Investigating text normalization and pronunciation variants for German broadcast transcription[*]

Martine Adda-Decker, Gilles Adda & Lori Lamel

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{madda,gadda,lamel}@limsi.fr

ABSTRACT

In this paper we describe our ongoing work concerning lexical modeling in the LIMSI broadcast transcription system for German. Lexical decomposition is investigated with a twofold goal: lexical coverage optimization and improved letter-to-sound conversion. A set of about 450 decompounding rules, developed using statistics from a 300M word corpus, reduces the OOV rate from 4.5% to 4.0% on a 30k development text set. Adding partial inflection stripping, the OOV rate drops to 2.9%. For letter-to-sound conversion, decompounding reduces cross-lexeme ambiguities and thus contributes to more consistent pronunciation dictionaries. Another point of interest concerns reduced pronunciation modeling. Word error rates, measured on 1.3 hours of ARTE TV broadcast, vary between 18 and 24% depending on the show and the system configuration. Our experiments indicate that using reduced pronunciations slightly decreases word error rates.

1. INTRODUCTION

The German language, more than other major western languages, exhibits a large variety of distinct lexical forms. This characteristic raises specific research issues for German speech recognition. In earlier work on German LVCSR (large vocabulary continuous speech recognition) it was proposed to use morphological decomposition to address the lexical coverage problem [4]. LVCSR systems must meet the following requirements to enable good performance: vocabulary, as well as the acoustic and language models must achieve good coverage under operating conditions. The lexicon should contain all or most words likely to appear during operation with a minimal out of vocabulary (OOV) word rate. Low lexical coverage entails high word error rates, hence the motivation for maximizing coverage.

ASR systems typically make use of full form word lexica, where each lexical entry is described by one or several pronunciations (phonemic transcriptions). For inflected languages partial morphological decomposition can help improve lexical coverage. Partial decomposition has been used to transcribe a German human-human scheduling corpus [6] containing about 120k words with a total of 6k distinct entries, and more recently for hypothesis driven lexical adaptation in a multipass decoding scheme [3]. For the Portuguese language [2] morphological decomposition has been used to improve lexical coverage and language modeling with newspaper corpora comprising 11M words and 160k distinct lexical entries. In this contribution German lexical variety is studied in very large corpora (300M words). Partial decomposition is investigated to increase lexical coverage for a fixed lexicon size. A secondary goal is to improve automatic letter to sound conversion by disambiguated cross-lexeme letter sequences. Partial decomposition is particularly well suited for transcription systems aiming at automatic archiving or retrieval applications, where stemming is routinely applied.

2. TEXT AND TRANSCRIPTS CORPORA

The written corpora used in this study come from different sources of newspaper and newswire texts and from transcripts of audio broadcasts.

• newspaper (∼ 200M words):		
Berliner Tageszeitung (TAZ):	1986-99	147 M
Die Welt:	1996-98	20 M
Frankfurter Rundschau:	1992-93	34 M
• news wire (∼ 100M words):		
Deutsche Presse Agentur (DPA):	1995-96	29 M
Agence France Presse (AFP):	1994-96	36 M
Associated Press Worldstream:	1993-96	40 M
• ARTE transcripts (∼ 200k words).		

The largest source is Berliner TAgesZeitung (**TAZ**) with almost 150 M words (years 1986-99) purchased directly from the newspaper. After a rough text preprocessing (sentence and word segmentation) a total of 300 M words (running text) produces an exhaustive word list of about 2.8 M different lexical items. In the list of 65k most frequent words (the size of a typical recognition lexicon) all items occur more than 160 times (the least frequent item included is "Wirtschaftsboom", meaning "economy boom"). There are 500k words with 5 or more occurrences, 2.3 M words occurring fewer than 5 times in the texts. The lexical coverage obtained on this 300M training text material and corresponding OOV rates using vocabularies of N most frequent words are shown in Table 1. In the following we note whether the coverage figures are computed on the 300M training corpus or on a 30k development test set, held out from ARTE transcripts.

N words	10k	30k	65k	100k	200k
OOV rate	14.3	8.3	5.2	3.9	2.4

Table 1: Lexical coverage and corresponding OOV rates on the 300M training text material using different vocabulary sizes N.

3. SOURCES OF LEXICAL VARIETY

In German, a major obstacle to high lexical coverage arises from word compounding, inflections and other derivations. For instance, a given adjective in German may be found with

[*] Part of this work is funded by the European LE4-Olive project.

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. In our 65k dictionary 10 distinct lexical entries are obtained for **schnell**, (engl.: quick) : schnell, schnelle, schnellem, schnellen, schneller, schnelles, schnellere, schnelleren, schnellste, schnellsten. Generatively more powerful than inflection is the word compounding process. Compounding may theoretically produce an infinite number of lexical items.

The German language also has, in minor proportions, graphemic variability, e.g. -ß- or -ss- writing, declension (e.g. genitive -s or -es : Ausstand-s or Ausstand-es). A non negligible part of the observed variability can also be linked to the text sources: word-internal capital letters are particularly frequent in the Berliner TAZ (e.g.SportlerInnen). The much debated German orthographic reform, which has been elaborated these last decades and is official since around 1996, will certainly increase graphemic variability in the coming years. Other less linguistic reasons stem from orthographical errors and insufficient text normalization. From all cited sources of lexical variety, compounding of nouns seems to be the most productive. A good indicator for compounds is word length. In Figure 1 text corpora in German, French and English are compared in terms of number of distinct lexical entries per word length (in number of characters). Figure 1 shows a significant lexical variety increase for German word lengths beyond 8 characters, where English and French curves drop severely. This significant lexical variety increase can be explained by word compounding. In addition to inflection, compounding is a very productive process for German.

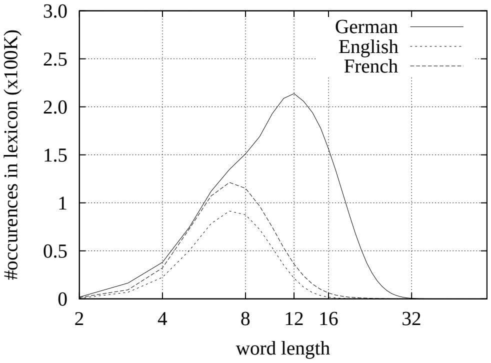

Figure 1: Number of distinct lexical entries extracted from 300M word corpora per language (German, French, English) against word length.

Whereas word compounding produces a huge number of additional entries, their relative occurrence in text corpora remains low (see Figure 2). Their contribution to OOV rates however is most significant: German OOV words have an average length of about 11 characters. Figure 3, focusing on German, with separate curves for words starting with a capital letter (UC curve) and lower-case words (LC curve), shows that lexical variety is largest for upper-case words (mainly nouns). There are more than 2 M of these items (80%) in the exhaustive word list.

4. COMPOUNDS & INFLECTIONS

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 variety in text corpora. A beneficial side effect of lexical variety reduction is the decrease of the sparse data problem for language model estimation. Concerning decomposition our interests were both OOV reduction and the production of accurate lexeme boundaries to improve automatic letter to sound conversion on ambiguous cross-lexeme letter sequences (e.g. nst in Aktions|tag, Bahn|steig). Based on the lexical variety analysis, upper-case initial words and long words were selected for decomposition investigation. Inflectional derivation applies to most lexemes, independent of case and word length.

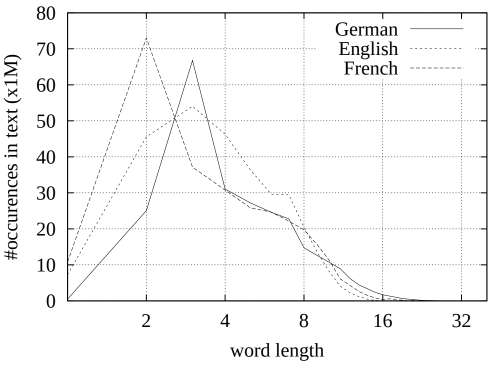

Figure 2: Number of observed words occurring in 300M word corpora per language (German, French, English) as a function of word length.

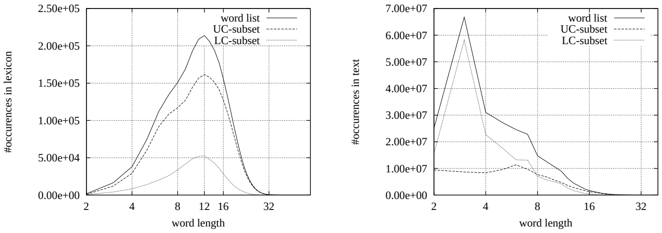

Figure 3: Left: Number of distinct lexical entries as occurring in the German corpora against word length. Separate curves are given for upper-case (UC) and lower-case (LC) word starts. **Right:** Number of observed words occurring in the text corpora.

Methods

For decomposition three approaches are being explored. Two approaches aim at developing **lexeme-based** decomposition rules, using word counts to select the potentially most promising rules. A third set of rules gathers more **general** decompositions. Finally a partial stripping of inflections is applied.

- **A-set rules**

All items starting with a capital A in the 65k word list have been checked, and decomposition rules have been manually developed. For a total of about 3k entries a set of 260 rules has been elaborated. Example rules are given in Table 2. In general the matching string is limited to one lexeme, 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 short lexemes (typically one syllable) are not decompounded (e.g. Alt in Altpapier).

Whereas only 3k such words are in the 65k lexicon, 137k distinct items are in the complete text corpora. This number is drasti-

matching string	/	exceptions
Abend	/	s|es|e|en
Altpapier	/	s|es|e|en
Arbeits minister	/	s|ium|iums|ien|in|n ˜_
Ausreise	/	n|ns|r|rn|rin|rinnen|nde|nden|nder

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 be empty). .

cally reduced (78k) by applying only 260 different decomposition rules. This partial result encourages us to continue the development of decomposition rules for the complete 65k list.

- **Most frequent word starts**

A similar approach has been explored to determine decomposition rules based on the most frequent word starts of a given length. The word starts must have at least 8 characters and are checked manually to develop appropriate rules. A set of 180 rules has been elaborated. Examples are shown in Table 3. Most rules have no exceptions and concern upper-case word starts, even if some rules occur for lower-case words (e.g. `zusammen`). This approach is very effective both for speeding up the rule elaboration process and for improving coverage.

matching string		exceptions
Wirtschafts	/	
zusammen	/	
Verkehrs	/	
Friedens	/	
Computer	/	n|s

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

- **General rules**

A limited set of general rules were also identified. A morpheme boundary can be hypothesized after the occurrence of letter sequences such as `-ungs`, `-hafts`, `-lings`, `-ions`, `-heits` with very few exceptions. The most productive rule is `-ungs`, occurring in 130k distinct items (`Regierungschef`, `Führungstor...`). These general rules, all with an `s`-ending, clarify the `s`- pronunciation during letter-to-sound conversion.

- **Inflection stripping**

Whereas in compounding a given lexeme can be combined with an open set of other lexemes, inflectional derivation is a general mechanism where a closed set of small items (inflections) can be added to most lexemes, thus producing many distinct lexical items. Inflection stripping should thus have a significant impact on coverage. We have experimented with stripping some of the most common German inflections: `-en`, `-es`, `-em`, `-er`, `-e`, `-s`, `-m`, `-r`. The only condition for stripping off the hypothesized inflection is based on word length: the base form (i.e. after stripping) must have at least 5 characters. An example of inflectional derivation, as observed in our corpora, is shown in Table 4.

Results

- **Coverage**

In Table 5, we compare the results of the 3 different decomposition rule sets on the whole text corpus. Combining the different rule sets (*all rules*) accumulates the individual gains in coverage. When a case-insensitive text normalization is applied an additional 0.4% absolute gain in coverage is achieved.

		#occur.
base	schmunzelnd	358
inflected	schmunzelnd-e	10
	schmunzelnd-en	8
	schmunzelnd-er	8
	schmunzelnd-em	4
	schmunzelnd-es	1

Table 4: Example of partial inflectional derivation for the adjective `schmunzelnd` (engl. `smiling`).

300M / train.	#rules	%coverage	%OOV	rel.
original	-	94.8	5.2	-
A set	260	94.9	5.1	2
8 char. most freq.	180	95.1	4.9	6
general rules	7	95.1	4.9	6
all-decomp	447	95.5	4.5	13
all-decomp (c.i.)	447	95.9	4.1	21

Table 5: Lexical coverage obtained with a 65k lexicon on the 300M corpus with the individual rule sets *A set*, *8 char. most freq.*, *general*, and the combination of the 3 rule sets (*all*). The last column gives relative OOV reduction as compared to the *original* configuration. Case-insensitive text processing results are added (c.i.).

Figure 4 shows that the number of distinct items reduces significantly for word lengths in the range of 11 to 20 characters. As expected the curves corresponding to words starting with capital letters display the most significant differences.

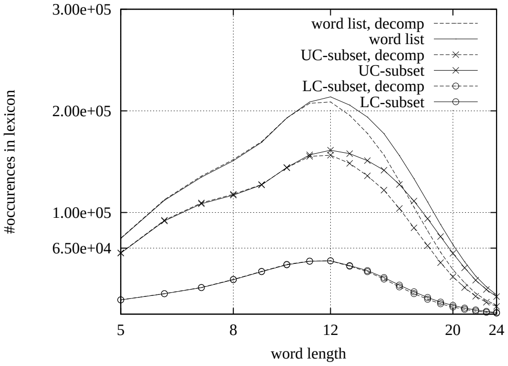

Figure 4: **Left:** Number of distinct words (as occurring in the exhaustive word list) against word length. A zoom is carried out on the region comprising words in the range of 5 to 24 characters.

Table 6 shows that decomposition rules are equally effective on independent development data. Inflection stripping produces a significant OOV reduction. Such syntactic derivation normalization is relatively straightforward to implement.

Figure 5 shows coverage versus minimum occurrence threshold. The curves corresponding to the processed text stay above the curve from the original text, the gap between the two curves increasing with the threshold. This means that decompounding and inflection stripping tend to produce already seen items for which the number of occurrences in the corpus is increased.

- **Language model perplexity**

Table 7 gives 4-gram language model (LM) perplexities (ppx) on the 30k development set from the transcripts. Perplexities are normalized to take into account changes in corpus size due to decompounding [5]. OOV words are discarded for ppx com-

30k / dev.	65k**n**	65k**n+t**	rel. (**n**/**n+t**)
original	5.2	4.5	-
all-decomp	4.7	4.0	10 / 11
all-decomp + inflect.	3.4	2.9	35 / 36

Table 6: OOV rates obtained with 65k lexica on 3 forms of the 30k development corpus *original*, *all-decomp*, *all-decomp + inflect.*. 65k **n** is obtained from the news texts without considering transcripts. The 65k **n+t** includes all words from the 170k ARTE training transcripts. Relative OOV reductions are included for both **n**, **n+t** lexica.

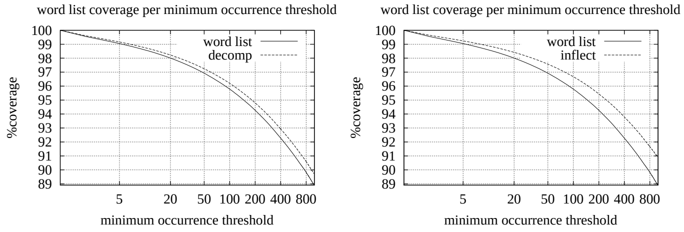

Figure 5: The curves indicate the coverage obtained by the set of items with more than **minimum occurrence threshold** occurrences in the text corpora. The two curves correspond to the original and processed texts (left: compounds, righ: inflections).

putation. The ppx values increase for *decomp-all* and *decomp-all+inflect* forms: as more and more infrequent words (OOVs in the *original* text) are included for the subsequent text versions, perplexities naturally tend to increase. The bias due to OOV reduction can be alleviated by simply replacing OOV words by the least frequent unigram in the language model. This allows an estimate of a realistic lower bound on the ppx during recognition (ppx_oov).

text form	ppx	#OOV	ppx_oov
original	213	1392	310
decomp-all	227	1259	310
decomp-all+inflect	235	915	280

Table 7: 4-gram LM perplexities on the 30k ARTE transcripts.

5. PRONUNCIATION VARIANTS

Our 65k pronunciation lexicon has been developed at LIMSI using a letter-to-sound conversion system, a PERL script of roughly 300 rules: general rules for German and a list of exceptions for the most common foreign words. Generated pronunciations are semi-automatically checked for errors due to ambiguous letter sequences (e.g.: `best, eng...`) or proper names.

Two pronunciation variant dictionaries have been derived from the *original* dictionary using the following morpheme reduction rules: /schwa-vowel+[lnm]/ are replaced by syllabic /[lnm]/ if they occur in word final position or if followed by a consonant. Mapping sequences may be either simply replaced (*reduced* dictionary) or added to allow for both full or reduced pronunciations (*optional* dictionary). For each dictionary, distinct acoustic models have been trained and used during recognition.

Experiments were carried out using 4 shows of news and documentaries each longer than 15 minutes (total of 1h20min). Prior to word recognition the continuous audio stream is partitioned into homogeneous acoustic segments using an iterative segmentation and clustering algorithm, and non-speech segments are identified and rejected [1]. The acoustic models are context-dependent, position-independent triphone-based cross-word phone models. Each phone model is a tied state left-to-right CD-HMM with Gaussian mixtures, where the state tying is obtained by means of a decision tree. The acoustic models depend on the pronunciation lexica used, but the number of parameters are comparable across the different model sets. Unsupervised cluster-adaption is carried out between decoding passes. The recognition LM is a 3-gram interpolation of a text LM and a transcript LM.

pronunciation	*orig.*	*reduced*	*optional*
2 news shows	21.6	21.3	21.8
2 doc. shows	26.0	26.0	26.6
all shows	23.9	23.7	24.2

Table 8: Word error rates on ARTE news and documentary shows using 3 different pron. dictionaries.

Table 8 shows recognition results. Only small differences in recognition rates are measured for the different experimental setups, with slightly better results for the reduced lexica. Recent experiments with a new decoder, a 4-gram LM, position-dependent acoustic models and additional variants, result in word error rates of 18% on the news shows and 24% on the documentaries.

6. CONCLUSIONS & PERSPECTIVES

We have investigated several ways to improve lexical coverage, and more generally lexical modeling in our German broadcast news transcription system. Concerning lexical coverage a 10% relative OOV reduction has been achieved using 450 decompounding rules and additional 25% relative OOV reduction using partial inflection stripping. For the different text forms language models have been estimated and an OOV sensitive perplexity measure shows interesting gains when using inflection stripping. Other investigations concern the pronunciation dictionaries and acoustic modeling. Slightly better recognition results have been obtained with reduced pronunciation dictionaries. Normalized text forms allow a 65k base form lexicon to represent a 170k full form lexicon. The extension of our pronunciation dictionaries is underway for transcription experiments using this new lexicon.

REFERENCES

[1] Gauvain J.L., Lamel L., Adda G., Jardino M. (1999), "*The LIMSI 1998 HUB-4E Transcription System*", DARPA Broadcast News Workshop, pp. 99-104, Herndon, VA.

[2] Martins C. et al. (1999), "*Using Partial Morphological Analysis in LM Estimation for Large Vocabulary Portuguese Speech Recognition*", Eurospeech, Budapest.

[3] Geutner P., Finke M., Waibel A. (1999), "*Selection Criteria for Hypothesis Driven Lexical Adaptation*", pp. 617-620, vol.II, ICASSP, Phoenix.

[4] Adda-Decker M., Adda G., Lamel L., Gauvain J.L. (1996), "*Developments in Large Vocabulary, Continuous Speech Recognition of German*", IEEE-ICASSP, Atlanta.

[5] Adda G., Adda-Decker M., Gauvain J.L., Lamel L. (1997), "*Text normalization and speech recognition in French*", ESCA Eurospeech, Rhodes.

[6] Geutner P. (1995), "*Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems*", IEEE-ICASSP, Detroit.