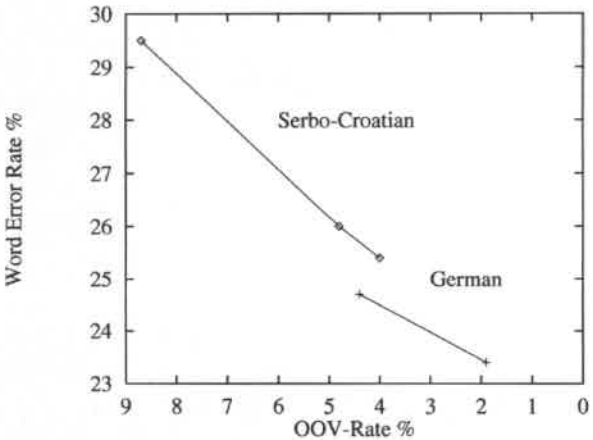

Figure 11.5: Resulting Word Error Rates with different OOV-Rates for Serbo-Croatian and German Broadcast News Data.

Bibliography

G. Adda, M. Adda-Decker, J.-L. Gauvain, L. Lamel (1997). Text Normalization and Speech Recognition in French. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2711–2714, Rhodos, Greece, September 1997.

J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang (1998). Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, Virginia, February 1998.

A. Asadi, R. Schwartz, J. Makhoul (1990). Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System. In *Proceedings of the IEEE 1990 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 125–128, Albuquerque, New Mexico, April 1990.

A. Asadi, R. Schwartz, J. Makhoul (1991). Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System. In *Proceedings of the IEEE 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'91)*, pages 305–308, Toronto, Canada, May 1991.

X. Aubert, C. Dugast, H. Ney, V. Steinbiss (1994). Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data. In *Proceedings of the IEEE 1994 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Volume 2, pages 129–132, Adelaide, Australia, April 1994.

A. Averbuch, L. Bahl, R. Bakis, P. Brown, J. Cohen, A. Cole, G. Daggett, S. Das, K. Davies, S. DeGennaro, P. de Souza, D. Fraleigh, M. Garrett, F. Jelinek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman (1985). A Real-Time Isolated-Word Speech Recognition System for Dictation

Transcription. In *Proceedings of the IEEE 1985 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'85)*, pages 858–861, Tampa, Florida, March 1985.

L.R. BAHL, J.K. BAKER, F. JELINEK, R.L. MERCER (1977). Perplexity – A Measure of Difficulty of Speech Recognition Tasks. In *94th Meeting of the Acoustical Society of America*, Miami Beach, Florida, December 1977.

L.R. BAHL, P.F. BROWN, P.V. DE SOUZA, R.L. MERCER (1989). A Tree-Based Statistical Language Model for Natural Language Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 37(7):1001–1008, July 1989.

L.R. BAHL, S. BALAKRISHNAN-AIYER, J.R. BELLGARDA, M. FRANZ, P.S. GOPALAKRISHNAN, D. NAHAMOO, M. NOVAK, M. PADMANABHAN, M.A. PICHENY, S. ROUKOS (1995). Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task. In *Proceedings of the IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 41–44, Detroit, Michigan, May 1995.

J.M. BAKER (1993). Dictation, Directories, and Data Bases; Emerging PC Applications for Large Vocabulary Speech Recognition. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93) - Keynote Speech*, pages 3–10, Berlin, Germany, September 1993.

L.E. BAUM (1972). An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, 3:1–8, 1972.

M.M. BEAULIEU, M. GATFORD, X. HUANG, S.E. ROBERTSON, S. WALKER, P. WILLIAMS (1997). Okapi at TREC-5. In *Proceedings of the 5th Text Retrieval Conference (TREC-5), NIST*, Gaithersburg, Maryland, January 1997.

A. BERTON, P. FETTER, P. REGEL-BRIETZMANN (1996). Compound Words in Large-Vocabulary German Speech Recognition Systems. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 1165–1168, Philadelphia, Pennsylvania, October 1996.

E. BLACK, F. JELINEK, J. LAFFERTY, D.M. MAGERMAN, R. MERCER, S. ROUKOS (1992). Towards History-based Grammars: Using Richer Models for Probabilistic Parsing. In *Proceedings of the ARPA Workshop on Spoken Language Technology*, March 1992.

J.-M. BOITE, H. BOURLARD, B. D'HOORE, M. HAESEN (1993). A New Approach Towards Keyword Spotting. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 1273–1276, Berlin, Germany, September 1993.

M. BOROS, W. ECKERT, F. GALLWITZ, G. GÖRZ, G. HANRIEDER, H. NIEMANN (1996). Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 1009–1012, Philadelphia, Pennsylvania, October 1996.

M. BOROS, M. ARETOULAKI, F. GALLWITZ, E. NÖTH, H. NIEMANN (1997). Semantic Processing of Out-Of-Vocabulary Words in a Spoken Dialogue System. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 1887–1890, Rhodos, Greece, September 1997.

D. CARTER, J. KAJA, L. NEUMEYER, M. RAYNER, F. WENG, M. WIRÉN (1996). Handling Compound Nouns in a Swedish Speech-Understanding System. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 26–29, Philadelphia, Pennsylvania, October 1996.

S. CHEN, M.J.F. GALES, P.S. GOPALAKRISHNAN, R.A. GOPINATH, H. PRINTZ, D. KANEVSKY, P. OLSEN, L. POLYMENAKOS (1998). IBM's LVCSR System for Transcription of Broadcast News used in the 1997 HUB4 English Evaluation. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 69–74, Lansdowne, Virginia, February 1998.

M. CHRISTEL, T. KANADE, M. MAULDIN, R. REDDY, M. SIRBU, S. STEVENS, H. WACTLAR (1995). Informedia Digital Video Library. *Communications of the ACM*, 38(4):57–58, 1995.

P. CLARKSON, R. ROSENFELD (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2707–2710, Rhodos, Greece, September 1997.

R.A. COLE, R.M. STERN, M.S. PHILLIPS, S.M. BRILL, A.P. PILANT, P. SPECKER (1983). Feature-Based Speaker-Independent Recognition of Isolated English Letters. In *Proceedings of the IEEE 1983 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'83)*, pages 731–733, Boston, Massachusetts, April 1983.

B. COMRIE (ed.) (1987). The World's Major Languages. Oxford University Press, New York, USA, 1987.

G. CORBETT (1987). Serbo-Croat. In B. COMRIE (ed.) (1987), *The World's Major Languages*, chapter 18, pages 391–409. Oxford University Press, New York, USA, 1987.

D.A. DAHL, M. BATES, M. BROWN, W. FISHER, K. HUNCKE-SMITH, D. PALLETT, C. PAO, A. RUDNICKY, E. SHRIBERG (1992). Expanding the Scope of the ATIS Task: The ATIS-3 Corpus. In *Proceedings of the ARPA Workshop on Spoken Language Technology*, pages 3–8, March 1992.

S.B. DAVIS, P. MERMELSTEIN (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 28(4):357–366, August 1980.

K.H. DAVIS, R. BIDDULPH, S. BALASHEK (1952). Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America (JASA)*, 24(6):637–642, November 1952.

S. DELIGNE, F. BIMBOT (1995). Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. In *Proceedings of the IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 169–172, Detroit, Michigan, May 1995.

S. DELIGNE, F. YVON, F. BIMBOT (1995). Variable-Length Sequence Matching for Phonetic Transcription using Joint Multigrams. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 2243–2246, Madrid, Spain, September 1995.

W. ECKERT, T. KUHN, H. NIEMANN, S. RIECK, A. SCHEUER, E.G. SCHUKAT-TALAMAZZINI (1993). A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 1871–1874, Berlin, Germany, September 1993.

P. FETTER, F. CLASS, U. HAIBER, A. KALTENMAIER, U. KILIAN, P. REGEL-BRIETZMANN (1995). Detection of Unknown Words in Spontaneous Speech. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 1637–1640, Madrid, Spain, September 1995.

P. FETTER, A. KALTENMAIER, T. KUHN, P. REGEL-BRIETZMANN (1996). Improved Modeling of OOV Words in Spontaneous Speech. In *Proceedings of the IEEE 1996 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, pages 534–537, Atlanta, Georgia, May 1996.

P. FETTER (1998). Detection and Transcription of OOV words. Ph.D. Thesis, Technische Universität Berlin, Berlin, Germany, August 1998.

E. FINEGAN (1987). English. In B. COMRIE (ed.) (1987), *The World's Major Languages*, chapter 3, pages 77–109. Oxford University Press, New York, USA, 1987.

M. FINKE, T. ZEPPENFELD (1996). Switchboard April 1996 Evaluation Report. In *Proceedings of the 1996 LVCSR Hub5-e Workshop*, Baltimore, Maryland, April 1996.

M. FINKE, J. FRITSCH, P. GEUTNER, K. RIES, A. WAIBEL (1997a). The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proceedings of the 1997 LVCSR Hub5-e Workshop*, Baltimore, Maryland, May 1997.

M. FINKE, P. GEUTNER, H. HILD, T. KEMP, K. RIES, M. WESTPHAL (1997b). The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proceedings of the IEEE 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, pages 83–86, Munich, Germany, April 1997.

J.W. FORGIE, C.D. FORGIE (1959). Results Obtained from a Vowel Recognition Computer Program. *The Journal of the Acoustical Society of America (JASA)*, 31(11):1480–1489, November 1959.

G.D. FORNEY (1973). The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3):268–278, March 1973.

S. FURUI (1986). Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 34(1):52–59, February 1986.

F. GALLWITZ, E. NÖTH, H. NIEMANN (1996). A Category Based Approach for Recognition of Out-of-Vocabulary Words. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 228–231, Philadelphia, Pennsylvania, October 1996.

J.S. GAROFOLO, J.G. FISCUS, W.M. FISHER (1997). Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, pages 15–21, Chantilly, Virginia, February 1997.

J.L. GAUVAIN, L. LAMEL, M. ADDA-DECKER (1995). Developments in Continuous Speech Dictation using the ARPA WSJ Task. In *Proceedings of the IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 65–68, Detroit, Michigan, May 1995.

J.L. GAUVAIN, G. ADDA, L. LAMEL, M. ADDA-DECKER (1997a). Transcribing Broadcast News: The LIMSI Nov96 Hub4 System. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, pages 56–63, Chantilly, Virginia, February 1997.

J.L. GAUVAIN, S. BENNACEF, L. DEVILLERS, L.F. LAMEL, S. ROSSET (1997b). Spoken Language Component of the Mask Kiosk. In K. VARGHESE, S. PFLEGER (eds.) (1997b), *Human Comfort and Security of Information Systems*, pages 93–103. Springer, Germany, 1997.

J.L. GAUVAIN, L. LAMEL, G. ADDA, M. ADDA-DECKER (1997c). Transcription of Broadcast News. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 907–910, Rhodos, Greece, September 1997.

J.L. GAUVAIN, L. LAMEL, G. ADDA (1998). The LIMSI 1997 Hub-4E Transcription System. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 75–79, Lansdowne, Virginia, February 1998.

P. GEUTNER, B. SUHM, F.-D. BUØ, T. KEMP, L. MAYFIELD, A.E. MCNAIR, I. ROGINA, T. SCHULTZ, T. SLOBODA, W. WARD, M. WOSZCZYNA, A. WAIBEL (1995). Integrating Different Learning Approaches into a Multilingual Spoken Language Translation System. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95) - Workshop on New Approaches to Learning for Natural Language Processing*, Montreal, Canada, August 1995.

P. GEUTNER, B. SUHM, F.-D. BUØ, T. KEMP, L. MAYFIELD, A.E. MCNAIR, I. ROGINA, T. SCHULTZ, T. SLOBODA, W. WARD, M. WOSZCZYNA, A. WAIBEL (1996). Integrating Different Learning Approaches into a Multilingual Spoken Language Translation System. In S. WERMTER, E. RILOFF, G. SCHELER (eds.) (1996), *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing - Lecture Notes in Artificial Intelligence*, pages 117–131. Springer, Berlin Heidelberg, March 1996.

P. GEUTNER, M. FINKE, P. SCHEYTT (1997a). Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News. Technical Report CMU-LTI-97-155, Carnegie Mellon University, Pittsburgh, Pennsylvania, December 1997.

P. GEUTNER, R. MALKIN, K. RIES (1997b). The JanusRTk Switchboard/Callhome System - Language Modeling. In *Proceedings of the 1997 LVCSR Hub5-e Workshop*, Baltimore, Maryland, May 1997.

P. GEUTNER, M. DENECKE, U. MEIER, M. WESTPHAL, A. WAIBEL (1998a). Conversational Speech Systems for On-Board Navigation and Assistance. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, December 1998.

P. GEUTNER, M. FINKE, P. SCHEYTT (1998b). Adaptive Vocabularies for Transcribing Multilingual Broadcast News. In *Proceedings of the IEEE 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 925–928, Seattle, Washington, May 1998.

P. GEUTNER, M. FINKE, P. SCHEYTT, A. WAIBEL, H. WACTLAR (1998c). Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 150–155, Lansdowne, Virginia, February 1998.

P. GEUTNER, M. FINKE, A. WAIBEL (1998d). Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, December 1998.

P. GEUTNER, M. FINKE, A. WAIBEL (1999). Selection Criteria for Hypothesis Driven Lexical Adaptation. To appear in: *Proceedings of the*

IEEE 1999 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99), Phoenix, Arizona, March 1999.

P. GEUTNER (1995). Using Morphology Towards better Large-Vocabulary Speech Recognition Systems. In *Proceedings of the IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 445–448, Detroit, Michigan, May 1995.

P. GEUTNER (1996). Introducing Linguistic Constraints into Statistical Language Modeling. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 402–405, Philadelphia, Pennsylvania, October 1996.

P. GEUTNER (1997). Fuzzy Class Rescoring: A Part-Of-Speech Language Model. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2743–2746, Rhodos, Greece, September 1997.

D. GODDEAU, E. BRILL, J. GLASS, C. PAO, M. PHILLIPS, J. POLIFRONI, S. SENEFF, V. ZUE (1994). GALAXY: A Human-Language Interface to On-Line Travel Information. In *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP'94)*, pages 707–710, Yokohama, Japan, September 1994.

J.J. GODFREY, E.C. HOLLIMAN, J. MCDANIEL (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the IEEE 1992 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, Volume 1, pages 517–520, San Francisco, California, March 1992.

J. GVOZDANOVIĆ (1980). Tone and Accent in Standard Serbo-Croatian (With a Synopsis of Serbo-Croatian Phonology). Österreichische Akademie der Wissenschaften, Vienna, Austria, 1980.

A.G. HAUPTMANN, H.D. WACTLAR (1997). Indexing and Search of Multimodal Information. In *Proceedings of the IEEE 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, pages 195–198, Munich, Germany, April 1997.

A.G. HAUPTMANN, M.J. WITBROCK (1997a). Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In M.T. MAYBURY (ed.) (1997a), *Intelligent Multimedia Information Retrieval*. AAAI Press, Menlo Park, California, 1997.

A.G. HAUPTMANN, M.J. WITBROCK (1997b). Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library, May 1997.

A.G. HAUPTMANN, R.E. JONES, K. SEYMORE, S.T. SLATTERY, M.J. WITBROCK, M.A. SIEGLER (1998). Experiments in Information Retrieval from Spoken Documents. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 175–181, Lansdowne, Virginia, February 1998.

J.A. HAWKINS (1987). German. In B. COMRIE (ed.) (1987), *The World's Major Languages*, chapter 4, pages 111–138. Oxford University Press, New York, USA, 1987.

S. HAYAMIZU, K. ITOU, K. TANAKA (1993). Detection of Unknown Words in Large Vocabulary Speech Recognition. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 2113–2116, Berlin, Germany, September 1993.

C.T. HEMPHILL, J.J. GODFREY, G.R. DODDINGTON (1990). The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the 1990 DARPA Speech and Natural Language Processing Workshop*, Pittsburgh, Pennsylvania, June 1990.

I.L. HETHERINGTON, V.W. ZUE (1993). New Words: Implications for Continuous Speech Recognition. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 2121–2124, Berlin, Germany, September 1993.

I.L. HETHERINGTON (1995). New Words: Effect on Recognition Performance and Incorporation Issues. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 1645–1648, Madrid, Spain, September 1995.

X. HUANG, F. ALLEVA, H.W. HON, M.-Y. HWANG, R. ROSENFELD (1993). The SPHINX-II Speech Recognition System: an Overview. *Computer, Speech and Language*, 7:137–148, 1993.

K. HWANG (1997). Vocabulary Optimization Based on Perplexity. In *Proceedings of the IEEE 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, pages 1419–1422, Munich, Germany, April 1997.

F. ITAKURA (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 23(1):67–72, February 1975.

A. ITO, M. KOHDA (1996). Language Modeling by String Pattern N-gram for Japanese Speech Recognition. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 490–493, Philadelphia, Pennsylvania, October 1996.

P. JEANRENAUD, K. NG, M. SIU, R.J. ROHLICEK, H. GISH (1993). Phonetic-Based Word Spotter: Various Configurations and Application to Event Spotting. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 1057–1060, Berlin, Germany, September 1993.

P. JEANRENAUD, E. EIDE, U. CHAUDHARI, J. MCDONOUGH, K. NG, M. SIU, H. GISH (1995). Reducing Word Error Rate on Conversational Speech from the Switchboard Corpus. In *Proceedings of the IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 53–56, Detroit, Michigan, May 1995.

F. JELINEK, R.L. MERCER, S. ROUKOS (1991). Principles of Lexical Language Modeling for Speech Recognition. In S. FURUI, M.M. SONDHI (eds.) (1991), *Advances in Speech Signal Processing*, chapter 21, pages 651–699. Marcel Decker, New York, USA, 1991.

F. JELINEK (1990). Self-Organized Language Modeling for Speech Recognition. In A. WAIBEL, K.F. LEE (eds.) (1990), *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

F. JELINEK (1991). Up From Trigrams! The Struggle for Improved Language Models. In *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech'91)*, pages 1037–1040, Genova, Italy, September 1991.

F. JELINEK (1997). Statistical Methods for Speech Recognition. MIT Press, Cambridge, Massachusetts, 1997.

S.M. KATZ (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 35(3):400–401, March 1987.

T. KEMP, A. JUSEK (1996). Modelling Unknown Words in Spontaneous Speech. In *Proceedings of the IEEE 1996 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, pages 530–533, Atlanta, Georgia, May 1996.

T. KEMP, A. WAIBEL (1998). Reducing the OOV Rate in Broadcast News Speech Recognition. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, December 1998.

T. KEMP, P. GEUTNER, M. SCHMIDT, B. TOMAZ, M. WEBER, M. WESTPHAL, A. WAIBEL (1998a). The Interactive Systems Labs View4You Video Indexing System. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, December 1998.

T. KEMP, M. WEBER, P. GEUTNER, J. GÜRTLER, P. SCHEYTT, B. TOMAZ, M. WESTPHAL, A. WAIBEL (1998b). Automatische Erstellung einer Video-Datenbank: das View4You-System. In *Proceedings of the 4th Conference on Natural Language Processing (KONVENS-98)*, Bonn, Germany, October 1998.

D.H. KLATT (1977). Review of the ARPA Speech Understanding Project. *The Journal of the Acoustical Society of America (JASA)*, 62:1345–1366, December 1977.

H. KLEMM, F. CLASS, U. KILIAN (1995). Word- and Phrase Spotting with Syllable-Based Garbage Modelling. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 2157–2160, Madrid, Spain, September 1995.

R. KNESER, H. NEY (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 973–976, Berlin, Germany, September 1993.

F. KUBALA, H. JIN, S. MATSOUKAS, L. NGUYEN, R. SCHWARTZ, J. MAKHOUL (1997a). Advances in Transcription of Broadcast News. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 927–930, Rhodos, Greece, September 1997.

F. KUBALA, H. JIN, S. MATSOUKAS, L. NGUYEN, R. SCHWARTZ, J. MAKHOUL (1997b). The 1996 BBN Byblos Hub-4 Transcription System. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, pages 90–93, Chantilly, Virginia, February 1997.

F. KUBALA, J. DAVENPORT, H. JIN, D. LIU, T. LEEK, S. MATSOUKAS, D. MILLER, L. NGUYEN, F. RICHARDSON, R. SCHWARTZ, J. MAKHOUL (1998a). The 1997 BBN Byblos System applied to Broadcast News Transcription. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 35–40, Lansdowne, Virginia, February 1998.

F. KUBALA, R. SCHWARTZ, R. STONE, R. WEISCHEDEL (1998b). Named Entity Extraction from Speech. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292, Lansdowne, Virginia, February 1998.

H. KUČERA, W.N. FRANCIS (1967). Computational Analysis of Present-Day American. Brown University Press, Providence, Rhode Island, 1967.

L.F. LAMEL, S.K. BENNACEF, H. BONNEAU-MAYNARD, S. ROSSET, J.L. GAUVAIN (1995). Recent Developments in Spoken Language Systems for Information Retrieval. In *Proceedings of the ESCA ETRW Workshop on Spoken Dialogue Systems*, pages 17–20, Visgo, Denmark, May 1995.

L. LAMEL, S. ROSSET, J.L. GAUVAIN, S. BENNACEF, M. GARNIER-RIZET, B. PROUTS (1998). The LIMSI ARISE System. In *Proceedings of the IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA)*, pages 209–214, Torino, Italy, September 1998.

LANGENSCHEIDTS SPRACHFÜHRER (1997). Kroatisch und Serbisch. ISBN 3-468-22312-9, 1997.

R. LAU, R. ROSENFELD, S. ROUKOS (1993). Trigger-based Language Models: A Maximum Entropy Approach. In *Proceedings of the IEEE 1993 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'93)*, Volume 2, pages 45–48, Minneapolis, Minnesota, April 1993.

W.A. LEA (ed.) (1980). Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, New Jersey, 1980.

K.F. LEE (1988). Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, April 1988.

K.F. LEE (1989). Automatic Speech Recognition: The Development of the SPHINX System. Kluwer Academic Publishers, Boston, Massachusetts, 1989.

D. LIU, L. NGUYEN, S. MATSOUKAS, J. DAVENPORT, F. KUBALA, R. SCHWARTZ (1998). Improvements in Spontaneous Speech Recognition. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 123–126, Lansdowne, Virginia, February 1998.

E. LLEIDA, J.B. MARIÑO, J. SALAVEDRA, A. BONAFONTE, E. MONTE, A. MARTÍNEZ (1993). Out-Of-Vocabulary Word Modelling and Rejection for Keyword Spotting. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 1265–1268, Berlin, Germany, September 1993.

H. LÜNGEN, M. PAMPEL, G. DREXEL, D. GIBBON, F. ALTHOF, C. SCHILLO (1996). Morphology and Speech Technology. In *Proceedings of the ACL-SIGPHON Workshop on Computational Phonology and Speech Technology*, Santa Cruz, California, 1996.

U. MANBER (1989). Algorithms Involving Sequences and Sets. In *Introduction to Algorithms - A Creative Approach*, chapter 6, pages 119–183. Addison-Wesley, Reading, Massachusetts, 1989.

L. MAYFIELD TOMOKIYO, K. RIES (1997). What makes a Word: Learning Base Units in Japanese for Speech Recognition. In *Proceedings of the ACL Workshop on Natural Language Learning*, 1997.

H. NEY, U. ESSEN (1993). Estimating 'Small' Probabilities by Leaving-One-Out. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 2239–2242, Berlin, Germany, September 1993.

H. NEY, U. ESSEN, R. KNESER (1994). On Structuring Probabilistic Dependencies in Stochastic Language Modelling. *Computer, Speech and Language*, 8(1):1–38, 1994.

H.F. OLSON, H. BELAR (1956). Phonetic Typewriter. *The Journal of the Acoustical Society of America (JASA)*, 28(6):1072–1081, November 1956.

D.S. PALLETT, J.G. FISCUS (1997). 1996 Preliminary Broadcast News Benchmark Tests. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, February 1997.

D.S. PALLETT, J.G. FISCUS, W.M. FISHER, J. GAROFOLO, B.A. LUND, A. MARTIN, M.A. PRZYBOCKI (1995). 1994 Benchmark Tests for the DARPA Spoken Language Program. In *Proceedings of the 1995 ARPA Workshop on Spoken Language Technology*, pages 5–36, Austin, Texas, January 1995.

D.S. PALLETT, J.G. FISCUS, A. MARTIN, M.A. PRZYBOCKI (1998). 1997 Broadcast News Benchmark Test Results: English and Non-English. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 5–11, Lansdowne, Virginia, February 1998.

D.S. PALLETT (1991). DARPA Resource Management and ATIS Benchmark Test Poster Session. In *Proceedings of the 1991 DARPA Speech and Natural Language Processing Workshop*, pages 49–58, February 1991.

D.S. PALLETT (1997). Overview over the 1997 DARPA Speech Recognition Workshop. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, February 1997.

D.B. PAUL, J.M. BAKER (1992). The Design for the Wall Street Journal-Based CSR Corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP'92)*, pages 899–902, Banff, Canada, October 1992.

H. PELZ (1996). Linguistik - Eine Einführung. Campe Verlag, Hamburg, Germany, 1996.

J.R. PIERCE (1961). Symbols, Signals and Noise. Harper, 1961.

P. PLACEWAY, J. LAFFERTY (1996). Cheating with Imperfect Transcripts. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 2115–2118, Philadelphia, Pennsylvania, October 1996.

P. PLACEWAY, S. CHEN, M. ESKENAZI, U. JAIN, V. PARIKH, B. RAJ, M. RAVISHANKAR, R. ROSENFELD, K. SEYMORE, M. SIEGLER, R. STERN, E. THAYER (1997). The 1996 Hub-4 Sphinx-3 System. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, pages 85–89, Chantilly, Virginia, February 1997.

X. POUTEAU, L. ARÉVALO (1998). Robust Spoken Dialogue Systems for Consumer Products: a Concrete Application. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, December 1998.

P.J. PRICE, W.M. FISHER, J. BERNSTEIN, D.S. PALLETT (1988). The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition. In *Proceedings of the IEEE 1988 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'88)*, pages 651–654, New York City, USA, April 1988.

P.J. PRICE (1990). Evaluation of Spoken Language Systems: The ATIS Domain. In *Proceedings of the 1990 DARPA Speech and Natural Language Processing Workshop*, pages 91–95, Hidden Valley, California, June 1990.

L. RABINER, B.-H. JUANG (1993). Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

L.R. RABINER, S.E. LEVINSON (1981). Isolated and Connected Word Recognition - Theory and Selected Applications. *IEEE Transactions on Communications (COM)*, 29(5):621–659, May 1981.

L.R. RABINER, R.W. SCHAFER (1978). Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

L.R. RABINER, S.E. LEVINSON, A.E. ROSENBERG, J.G. WILPON (1979). Speaker-Independent Recognition of Isolated Words Using Clustering Techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 27(4):336–349, August 1979.

L.R. RABINER (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 1989.

R. REDDY (1976). Speech Recognition by Machine: A Review. *Proceedings of the IEEE*, 64(4):501–531, April 1976.

K. RIES, F. BUØ, A. WAIBEL (1996). Class Phrase Models for Language Modeling. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, pages 398–401, Philadelphia, Pennsylvania, October 1996.

R. ROSENFELD (1995). Optimizing Lexical and N-gram Coverage via Judicious Use of Linguistic Data. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 1763–1766, Madrid, Spain, September 1995.

P. SCHEYTT, M. FINKE, P. GEUTNER (1997). Multilingual Informedia Project: Speech Recognition on Serbo-Croatian Dictation and Broadcast News Data. Technical Report CMU-LTI-97-154, Carnegie Mellon University, Pittsburgh, Pennsylvania, December 1997.

P. SCHEYTT, P. GEUTNER, A. WAIBEL (1998). Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain. In *Proceedings of the IEEE 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 897–900, Seattle, Washington, May 1998.

P. SCHEYTT (1997). Serbo-Croatian Speech Recognition of Broadcast News within a Multilingual Informedia Project. Master's thesis, Universität Karlsruhe, Karlsruhe, Germany, December 1997.

E.G. SCHUKAT-TALAMAZZINI, T. KUHN, H. NIEMANN (1994). Speech Recognition for Spoken Dialog Systems. In H. NIEMANN, R. DE MORI, G. HANRIEDER (eds.) (1994), *Progress and Prospects of Speech Research and Technology,* Proceedings in Artificial Intelligence, pages 110–120. Infix, 1994.

E.G. SCHUKAT-TALAMAZZINI (1995). Automatische Spracherkennung. Vieweg, Wiesbaden, Germany, 1995.

T. SCHULTZ, M. WESTPHAL, A. WAIBEL (1997). The GlobalPhone Project: Multilingual LVCSR with Janus-3. In *Proceedings of the 2nd SQEL Workshop*, pages 20–27, Plzeň, Czech Republic, April 1997.

K. SEYMORE, S. CHEN, S. DOH, M. ESKENAZI, E. GOUVÊA, B. RAJ, M. RAVISHANKAR, R. ROSENFELD, M. SIEGLER, R. STERN, E. THAYER (1998). The 1997 CMU Sphinx-3 English Broadcast News Transcription System. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 55–59, Lansdowne, Virginia, February 1998.

M. SPIES (1995). A Language Model for Compound Words in Speech Recognition. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 1767–1770, Madrid, Spain, September 1995.

B. SUHM, M. WOSZCZYNA, A. WAIBEL (1993). Detection and Transcription of New Words. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, pages 2179–2182, Berlin, Germany, September 1993.

B. SUHM, P. GEUTNER, T. KEMP, A. LAVIE, L. MAYFIELD, A.E. MC-NAIR, I. ROGINA, T. SLOBODA, W. WARD, M. WOSZCZYNA, A. WAIBEL (1995). Janus: Towards Multilingual Spoken Language Translation. In *Proceedings of the 1995 ARPA Workshop on Spoken Language Technology*, pages 221–226, Austin, Texas, January 1995.

D. VAN COMPERNOLLE (1997). Speech Recognition in the Car: From Phone Dialing to Car Navigation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2431–2434, Rhodos, Greece, September 1997.

A.J. VITERBI (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(4):260–269, April 1967.

N. WAEGNER, S. YOUNG (1992). A Trellis-Based Language Model for Speech Recognition. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP'92)*, pages 245–248, Banff, Canada, October 1992.

W. WAHLSTER (1993). VERBMOBIL: Translation of Face-To-Face Dialogs. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93) - Opening Ceremony and Plenary Session*, Berlin, Germany, September 1993.

A. WAIBEL, K.F. LEE (eds.) (1990). Readings in Speech Recognition. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

W. WARD (1991). Understanding Spontaneous Speech: The Phoenix System. In *Proceedings of the IEEE 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'91)*, pages 365–367, Toronto, Canada, May 1991.

C.L. WAYNE (1998). Topic Detection & Tracking (TDT) - Overview and Perspective. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 191–193, Lansdowne, Virginia, February 1998.

R. WILKINSON, J. ZOBEL, R. SACKS-DAVIS (1995). Similarity Measures for Short Queries. In *Proceedings of the 4th Text Retrieval Conference (TREC-4), NIST*, Gaithersburg, Maryland, November 1995.

J.G. WILPON, L.R. RABINER, A. BERGH (1982). Speaker-Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary. *The Journal of the Acoustical Society of America (JASA)*, 72(2):390–396, August 1982.

P.C. WOODLAND, C.J. LEGETTER, J.J. ODELL, V. VALTCHEV, S.J. YOUNG (1995). The 1994 HTK Large Vocabulary Speech Recognition System. In *Proceedings of the IEEE 1995 International Conference on*

Acoustics, Speech, and Signal Processing (ICASSP'95), pages 73–76, Detroit, Michigan, May 1995.

P.C. WOODLAND, M.J.F. GALES, D. PYE, S.J. YOUNG (1997). The Development of the 1996 Broadcast News Transcription System. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*, pages 73–78, Chantilly, Virginia, February 1997.

P.C. WOODLAND, T. HAIN, S.E. JOHNSON, T.R. NIESLER, A. TUERK, E.W.D. WHITTAKER, S.J. YOUNG (1998). The 1997 HTK Broadcast News Transcription System. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 41–48, Lansdowne, Virginia, February 1998.

S. YOUNG (1996). Large Vocabulary Continuous Speech Recognition: a Review. Technical Report, Cambridge University, Cambridge, England, April 1996.

V. ZUE, S. SENEFF, J. POLIFRONI, M. PHILLIPS, C. PAO, D. GODDEAU, J. CLASS, E. BRILL (1994). PEGASUS: A Spoken Language Interface for On-Line Air Travel Planning. *Speech Communication*, 15:331–340, 1994.

V. ZUE, S. SENEFF, J. GLASS, L. HETHERINGTON, E. HURLEY, H. MENG, C. PAO, J. POLIFRONI, R. SCHLOMING, P. SCHMID (1997). From Interface to Content: Translingual Access and Delivery of On-Line Information. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2227–2230, Rhodos, Greece, September 1997.

V. ZUE (1997). Conversational Interfaces: Advances and Challenges. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97), Keynote Speech*, pages KN 9–18, Rhodos, Greece, September 1997.