

unseen test text was determined through cross validation on all available text data. As a desirable baseline, word accuracy was also tested on a closed-vocabulary scenario yielding a performance of 66.9%.

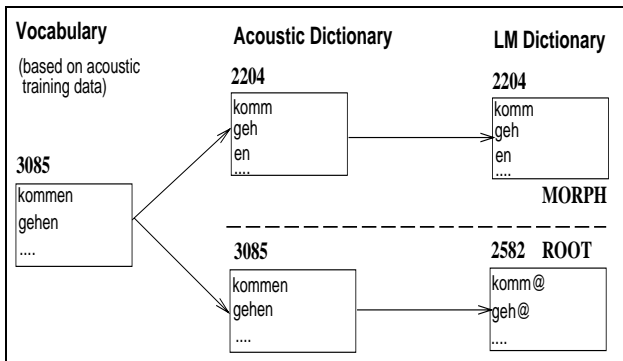


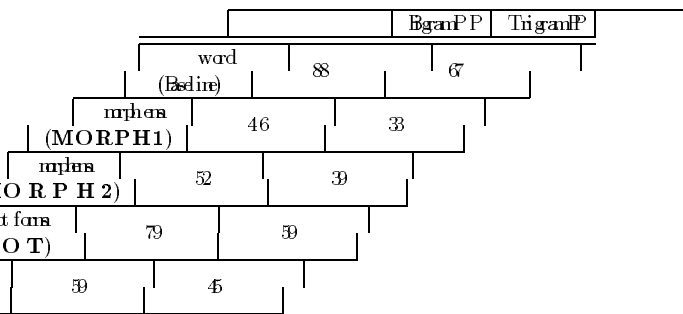
Figure 4. Mapping of Acoustic and Language Modeling Dictionaries during Recognition Process

5.1. Morphem-based Decapitiation

Pure morphem based recognition (as described in MORPH2) measured on word basis slightly outperforms the result achieved with word bigram models by 0.7% (table 6). As the vocabulary size of the acoustic dictionary used within the recognition process is much smaller than on word basis, recognition speed is accelerated by one third.

5.2. Root Form Decapitiation

Using root forms only reduces the original language model dictionary from 3821 words to 3205 root forms. This means a 16% reduction in the vocabulary used as basis for language modeling. The relatively small decrease results in 10% perplexity improvement and thus a slightly smaller language model. However, root forms serve as acoustic dictionary for the recognition process. All inflections also have to be included in the acoustic dictionary, thus increasing the recognition speed.



and Morphem Perplexity

Utterans

n the morphems in- y, as it can be seen range of morphems almost ap- range found in the English language man word coverage of 88%by 3%based on ni ng data.

ng the number of tokens in table 4, we see that he average one word becomes 1.25 tokens within the morphem based framework. All available 225 training di- alogues were used for building two overall language models: One based on words, the other on their morphem decom positions. Smoothing was done by absolute discounting [2] in both cases.

As to be expected the reduction in vocabulary growth leads to a significant perplexity reduction when comparing morphem based language models with word models. Tak- ing into account that only every fourth word has been de- composed the perplexity results are surprising: Mrphem bigramperplexity is 48%lower than word bigram for tri- grams there is a 51%reduction (see table 5).

4.1 Morphem based Composition

Even though perplexity reduction (and also the restric- tion of dictionary growth) is highest when using a s- linguistic-based decomposi ti on of words, (see table 6) are degrading c- recognition process. V- from a very small

Also the German language has an uncountable number of compound words. Nouns can be concatenated to long noun chains, even creating a word with a new meaning, e.g. :

- Sprach-er-kennungs-modul² (speech recognition module)
- Sprach-er-kennungs-genauigkeit (speech recognition accuracy)

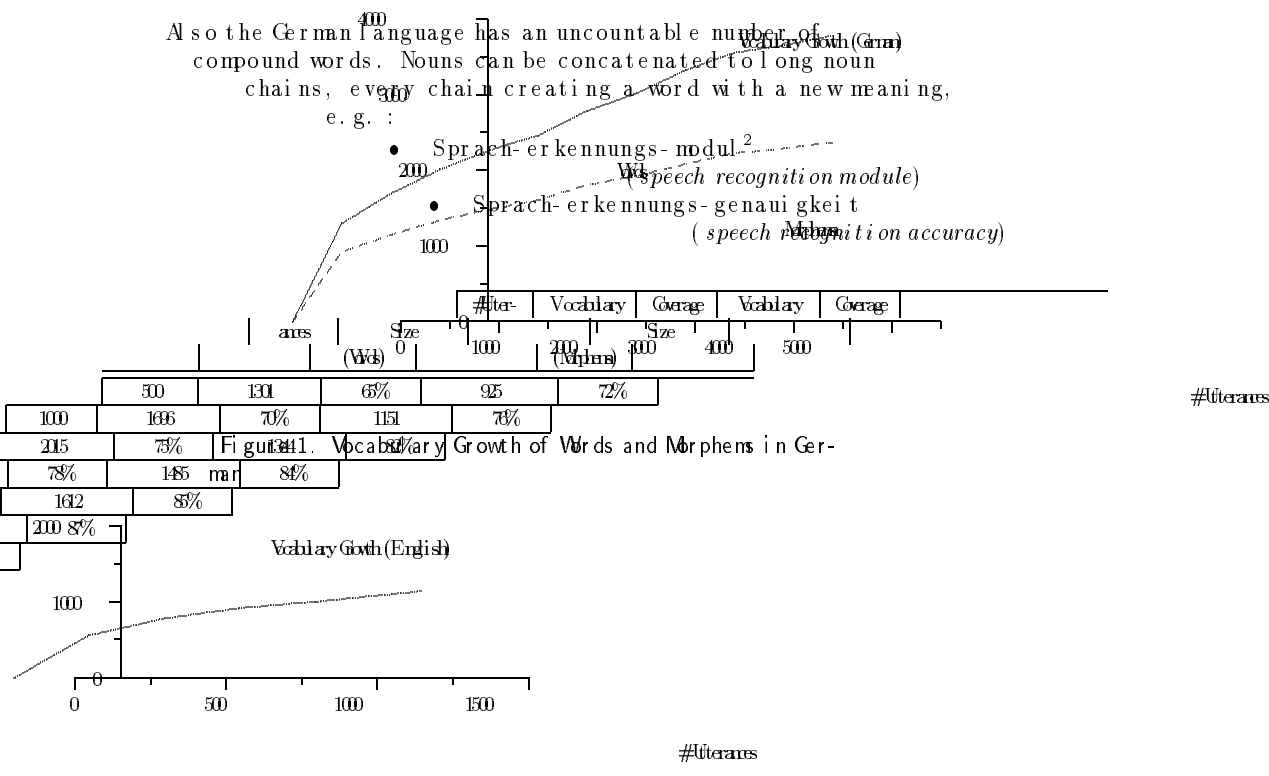


Figure 2. Vocabulary Growth in English

THE MBSOLACEMBS

more robust probabilistic databases and also limit the growth with increasing training material. different ways of decompositions can be performed:

1. strictly morphem based decomposition, e.g. :
 - weggehen → weg-geh-en² (to go away)
 - Spracherkennung → Sprach-er-kenn-ung (speech recognition)
2. decomposition in root forms:
 - weggehen → weggeh@ (to go away)
 - Dialoge → Dialog@ (dialogues)
3. combination of strictly morphem based decomposition and root forms

steadily For the German Spontaneous Scheduling Task (GSST) words that might appear the decomposition of training texts in strictly linguistically can be foreseen. Tables 2 based morphemes (MBS) results in a reduction of vocabulary coverage of the German test based morphemes (MBS) results in a reduction of vocabulary size by 37%(see figure 1). Whereas the word dictionary already covers 92% of English words in the test dictionary consists of only 2391 entries (see table 4). This reduction will certainly get bigger the more data will be available 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.

	Words	Morphemes
#classes	11749	14390
vocabularysize	3821	2391

²Morphemes are used for didactic purposes as decomposition makes only a difference in the actual German spelling Table 4. Comparing Word and Morphem Vocabulary

USING MORPHOLOGY TOWARDS BETTER LARGE-VOCABULARY SPEECH RECOGNITION SYSTEMS

P. Geutner

Interactive Systems Laboratories
Department of Computer Science,
University of Karlsruhe,
76128 Karlsruhe, Germany

ABSTRACT

To guarantee unrestricted natural language processing, state-of-the-art speech recognition systems require huge dictionaries that increase search space and result in performance degradations. This is especially true for languages where there do exist a large number of inflections and compound words such as German, Spanish, etc. One way to keep up decent recognition results with increasing vocabulary is the use of other base units than syllables. This paper differentiates different decomposition methods and compares morphological decomposition with syllable decomposition. Notation and data are given.