Abstract
This paper presents a new approach for grapheme-to-phoneme conversion based on morphology. With this approach, a high accuracy can be obtained, although not for all words a transcription is achieved.

The principle of this approach is to automatically decompose an existing pronunciation lexicon into morpheme-similar units called pseudo-morphological units. The pronunciation of the pseudo-morphological units is also generated from the pronunciation lexicon. The pronunciation of unknown words is composed from the pronunciation of its morphemes. With this approach, it is possible to transcribe also abbreviations and words, where a differentiation between spelling and pronunciation-as-a-word is difficult for other approaches. Another advantage of this approach is that it is performed automatically. It can further be used for any language as long as a word based pronunciation lexicon is available.

1. INTRODUCTION
An important aspect for the use of automatic speech recognition has been its portability across different applications. Especially, the phonetic transcription of words of a new domain is a task that makes it difficult to adapt existing systems to new domains.

The manual phonetic transcription takes much time, therefore many approaches have been pursued in order to automatically generate the phonetic description, e.g. [1], [2], [4].

For a new application, we need a phonetic transcription for the words of the new domain. Phonetic lexica are available with the transcription of names and a general lexicon. However, we need words to be transcribed for the Wap (wireless application protocol) domain. The vocabulary of words of German Wap pages consists mainly of German words, still, a couple of English words and some words consisting of elements of both languages can be found like for example BetAndWin-Wetten, where Wetten is German for bets.

The approach presented here uses subword units similar to morphemes in order to generate an automatic transcription of words. This approach is thought to provide a high accuracy for the words that are transcribed, although this method does not provide a complete coverage, i.e. it does not contain all words to be transcribed. Thus, this approach can be used as an additional step providing a higher overall accuracy in the transcription process.

The following section will present the pseudo-morphological units that have already been used before for language modeling [3]. Especially, the generation of a pronunciation lexicon of these units will be presented.

Section 3 will present the application in the Wap domain and the data used for this study. Section 4 shows the experiments that were carried out and the obtained results. Finally, the future usage of this method as well as possible improvements are discussed in section 5.

2. PSEUDO-MORPHOLOGICAL UNITS
2.1. Pseudo-morphological units
Several approaches have been studied in order to obtain a phonetic transcription of words. However, it is sometimes difficult to obtain the correct pronunciation, especially in languages that have irregular pronunciations. Often, the pronunciation is consistent for the same morpheme in different words, thus the transcription of words using morphemes could be a good solution. The problem with morphemes, however, is, that the decomposition of words into morphemes is also error-prone and therefore errors can occur at another level of the transcription process.

A compromise could be to find units that are similar to morphemes and to use the pronunciation of these units for an automatic transcription of words. Especially for German, which consists of many compound words, the decomposition of words may be useful. The decomposition of words into pseudo-morphological units can be performed automatically [3], thus this algorithm could also be used for other languages without any additional work. For simplification, we will further refer to the pseudo-morphological units as morphemes.

The main principle is to decompose the words of the pronunciation lexicon into morphemes. This decomposition results in a morpheme lexicon, where the pronunciation of each morpheme is given. A word that shall be transcribed phonetically is composed of the morphemes of the morpheme lexicon.
2.2. Construction of morphemes

The morphemes are generated automatically from an existing pronunciation lexicon. Thus, the algorithm can be used for any language, once a pronunciation lexicon is available in that language. Instead of using morphological knowledge for the decomposition, the units are obtained by counting the occurrence of subword units in the pronunciation lexicon.

The current version of this algorithm counts only subword units that are found at the beginning and at the end of a word. The algorithm could be extended to finding subword units also within words, but so far this study is limited to beginnings and endings in order to reduce complexity.

The algorithm is performed in the following steps:

1. look for words that have the same orthographic beginning. If more than $n$ words with the same beginning are found, then accept this as a morpheme.

   altersheim  $\Rightarrow$ altersheime $\Rightarrow$ a l t 6 s h a l m @
   altersheimstaette  $\Rightarrow$ altersheim  $\Rightarrow$ a l t 6 s h a l m S t E t @
   altersklassen  $\Rightarrow$ altersklasse $\Rightarrow$ a l t 6 s k l a s @ n
   alterskontrollen  $\Rightarrow$ alterskontrolle $\Rightarrow$ a l t 6 s k O n t r O l @ n
   alterssicherung  $\Rightarrow$ alterssicherung $\Rightarrow$ a l t 6 s z I C @ r U N
   alters
   altersk
   altersheim are morphemes

2. Check the pronunciation of these words. If they all share the same pronunciation for more than the first $m$ phonemes, take the complete common phoneme sequence as the pronunciation of the morpheme.

   a l t 6 s
   a l t 6 s k
   a l t 6 s h a l m are the corresponding pronunciations

3. Save this morpheme-phoneme combination in the morphological pronunciation lexicon.

   alters
   altersk
   altersheim  $\Rightarrow$ a l t 6 s h a l m are morphemes

4. Repeat 1-3 for endings of words.

   fuerbringer  $\Rightarrow$ f y: 6 b r I N 6
   gebringer  $\Rightarrow$ g e: b r I N 6
   obringer  $\Rightarrow$ o: b r I N 6
   bringer  $\Rightarrow$ b r I N 6

   leads to

   leads to

   The entries in the morphological pronunciation lexicon are not morphemes in the linguistic sense but rather beginnings and endings of words sharing the same pronunciation.

   Depending on the thresholds set for minimal and maximal lengths of morphemes and phoneme sequences, quite different morphemes are obtained.

   Unfortunately, the resulting pronunciation is not always correct. In case of a difference in one of the last phonemes of a bunch of words, often only a part of the pronunciation is recognized as the pronunciation of the complete morpheme. For example,

   altenberg  a l t @ n b E r k
   altenberger  a l t @ n b E r g 6

   leads to the entry

   altenberg  a l t @ n b E r

   In later versions, a control mechanism could be employed that checks on possible pronunciations (e.g. /g/, /k/) of the last letter (e.g. g), and would thus eliminate the last letter of the morpheme.

   Further on, this algorithm is very sensitive to inconsistencies in the used pronunciation lexicon. A very important precondition is to have a correct and consistent pronunciation lexicon. One measure to avoid errors by inconsistencies might be to accept a morpheme for the morpheme lexicon as soon as the majority of the words share the same pronunciation.

   Using a lexicon of 350,000 entries (and $m=n=2$), a morpheme lexicon in the range of 50,000 beginning morphemes and 40,000 ending morphemes is obtained, depending on the parameters used for establishing the lexicon.

3. APPLICATION DATA

3.1. Data to be transcribed

The application where this algorithm was tested, consists of Wap data. The links of Wap pages were collected, starting from German Wap server pages and recursively following the links. Thus, the language in most cases is German. Some words, however, are English like in many computer applications, some words are a mixture of German and English, like the already mentioned BetAndWin-Wetten.

These words contain numbers, other symbols, thus the data must be processed in order to obtain a normalized orthography. This standardized orthography is necessary for comparison and, especially, for a lexicon lookup. We also eliminated numbers. After this processing, we obtained two sets of Wap data obtained with different starting pages:

   wap-1    with 2113 different words
   wap-2    with 977 different words.

3.2. Pronunciation Lexica

The pronunciation lexica are used as a basis for the morphological decomposition. We use two different lexica as shown in Table 1.
The lexicon Names (N) consists of proper names, the lexicon General (G) consists of words from different applications (Verbmobil, Speecon), thus would have a wider range with respect to the vocabulary of a new application. Both lexica cover German words, thus English words of the Wap domain cannot be covered by either lexicon.

The coverage of the wap-1 words in the lexicon is given in Table 1. For example, the pronunciation of 31 % of the Wap words can be found in the lexicon N, 35 % in the lexicon G. Using both lexica, a coverage of 49 % is obtained.

<table>
<thead>
<tr>
<th>LEXICON</th>
<th>N</th>
<th>G</th>
<th>N+G</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Words</td>
<td>329,853</td>
<td>20,916</td>
<td>350,526</td>
</tr>
<tr>
<td>Wap words</td>
<td>31 %</td>
<td>35 %</td>
<td>49 %</td>
</tr>
</tbody>
</table>

Table 1 Coverage of Wap words in two lexica

The remaining words of the Wap domain must be phonetically transcribed with another technology.

4. EXPERIMENTS AND RESULTS

4.1. Experiments with morphemes on words of the lexicon used for decomposition

The quality of the pseudo-morphological units achieved by the decomposition is measured by

- the coverage, i.e. the amount of words whose pronunciation can be obtained by composing the words out of the morphemes (where the morphemes are smaller than existing words).
- the accuracy of the phonetic transcription of words whose pronunciation is generated with morphemes. The accuracy is measured as phoneme accuracy (1−(ins+del+sub/all), in analogy to the word accuracy) and vocabulary accuracy, i.e. the amount of correctly transcribed words.

First, the available pronunciation lexica are used for a decomposition into morphemes, leading to a morpheme-based phonetic lexicon. The coverage and accuracy of the morphemes is first tested on the training lexicon, i.e. on the words that were used for generating the morpheme lexicon.

With n=sm=2 and a minimum of 2 characters for the length of a morpheme the highest coverage and also the highest accuracy was found. For example, with the G lexicon for both decomposition and test we obtain a coverage of 35 %. The phoneme accuracy of the transcribed words is 94 %, and the vocabulary accuracy is 79 %.

4.2. Experiments with Morphemes on unseen words

Now we tested the morphemes obtained from one lexicon with the vocabulary of the other lexicon. Using the G lexicon for the decomposition into morphemes and the N lexicon for test, we obtain a poor coverage of 4 %, whereas the other way (N for decomposition and G for test) we obtained a coverage of 52 %.

This shows that the lexicon that is used for decomposition is essential for the coverage of the vocabulary to be transcribed. If the available lexicon is too specialized it will be difficult to obtain frequently occurring morphemes. On the other hand, if the vocabulary to be transcribed is very special, a similar problem will arise, such as that the needed morphemes have not occurred in the phonetic lexicon.

The accuracy of the transcribed words is quite high again, the phoneme accuracy ranges between 91 and 92 %, the vocabulary accuracy is around 70 %. In contrast to section 4.1, however, the vocabulary accuracy is increased when using longer morphemes only. In this case, of course, the coverage decreases. For example, with a minimal length of 6 characters, a coverage of 3.5 % (in contrast to 4 %; N for decomposition, G for test) is obtained that leads to a vocabulary accuracy of 78 %.

4.3. Experiments with Wap words

4.3.1. Coverage

For the transcription of words of the application, i.e. of words from the Wap domain, we use both lexica as input to the generation of morphemes.

The transcription of the Wap words is done in a three-step-process:

1. lexicon look-up. The word is searched in the pronunciation lexicon resulting from the N+G lexicon.
2. The words remaining after step 1 are transcribed with morphemes from the morpheme lexicon, if possible.
3. The words remaining after step 2 are transcribed with an robust algorithm that provides a transcription for any word. Here, we used a standard grapheme-to-phoneme converter, based on decision trees.

From the wap-1 data,

50 % were found in the N+G, from the remaining words, 13 % were found with morphology (using N+G), 37 % were transcribed with the grapheme-to-phoneme converter.

The wap-2 data were additionally searched in the wap-1 lexicon:

48 % were found in N+G, 9 % were found in the wap-1 lexicon, 33 % were transcribed with the grapheme-to-phoneme converter.

When adding the wap-1 lexicon for generating morphemes, we obtain the following coverage:

48 % were found in N+G, 9 % were found in the wap-1 lexicon, 32 % were transcribed with the grapheme-to-phoneme converter.

These results show that with the small wap-1 lexicon from the same domain, additional 9 % are found with a lexicon lookup. Furthermore, 1 % is additionally obtained with morphology when the Wap lexicon is included in the generation of morphemes.
4.3.2. Accuracy

The accuracy of the morpheme transcription uses the lexicon entry compared to a manual phonetic transcription of the wap-1 words as reference. Table 2 shows the performance of the morpheme transcription and of the grapheme-to-phoneme converter.

<table>
<thead>
<tr>
<th>Morphology</th>
<th>Grapheme2Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme/Vocabulary</td>
<td>Phoneme/Vocabulary</td>
</tr>
<tr>
<td>89 %</td>
<td>46 %</td>
</tr>
<tr>
<td>79 %</td>
<td>21 %</td>
</tr>
</tbody>
</table>

Table 2 Accuracy of morpheme transcription

The phoneme accuracy of the morpheme transcription is 10% higher than using the grapheme-to-phoneme converter. The amount of correctly transcribed words is more than twice higher for the morpheme transcription.

To a certain extent, the morphology approach is in some way domain-dependent. A large amount of the unseen words in German is caused by inflected word forms and compound words which modify the frequent words of a domain. Thus, the coverage of the morphology approach is higher when transcribing words of a known domain.

If the pronunciation of a word can be obtained by a combination of morphemes the accuracy is higher than using only a grapheme-to-phoneme converter. The quality of the transcription depends much on the quality of the lexicon that is used for the decomposition into morphemes. Errors in the lexicon are not detected and are further used also in the morphemes. Due to the algorithm of decomposition, morphemes cannot be found if the lexicon provides different pronunciations for the same morpheme.

A big advantage of the morpheme approach is the use of domain-specific information for the pronunciation of abbreviations, e.g. in the transcription of words containing WWW, Url or Wap. Another approach like the grapheme-to-phoneme converter would have to decide if the spelling mode or a pronunciation as a word should be used which could cause transcription errors.

5. DISCUSSION

We have presented a new algorithm for obtaining a phonetic transcription of new words. This algorithm makes use of a transcription lexicon and uses pseudo-morphological units of the words in the lexicon for the transcription of new words. This algorithm is especially useful for an inflecting language like German, since the pronunciation of a word differing e.g. only in the ending can be used for transcription.

This algorithm should work well for inflecting languages like Finnish and Turkish and with a word-based lexicon, since in this case morphemes with a reliable pronunciation can be generated.

The algorithm does not provide a complete coverage of the pronunciation for all words, thus there must always follow an algorithm that works for any word.

The algorithm performs better than the robust algorithm it was compared with, but there are some improvements to be made:

1. the lexicon itself should be improved, since there are some inconsistencies in the pronunciation. In German, especially, the transcription often differs between the phonemes /s/ and /z/.
2. Words listed several times in a lexicon due to pronunciation alternatives, have not been taken into account yet.
3. Post-processing of pseudo-morphemes could avoid errors like in the above shown example of Altenberger.

Furthermore, it would be best to increase the lexicon size used for developing the morpheme lexicon. With words of different applications, the coverage of the morpheme of a language can be increased and with that the coverage of the transcription algorithm.

A major advantage of this algorithm is that it can be used for any language without needing any knowledge of that language. No manual work is needed for generating a pronunciation lexicon of words of a new language. The only need of this algorithm is to have a pronunciation lexicon that is as large as possible.

The largest advantage of this algorithm is that, in a certain way, knowledge is incorporated in this completely automated algorithm. Having words of the domain already transcribed, the algorithm is capable of correctly transcribing words or abbreviations that would hardly be transcribed correctly with any other algorithm.

Still, this algorithm will never provide a complete coverage of all words to be transcribed, thus, an additional step for the remaining untranscribed words is necessary with a robust algorithm that produces a good transcription for any word.

6. REFERENCES