MORPHOLOGICAL DECOMPOSITION FOR ASR IN GERMAN

Martine Adda-Decker & Gilles Adda

Spoken Language Processing Group, LIMSI-CNRS, Orsay, France
{madda,gadda}@limsi.fr, http://www.limsi.fr/TLP

Abstract

In this contribution we report on our ongoing work in lexical decomposition for automatic speech recognition (ASR)[1]. Lexical decomposition is investigated with a twofold goal: lexical coverage optimization and improved automatic letter-to-sound conversion. Whereas morphological decomposition is a widely-studied domain in linguistics, our interest is limited here to identifying and processing the statistically most relevant sources of lexical variation in text corpora. Lexical variation is shown to be particularly important for nouns, due to compounding. A set of about 340 decomposition rules has been developed using statistics from 300M words from different newspaper sources (primarily 14 years from the TAZ, the Berliner TAgesZeitung). The out-of-vocabulary (OOV) rate on the same 300M words is reduced from 5.2 to 4.6% in case-sensitive form and to 4.2% in case-insensitive form. For letter-to-sound conversion cross-morpheme letter sequences are a major source of ambiguity. Decomposition, by reducing these ambiguities, contributes to producing more consistent phonemic transcriptions for pronunciation dictionaries.

1. Introduction

The German language, more than other major western languages, produces a large variety of distinct lexical forms. This characteristic raises specific research issues for German speech recognition. In previous work on German ASR the problem of lexical

[1] Part of this work is funded by the European LE-OLIVE project on multilingual broadcast transcription.

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 129-143.

coverage and morphological decomposition has been highlighted (Adda-Decker et al., 1996).

Generally ASR systems have to meet the following requirements to enable good performance: vocabulary, as well as acoustic and language models must achieve good coverage during the system's operating conditions. The vocabulary should contain all or most words likely to appear during operation, with the out of vocabulary (OOV) rate staying minimal. ASR systems typically make use of full form word lexica, where each lexical entry is described by one or more pronunciations (phonemic transcriptions). Most large vocabulary continuous speech recognition (LVCSR) systems are limited to 65k lexical entries. Recognition performance in terms of the measured word accuracy rate is known to depend on the coverage of the system's lexicon during operation. Low lexical coverage necessarily results in high word error rates, hence the motivation for lexical coverage maximization. Recognition performance is also dependent on the acoustic model quality. Acoustic models have to accurately model the vocabulary words. A required condition for accurate acoustic phone model estimation is a good pronunciation lexicon which makes consistent use of the same phone symbol sequence for identically pronounced sub-words.

Whereas morphological decomposition is a widely-studied domain in linguistics (Hausser, 1994), our interest in decomposition is limited here to identifying and processing the statistically most relevant sources of lexical variation in text corpora. In the following we start with a description of lexical variation as observed in 'standard' German and then, more specifically, in the processed corpora. Next, decomposition is addressed, with a first section describing different methods for decomposition rule development followed by a results section, where we want to quantify gains in coverage by applying partial morphological decomposition. For letter to sound conversion only qualitative results are provided.

2. Lexical variation in German text sources

The German language produces a large variety of distinct lexical forms. To give an idea of lexical variation across different languages we have reproduced some of the results obtained in former multilingual studies (Young et al., 1997; Lamel et al., 1995) using comparable amounts of text corpora[2] in Table 1. Coverage figures of Japanese

[2] The newspaper text corpora compared are the *Wall Street Journal* (American English), *Le Monde* (French), *Frankfurter Rundschau* (German) from the ACL-ECI cdrom, *Il Sole 24 Ore* (Italian) and *Nikkei* (Japanese).

(Matsuoka et al., 1996) are very close to those obtained for Italian. Whereas the highest lexical coverage (close to 100% for a 65k vocabulary) is achieved for English, German has the highest OOV rate of about 5%.

Table 1. Comparison of 5 languages (*Frankfurter~Rundschau* with *WSJ, Il Sole 24 Ore, Le Monde* and *Nikkei* text corpora) in terms of number of distinct words and lexical coverage of the text data for different lexicon sizes. OOV (Out Of Vocabulary) rates are shown for 65k lexica.

	German	**English**	**Italian**	**French**	**Japanese**
corpus	*FR*	*WSJ*	*Sole 24*	*Le Monde*	*Nikkei*
#words	36M	37.2M	25.7M	37.7M	180M
#distinct	650k	165k	200k	280k	623k
5k cover. %	82.9	90.6	88.3	85.2	88.0
20k cover. %	90.0	97.5	96.3	94.7	96.2
65k cover. %	95.1	99.6	99.0	98.3	99.2
65k-OOV %	4.9	0.4	1.0	1.7	0.8

The written corpora used in this study come from different sources of news and newspaper texts with a total of about 300M words. The major sources are the following: **Deutsche Presse Agentur (German Press Agency)** with about 30M words (years 1993-96, distributed by the LDC), **Frankfurter Rundschau** newspaper (about 35 M words) from the ECI (European Corpus Initiative), Berliner TAgesZeitung (**TAZ**) with about 150 M words (years 1986-99) purchased directly from the newspaper **Die Welt**, years 1996-98, including 20 M words obtained via the Web.

These various text sources are gathered in different formats with different mark-ups. Therefore each source requires different manipulations. Once the texts are available, further cleaning and processing are needed to prepare them for vocabulary selection and language modeling. Normalization decisions are generally language-specific (Adda et al., 1997; Habert et al., 1998). Much of speech recognition research for American English has been supported by ARPA and has been based on text materials which were processed to remove upper/lower case distinction and compounds. We have chosen to maintain case distinction for German. German text

normalization is still under development. At present all sentence-initial words are registered with their capital initial. After a rough text preprocessing (sentence and word segmentation) a total of 300 M words (running text) produces an exhaustive word list of about 2.6 M different lexical items. In the list of 65k most frequent words all items occur at least 163 times (Wirtschaftsboom, Engl. economy boom), with the highest number of occurrence of several millions for articles and conjunctions (8.5 M (der), 8 M (die), 7 M (und)). There is a total of about 500k words which occur at least 5 times and more and thus more than 2 M words occurring less than 5 times in the texts.

This high lexical variation can be explained by various known factors, such as word compounding, inflexion and derivation. For instance, a given adjective in German may be found with more than 10 distinct forms in a speech recognizer's lexicon. These are due to declensions, which may be combined with comparative and superlative forms (e.g.: selbstbewußt, selbstbewußte, selbstbewußten, selbstbewußtes, selbstbewußtem, selbstbewußter, selbstbewußtere, selbstbewußteren, selbstbewußtesten, selbstbewußteste...). The word compounding process is significantly more powerful than inflexion, as it is generative. Theoretically, compounding can produce an infinite number of lexical items.

In addition the German language admits graphemic variability in some particular situations, e.g. ß-writing (words written with ß according to proper German orthography often appear with ss, which is a generally accepted (but non-standard) written form, declension (e.g. Genitive -s or -es: Ausstand-s (#144)[3] or Ausstand-es (#145)), compounding (characters may appear or disappear when items are linked together). No clear rules exist for the use and the choice of such characters, called *Fugenelemente* (s, n, e...). Cases occur where two given words may be compounded with or without *Fugenelemente*. For example Bahnhof (Engl. station) compounds are formed using a Fugen-s in a large majority of cases, but for the compound Bahnhof (Engl. station), Straße (Engl. street) the version without -s-, Bahnhof-straße (Engl. station street) (#359), appears significantly more often than the expected version Bahnhof-s-straße (#49). Semantic differences may sometimes explain variants in the *Fugenbildung* process (Bauersfrau (#43), Bauernfrau (#4)). We can cite other types of compounding variants: Aussterbe-rate or Aussterben-s-rate, Autoren-schaft (#115) or Autor-schaft (#80), nouns with more than one capital letter:

[3] (#N) indicates the number of occurrences in the 300 M word text corpus.

HipHop (#2626), SonntagsZeitung (#184), MultiKulti (#81), StudentInnen (#4923). Word-internal capital letters are particularly frequent in the Berliner TAZ. Although the latter forms are not officially correct, they may facilitate decompounding (and understanding) while reading. Especially the feminine ending -Innen is more frequent than the classical form -innen. There are more than 20k distinct items ending with -Innen (together occurring more than 200k times) and less than 17.5k items ending with -innen (occurring 150k times). According to a 3 consonant rule, only 2 of them are kept if compounding generates a sequence of 3 identical letters: Schiff⊕Fahrt → Schifffahrt, Ballett⊕Tänzer → Balletttänzer. Special care must be taken when decompounding such items. The German orthographic reform cancels the 3 consonants rule, recommending the writing of all the 3 letters. In our text corpora Schifffahrt occurs about 2000 times, Schifffahrt about 100 times. The orthographic reform, which was worked on for several decades and has been official since around 1996, will certainly increase graphemic variability in the coming years (due to different rules being applied by different parts of the population).

Other less linguistic reasons which help to explain part of the observed variation are typing errors and insufficient text normalization. Foreign words and long compounds are particularly prone to misspellings. There may be consonants missing at boundaries, missing Fugen-s, truncated endings (Aussiedlerwohnun) and spurious or swapped letters. For the entry Auseinandersetzungen more than 10 different (wrong) forms could be found: Ausneinandersetzungen, Ausenandersetzungen, Auseinsetzungen, Auseinsandersetzungen, Auseindersetzungen, Auseinderansetzungen, Auseinansetzungen, Auseinansersetzungen, Auseinandesetzungen, Auseinanderzu-setzungen, Aauseinandersetzungen, AAuseinandersetzungen. It is difficult to give an accurate estimate of the rate of typing errors in the 2.6 M word list. In our opinion at least 0.3 M entries can be considered as errors.

From all cited sources of lexical variation, the compounding of nouns seems to be one of the most important. For example, more than 4000 distinct compounds, all starting with the noun Stadt can be found in our newspaper corpus (e.g. Stadt-teil, Stadt-bibliotheks-leiterin, Stadt-bezirks-verordneten-versammlungen). Some words compound very easily, others tend to stay isolated. The word Beduinen, while rarely occurring on its own in the corpus (#280) can nonetheless be found in more than 20 distinct compound forms. The word Bedrängnis occurring more than 1250 times as such or in an inflected form, appears in only one compound Bedrängnisgesängen. A good indicator for

compounds is word length. Figure 1 compares lexical variation as a function of word length (in number of characters) in German, French and English using similar text corpora of about 300M words each. French and English curves are similar with a peak for word length of 7. The higher French curve can be attributed to the French inflexion mechanisms. Both languages are known to produce only few compounds. The German curve is always higher than the other two and increases even more steeply after word length of 9. Compounding is probably the main reason, especially noun compounding.

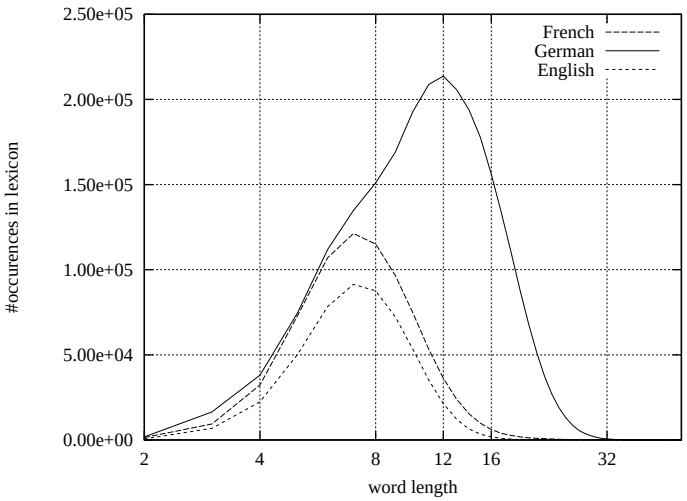

Figure 1. Comparison of lexical variation in English, French and German. The number of distinct lexical items as a function of word length (in # of characters) is measured from news text & transcripts corpora of about 300M words per language.

It is relatively easy to measure the proportion of nouns (roughly speaking, words starting with an upper-case letter) in German. There are about 2M of these items (80%). In Figure 2 separate curves are added for capital initial words (UC curve) and lower-case words (LC curve).

The left part of Figure 2 plots lexical variation in the word list (#occurrences in lexicon). The right part gives the same information weighted over the text material #occurrences in text). The highest lexical variation can be observed for words of 12 characters. Nouns produce significantly more different forms than the other word classes, but are globally less frequent in running text. For words of length 8 and more, UC words are more frequent than LC words. In the following we have mainly concentrated our decompounding efforts on words starting with upper-case letters.

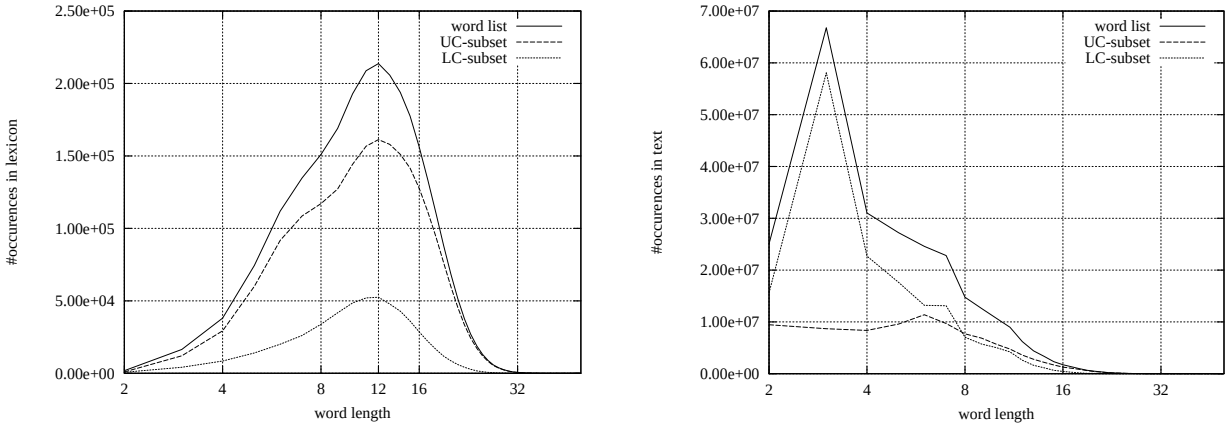

Figure 2. **Left:** number of distinct words as occurring in the exhaustive word list against word length. **Right:** number of observed occurrences in the text corpora.

3. Decomposition

As stated above our goal here is twofold. Firstly, decomposition aims at lexical variation reduction, so as to increase coverage of fixed-size ASR lexica and to reduce the sparse data problem for language model training. Secondly, known morpheme boundaries are helpful for improving letter-to-sound conversion, resulting in more consistent pronunciation lexica and hopefully in more accurate acoustic phone models.

We carried out a brief, and certainly incomplete, investigation of existing resources for automatic morphological decomposition, and solutions seemed either too limited or too slow for the very large corpora we wanted to process. But we must admit that we were mainly interested in looking into the quite complex decomposition problem, rather than applying any readily available solution.

3.1. Methods

Three approaches have been partially explored. Two approaches aim at developing **word-based** decomposition rules, using word counts to select the potentially most promising decomposition rules. A third set of rules gathers more **general** decompositions.

- **A-set rules**

  All items starting with a capital A in the 65k word list have been checked, and decomposition rules have been developed manually. For a total of about 3k such items a set of 260 rules has been elaborated. Example rules are given in Table 2. In general the matching string is limited to one word, but word sequences are also accepted (`Arbeits minister`). If a left-justified matching string is found, it is split off from the remaining string, except if the remaining part matches one of the possible exceptions. For the moment too short morphemes (typically one syllable) are not decompounded (e.g. `Alt` in `Altpapier`).

Table 2. Example decomposition rules. A rule has two parts: the left part contains the matching string, the right part a list of exceptions (which may remain empty).

matching string		**exceptions**
Abend	/	s|es|e|en
Altpapier	/	s|es|e|en
Arbeitslosen	/	
Arbeits minister	/	s|ium|iums|ien|in|In|n ~_
Ausreise	/	n|ns|r|rn|rin|rIn|rinnen|rInnen|nde|nden|nder
Ausverkaufs	/	

- **Most frequent word starts**

  A similar approach to derive decomposition rules based on the most frequent word starts of a given length has been explored. The word start length has been fixed to at least 8 characters. These word starts are checked manually to develop appropriate rules. A set of 70 rules has been applied. Example rules may be found in Table 3. Most rules concern upper-case word starts, even if some rules occur for lower-case words (`zusammen`, `entgegen`, `herunter`...). This approach is very effective both for speeding up the rule derivation process and for the measured lexical variation reduction.

Table 3. Example decomposition rules for the most frequent word start approach

matching string		exceptions
Wirtschafts	/	
zusammen	/	
Verkehrs	/	
Sicherheits	/	
Familien	/	
Betriebs	/	
Friedens	/	
Menschen	/	s
Computer	/	n|s
Informations	/	

- **General rules**

  A limited set of general rules could be identified. A morpheme boundary can be hypothesized after the occurrence of sequences such as `-ungs`, `-hafts`, `-lings`, `-tions`, `-heits` with a very reduced number of exceptions (e.g. `-ungs` occurs as a boundary in more than 130k distinct forms: `Regierungschef`, `Führungstor`..., `-tions` in 32.6k and `-hafts` in 13.5k distinct forms). These general rules, all with an `s`-ending, clarify the `s`-pronunciation during letter-to-sound conversion.

3.2. *Results*

In table 4, we summarize the decomposition figures obtained on the capital A word set. Whereas only 3k such words are in the 65k lexicon, 136.6k distinct items are in the complete text corpora. This number is severely reduced (78.2k) by applying only 260 different decomposition rules. Lexical coverage is hence significantly improved on the A subset. The global OOV reduction of the new (decompounded) 65k word list corresponds to a 2.5% proportion. This partial result is very encouraging to continue the development of decomposition rules for the complete 65k list.

Table 4. Description of the **A word set** (words starting with capital A in the 300M text corpus) before and after applying the 260 decomposition rules. The coverage is computed on the full corpus.

	original A set	**decompounded A set**
#distinct	136.6k	78.2k
#distinct (#occ. > 5)	26.6k	16.4k
#total	7.44M	7.72M
#total (#occ. > 5)	7.26M	7.62M
%coverage (65k, full corpus)	94.8	94.9

In Table 5, we compare the results of the 3 different decomposition rule sets (which are still under development) on the whole text corpus. Results are given for the whole word list, the word list comprising all words occurring at least 5 times (#occ. > 5) and for the 65k most frequent word lists, which represent the portion of the vocabulary typically modeled in an LVCSR system. Combining the different rule sets (*all rules*) allows the individual gains in coverage to be summed.

Table 5. Comparative results between *A set rules, most frequent rules, general rules.* The global results obtained by the combination of the 3 rule sets (*all rules*) correspond to the last column.

	original	**A set rules**	**most freq. rules**	**general rules**	**all rules**
#rules		260	70	5	335
#distinct	2,620k	2,576k	2,543k	2,463K	2,352k
#distinct (#occ. > 5)	538k	529k	522k	519k	495k
#total	314.4M	315.4M	316.1M	316.2M	318.5M
#total (#occ. > 5)	311.0M	312.1M	312.8M	313.1M	315.5M
#total (65k)	297.9M	299.3M	300.3M	300.3M	303.8M
%coverage (#occ. > 5)	98.9	99.0	99.0	99.0	99.1
%coverage (65k)	94.8	94.9	95.0	95.0	95.4

In table 6, we summarize the global decomposition results in terms of OOV rates and measure the impact of case-insensitivity. Despite the small set of decomposition rules the OOV rate is seen to be reduced significantly. An additional important OOV reduction may be achieved using case-insensitive text.

Table 6. Summary of lexical variation measures (#distinct items) for case-sensitive and case-insensitive texts before and after decomposition with the *all rule* set along with the ASR related OOV measures. #total indicates the total number of words in the corpus.

	#distinct (M)	**#total (M)**	**%OOV 65k**
original	2.62	314.35	5.2
original (case i.)	2.47	314.35	4.8
all rules	2.35	318.52	4.6
all rules (case i.)	2.20	318.52	4.2

It is interesting to compare the word lengths of the original word list with the decompounded word list. Figure 3 (left) shows that the number of distinct items is significantly lower for word lengths in the range of 11 to 20 characters. A zoom of this region is given in the right part of Figure 3. As expected, the curves corresponding to the capital initial words display the most significant differences. Figure 4 gives the same information weighted over the text material (#occurrences in text). For short word lengths the decompounded text produces more occurrences than the original text, whereas for word lengths above 12 occurrences less are observed.

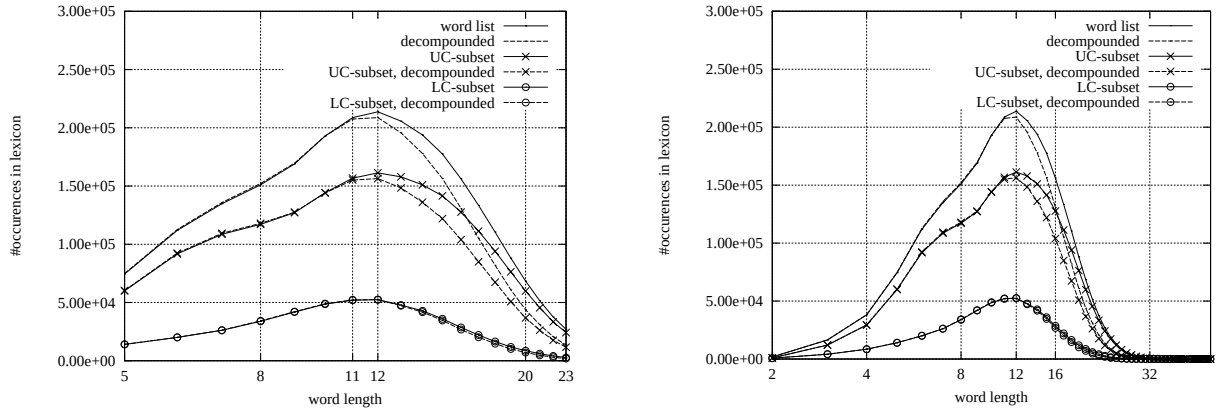

Figure 3. **Left**: Number of distinct words (as occurring in the exhaustive word lexicon) against word length. **Right**: A zoom is carried out on the region comprising words in the range of 5 to 23 characters.

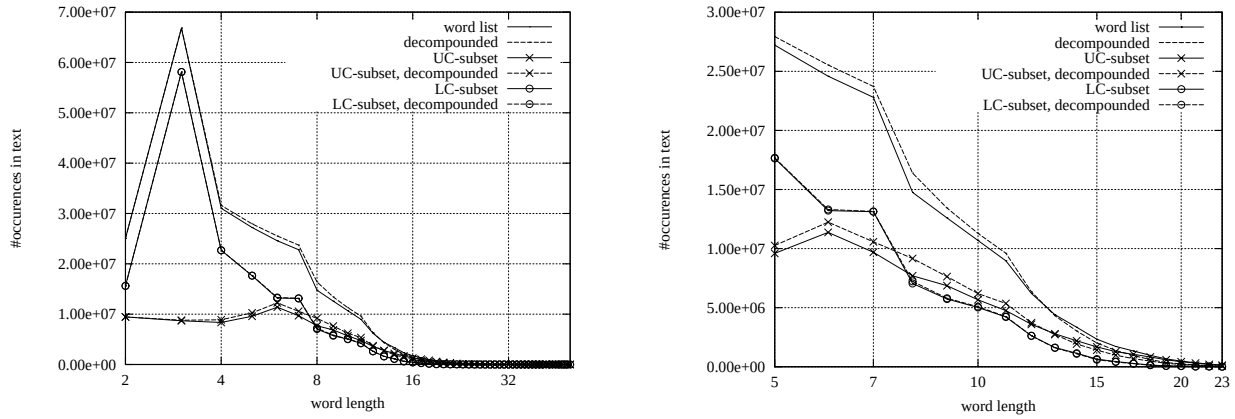

Figure 4. **Left**: Total number of word occurrences against word length (computed using 300M words). Different curves are drawn for the sets of lower-case words (LC), upper-case words (UC) and the full word set. For each set the total number is computed before and after decomposition. **Right**: A zoom is carried out on the region comprising words in the range of 5 to 23 characters.

Table 7 shows relative reduction rates obtained after decomposition on the #distinct, #total and %OOV measures. Separate results are provided for capital initial words and for lower case words. As expected, the best reduction in lexical variation is obtained for upper-case initial words (12.1%) with a relative OOV reduction of 17.7%. The impact of case-insensitivity is also measured.

Table 7. Relative reduction (%) of lexical variation between the original word list and the decompounded list

	#distinct	**#distinct (#occ. > 5)**	**#total**	**#total (#occ. > 5)**	**%OOV 65k**
Upper-case	12.1	10.0	-4.0	-4.4	17.7
Lower-case	2.8	1.4	-0.2	-0.2	3.5
Total	10.2	7.9	-1.3	-1.4	11.4
Tot. (case-ins.)	10.7	8.4	-1.3	-1.4	12.3

Figure 5 displays the number of distinct items against a minimum occurrence threshold. This figure shows that decomposition reduces the number of items with a low occurrence threshold. This means that decomposition reduces the number of rarely occurring items (the decomposition does not create additional rare items). The right part of Figure 5 shows coverage rates against the minimum occurrence threshold. The curve corresponding to the decompounded text stays above the curve obtained from the original text, the gap between the two curves increasing with the threshold. This means that the decomposition produces known items, for which the number of occurrences in the text corpus is increased.

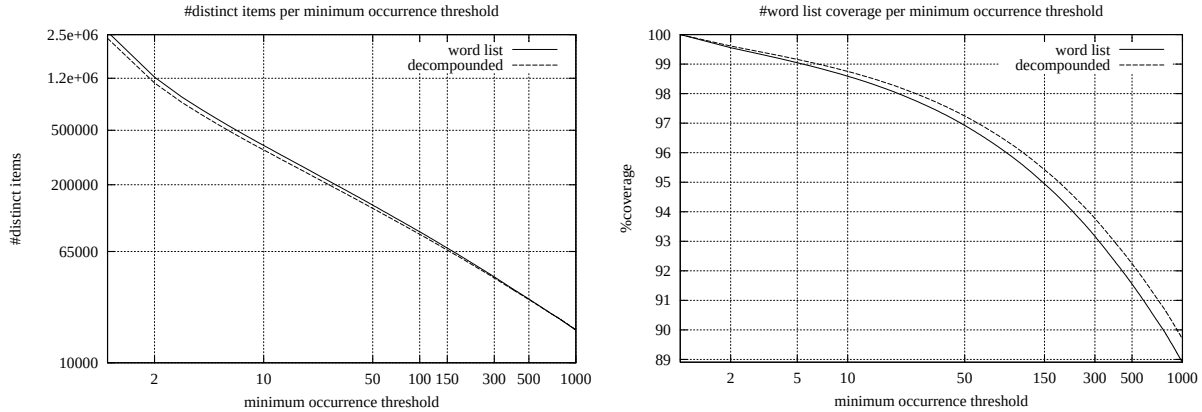

Figure 5. **Left**: The curves indicate the *number of distinct items* with a number of occurrences in the text corpora higher than the *minimum occurrence threshold*. The two curves correspond to the original texts and the decompounded texts. **Right**: For the original and the decompounded text the curves indicate the *coverage* as a function of the *minimum occurrence threshold*.

Decomposition information can significantly improve automatic letter-to-sound conversion for ASR pronunciation lexica. This assertion is based on our experience in manually verifying the automatically generated phonemic transcriptions. A large proportion of problems could be linked to unknown morpheme boundaries. Examples of problematic conversions are : Schuld|enerb|last, Berlin|erb|laus, Schwäche|tendenz|en, Konkurr|entente|lefone, Entwicklungs|ten-den|zen, R|ei|nkarnation, Aktion|st|ag, Bahn|st|eig, Bakt|eriengen|... We can cite some more exotic examples of ambiguous conversions where problems of words of foreign origin coincide with the morpheme boundary problem: Beatmeister versus Beatmen, Beachten versus Beachball.

Common ambiguous consonants include /b d g s v/ which tend to be pronounced differently depending on their position in the syllable. Ambiguous sequences are -er- , -st-, -sp-, -ng-, -ge-, -ea-, -ei-,... for which morpheme boundary information is mandatory to make the right decision. This determines whether a consonant is in syllable-initial or syllable-final position, which in turn allows the letter-to-sound converter to use the right rule.

4. Conclusion

In this presentation our current investigations on German morphological decomposition for ASR are described. Whereas one of the motivations of this work is improved letter-to-sound conversion, the paper clearly focuses on decomposition for lexical variation reduction. Using 300 M words of text corpora, major sources of lexical variation are discussed in terms of word length and case-sensitivity. Compounds, especially noun compounds, are identified as the major source of lexical variation. Decomposition rules have been developed and a significant coverage improvement has been achieved. The original OOV rate of 5.2% is reduced to 4.6% with a very limited set of 335 rules. Case-insensitivity produces a further reduction, resulting in an OOV rate of 4.2%. Using this decomposition rule set prior to letter-to-sound conversion will disambiguate a large number of ambiguous character sequences. The rule development is being continued. Future work includes validation of the decomposition rules on independent text data and comparison with other existing tools, experiments with decompounded texts for language modeling and speech recognition, and evaluating the decomposition for automatic pronunciation dictionary generation.

5. References

Adda, G., Adda-Decker, M., Gauvain, J.L. & Lamel, L. (1997). Text normalization and speech recognition in French. *Proc. 5th Conf. on Speech Comm. and Techn. (Eurospeech'97)*, Rhodes.

Adda-Decker, M., Adda, G., Lamel, L. & Gauvain, J.L. (1996). Developments in large vocabulary, continuous speech recognition of German. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta.

Habert, B., Adda, G., Adda-Decker, M., Boula de Mareuil, P., Ferrari, S., Ferret, O., Illouz, G. & Paroubek, P. (1998). The need for tokenization evaluation. *Proc. 1st Int. Conf. on Language Resources and Evaluation (LREC'98)*, Granada.

Hausser, R. (ed.) (1994). *Computer-Morphologie, Dokumentation zur Ersten Morpholympics.*

Lamel, L.F., Adda-Decker, M. & Gauvain, J.L. (1995). Issues in large vocabulary, multilingual speech recognition. *Proc. 4th Conf on Speech Comm. and Techn. (Eurospeech'95)*, Madrid.

Matsuoka, T., Ohtsuki, K., Mori, T., Furui, S. & Shirai, K. (1996). Large vocabulary continuous speech recognition using a Japanese business newspaper (Nikkei). *Proc. DARPA Speech Recognition Workshop*, Harriman, 137-142.

Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.-L., Kershaw, D., Lamel,, L., Leeuwen, D., Pye, D., Robinson, A., Steeneken, H. & Woodland, P. (1997). Multilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language* **11**(1), 73-89.