Compound Words in Large-Vocabulary German Speech Recognition Systems

André Berton *Pablo Fetter* *Peter Regel-Brietzmann*

Daimler-Benz AG, Research and Technology, Wilhelm-Runge-Str.11
D-89081 Ulm, Germany
e-mail: {berton,fetter,regel}@dbag.ulm.daimlerbenz.com

ABSTRACT

This paper analyzes the impact of German compound words on speech recognition. It is well known that, due to an idiosyncrasy of German orthography, compound words make up a major fraction of German vocabulary. And most **O**ut-**O**f-**V**ocabulary (OOV) compounds are composed of frequent words already in the lexicon. This paper introduces a new method for handling the components of compounds rather than the compounds themselves. This not only reduces the vocabulary, and therefore the perplexity, but also improves word accuracy. And reduced perplexity means a more robust language model.

1. INTRODUCTION

In every language there are three main mechanisms for creating new words: namely inflections, derivations, and compounding. In many languages, including German, inflections and derivations are formed with a limited number of often short prefixes and suffixes. Being short means they are difficult to recognize in isolation because of the coarticulation effect. Being limited in number means that, although the full-form vocabulary size may be very large, it is nonetheless finite. In compounding, however, the set of constituents in every position is potentially unlimited, so that the number of potential compounds is infinite.

There are basically two types of compounds: semantic and orthographic. In semantic compounds, which tend to be written together in many languages, the semantic content of the compound cannot be derived from the semantic content of its constituent parts, e.g. "streetcar" and "post office." In orthographic compounds the constituent parts retain their original semantic content, e.g. "apple tree" which in German is written "Apfelbaum." German orthography requires that both types of compounds be written together, thus resulting in an exceedingly large and potentially unlimited set of compounds.

How can this knowledge be used to reduce the vocabulary size in German? A complete morphological decomposition would significantly reduce vocabulary size, but due to the shortness of many (in German most) inflectional and derivational prefixes and suffixes, it would also reduce the performance of the acoustical recognizer[1]. For compound words this performance degradation does not apply since the components are normally sufficiently long while the benefit of reducing vocabulary size still holds. For these reasons we decided to focus only on compound words.

Our study is based on the Verbmobil database [5] (called the German Spontaneous Scheduling Task), which consists of over 400 human-to-human dialoges collected at various German universities. The entire corpus contains over 200,000 words, including such "non-words" as word fragments, repairs, etc.

We will first present the results of the analysis on our training and test data to evaluate the importance of compound words. Then we will describe a method for recognizing components and generating compounds from them. Afterwards we will show how the results can be optimized. Finally we will present some preliminary results of detecting OOV compounds.

2. FREQUENCY ANALYSIS

We first analyzed word frequencies in a large text corpus (Evaluation '95 in the Verbmobil project) in order to show that a significant vocabulary reduction can be achieved by decomposing compound words. The training and test data contain about 96,000 and 7,000 words of running text and a vocabulary of about 3,000 and 1000 words respectively.

Next, the relative frequencies of compound words within the training and the test material were analyzed. The weighted and unweighted results are shown in Table 1. As can be seen, even though the number of compound words in the running text is relatively low, they make up a major fraction of the vocabulary. Also note the relatively large number of compounds in the set of OOV words.

Since we are particularly interested in utilizing these morphological methods to detect OOV words, all OOV compounds of the test set were analyzed (see Figure 1). It was found that 54% of these compound words are formed from known components, which means although a compound may be rare, its

freq. of comp. in →	training	test	OOV in test
weighted	11.4%	12.0%	42.9%
unweighted	39.5%	29.9%	41.6%

Table 1: Relative frequencies of compound words

components more likely than not will be frequent enough to be in the lexicon.

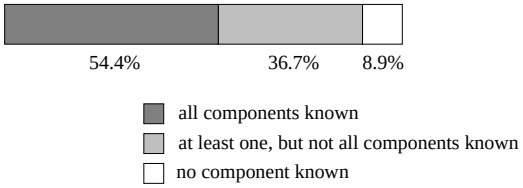

Figure 1: Percentage of compounds with known components

We were also interested in how word decomposition effects vocabulary size. For this, the official Verbmobil lexicon database[2] was used. When all compounds in the vocabulary are replaced by their component parts, the vocabulary was reduced by about $\approx 24\%$ to 2,561 entries (see Table 2).

In order to study the importance of compounds in very large vocabularies, we used a database from the TAZ newspaper published in Berlin. This database contains the text of everything published by TAZ from 1988 to the present—the largest collection of German texts in electronic form. It contains about 76 million words of running text. We constructed several vocabularies, each one consisting of the n most frequent words in the TAZ corpus, and then determined[1] the number of compounds in each vocabulary. Figure 2 shows that the percentage of compounds based on vocabulary size (weighted percentage) increases faster than the percentage of the same compounds in the running text (unweighted frequency analysis). From this we conclude that with increasing vocabulary the decrease in lexicon size will become all the more significant.

3. RECOGNITION PERFORMANCE

3.1. Component Recognition

First, all compound words in the Verbmobil lexicon database[2] were decomposed into their component parts. The remarkable reduction of the vocabulary size was already mentioned in Section 2. A bigram language model based on these components was then trained and tested on unseen

[1] For this database we used the program *Mikrogrammatik*[4], which decomposes any German word along phonological lines. For the rest of this work we used the hand-corrected Verbmobil lexicon database[2].

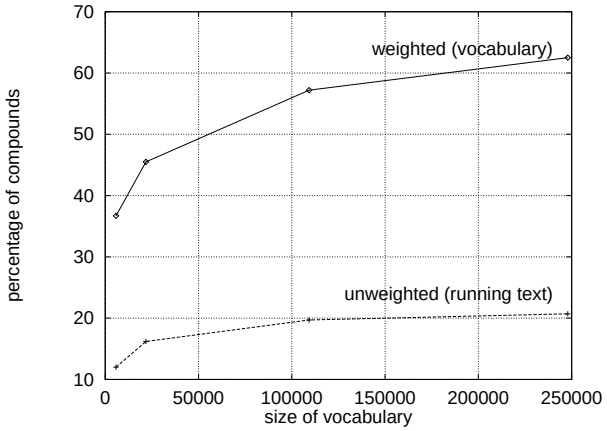

Figure 2: word vs. component vocabulary size

texts (again the material of Evaluation '95). This language model turned out to be much more robust than the word-based one. Both the perplexity (PP) and OOV rate were significantly reduced (see Table 2). The reduction in perplexity was the result of a smaller vocabulary, a lower OOV rate, and a more robust language model. Of these three, the robust language model contributed the most to this reduction.

Sys	*vocab.*	*OOV*	*perplex.*	*word accuracy* (1st best)
Word	3350	2.5%	117	62.7%
Comp	2561	1.6%	77	64.0%
Δ	-23.6%	-0.9%	-34.2%	+1.3%

Table 2: word vs. component based system

After running the HMM-based recognizer on both the word and the component-based system, first-best strings as well as word graphs were produced. The word accuracy results in Figure 3 illustrate the small difference of $\approx 1\%$ between these two systems.

3.2. Word Recognition

In order to provide a fair comparision with the baseline system, the components must be recomposed into compound words. For this, a lexical search on the word graph provided by the recognizer was performed[2]. When a compound candidate is detected by the search algorithm, an additional edge is inserted into the graph, yielding the compound word hypothesis. The acoustic scores of the components were simply added. The following example (see Figure 4) illustrates this method:

[2] In these experiments we chose a word graph with an average density of 30 hypotheses per word, because larger graphs are more likely to contain the right hypothesis. We plan to integrate this search in the second stage of our speech recognition system [3].

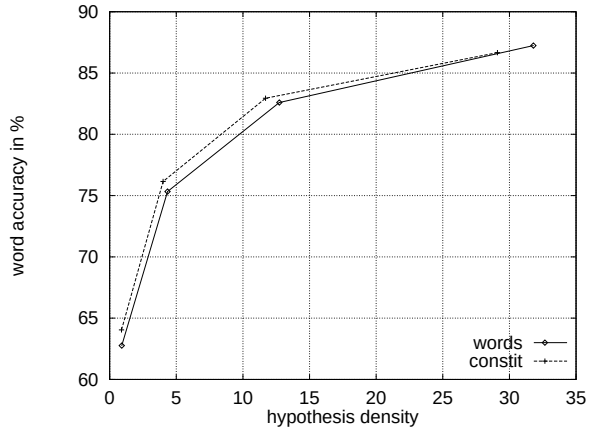

Figure 3: Word accuracy of both systems

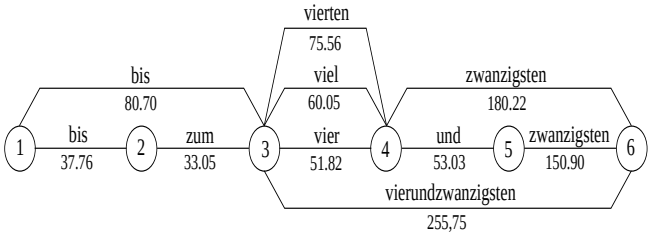

Figure 4: Example of a word graph

- Spoken utterance: *bis zum vierundzwanzigsten*[3]
- Recognizer output: *bis zum vier und zwanzigsten*
- After postprocessing: *bis zum vierundzwanzigsten*

After inserting the compound candidates, the Dynamic-Programming-based pruning algorithm described in [3] was applied with the word-based language model to extract the best sentence. The whole-word recognition system (our baseline) yielded a 62.2% word accuracy[4] after graph pruning, whereas the component-based system could only detect 60.7% of the words in the test set correctly. The 1.5% deterioration in word accuracy was caused mainly by the worse recognition of the individual components of the compound words. If even one of the components is not hypothesized in the wordgraph, then the compound word containing this component cannot be hypothesized.

4. OPTIMIZATION

As opposed to the baseline system where no compounds are decomposed, in the "basic-component" system just described, all compounds were decomposed. The results, however, did not meet our expectations. We then experimented with three criteria for selecting a subset of compounds to be decomposed: phone-based selection, selction based on component frequency, selection based on component interdependence.

4.1. Number of Phones

Some of the compound words in the text contain very short components, which like inflectional and derivational morphs tend to be easily misrecognized. This is particularly true of 2-phone components because of coarticulation effects. In our first experiment we therefore suppressed the decomposition of words containing components with less than a given number of phones. This experiment shows (see Figure 5, *phones*) that word accuracy increases almost linearly with vocabulary size but never exceeds the baseline word accuracy. We thus concluded that this was not a suitable method of optimization.

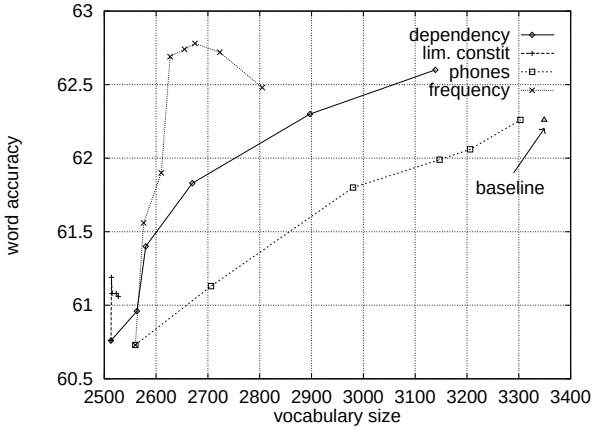

Figure 5: Effect of the optimization stategies (recognition vocabulary size and the word accuracy after recomposing components)

4.2. Frequency of Compounds

Some compounds occur so frequently in the training that their acoustic and language models can be trained well enough. The advantage of a component-based system, namely the higher frequency of each entry, decreases for such words. Hence we concluded that very frequent words should not be decomposed.

This more promising optimization criterion improves word accuracy by 2% with a vocabulary increase of 4% compared to the "basic-component" system, as can be seen from curve *frequency* in Figure 5. For the best system, all compounds that occur more than 15 times were not decomposed. This result is very promising because the vocabulary is 20% smaller and word accuracy increases by 0.5% over the baseline system.

[3] in English: up to the twenty-fourth

[4] These are not the best possible results achieved with our system on the test set, because we usually work with bigger word graphs.

4.3. Component Interdependence

It was found that some components occur only in conjunction with a very limited number of other components. They cannot be freely combined and thus do not increase performance if they are decomposed. **C**omponent **i**nter**d**ependence *cid* is defined as the ratio of the bigram count $C_{(w_1,w_2)}$ to the number of times the components have been seen in the training:

$$cid(w_1, w_2) = \frac{C_{(w_1,w_2)}}{\sqrt{C_{w_1} C_{w_2}}} \quad (1)$$

Optimization based on component interdependence consists of not decomposing those compounds for which their *cid* exceeds a given threshold. The system was tested for several *cid* thresholds. System performance was slightly improved (see in Figure 5, *depencency*), although not as much as for the frequency optimization.

A component w_1 was dubbed *limited*, if the bigram count equaled the absolute count of the component: $C_{(w_1,w_2)} = C_{w_1}$. We tested the system for the case that no compounds with limited components were decomposed. Compared with the non-optimized component-based system, word accuracy improved by 0.5% and the lexicon was reduced by 2% (see Figure 5, *lim.constit*).

5. DETECTING OOV COMPOUNDS

In Section 2 we saw that the components of over half of all OOV compounds are already in the lexicon. None of these OOV compounds was correctly hypothesized as unknown in the systems presented above. One statistical approach to detecting potential OOV compounds consists of determining the probability that a given component can form the head or the tail of a compound. These two probabilities were defined as the frequency that a given component occurred at the head or at the tail of a compound divided by the overall frequency of the component in the test word. In preliminary tests, up to 33% of OOV compounds were detected, but the large number of words incorrectly classified as compounds decreased the word accuracy by $\approx$ 5%, which is absolutely unacceptable.

Compounding is constrained by certain syntactic rules. For example noun-noun compounds are very common in German, whereas verb-verb compounds are very rare. We are therefore now experimenting with a classed-based system consisting of 13 words classes for German. This allows us to generalize the construction of compounds. The bigram probabilities of all class combinations within compounds were computed using the training set. These probabilities describe a measure for the likelihood that two adjacent words of given classes can form a compound. Using several cut-off probabilities for decomposing compounds, we were able to find 5% of the OOV compounds without decreasing word accuracy, 40% with an accuracy decrease of 2%, and 65% with an accuracy decrease of 5%.

None of these preliminary results satisfy our expectations. It has proven to be difficult to generalize from such a small data set. Future experiments will include further attributes of components such as `date`, `place` or `thing` to allow a more precise calculation of potential compounds.

6. CONCLUSIONS

It was found that compound words make up a relatively small fraction of the texts, but a major fraction of the lexicon. We analyzed a way of decomposing compound words into their components, reducing the vocabulary size by 24%. The language model turned out to be more robust. However the word accuracy decreased slightly. We have therefore suggested some optimizations. The most effective method was to suppress decomposition of very frequent compounds which improved word accuracy by 0.5% and reduced the vocabulary by 20%. More work is required to increase the number of OOV compounds correctly hypothesized without decreasing word accuracy. So far, only 5% of these OOV compounds can be correctly hypothesized without decreasing overall word accuracy.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF). Any opinions and conclusions expressed in this paper are those of the authors.

The authors gratefully thank all colleagues that contributed to our large-vocabulary speech recognition system, without which this work could not have been possible: Fritz Class, Alfred Kaltenmeier and Thomas Kuhn. We especially thank David Stall for his invaluable advice and ideas.

Furthermore, we thank the research group of Prof. David Gibbon for making the lexicon database available for our work [2].

REFERENCES

[1] P. Geutner. *Using Morphology Towards Better Large Vocabulary Speech-Recognition Systems.* In Proc. ICASSP'95, pages 445–448.

[2] D .Gibbon. *The Verbmobil Lexicon.* Verbmobil Technical Document 21, Bielefeld, 1995.

[3] T. Kuhn, P. Fetter, A. Kaltenmeier, and P. Regel-Brietzmann. *DP-Based Wordgraph Pruning.* In Proc. ICASSP'96, Atlanta, USA.

[4] H. Mangold, D. Stall und R. Zelinski. *Sprach-Ein-/Ausgabe.* In BMFT-Programm zur Förderung der Datenverarbeitung, Signal- und Mustererkennung, Forschungsbericht, Ulm, 1978.

[5] W. Wahlster. *Verbmobil, Translation of Face-To-Face Dialogs.* In Proc. EuroSpeech'93, Opening and Plenary Sessions, pages 29-38.