On the use of speech and text corpora for speech recognition in French

Martine Adda-Decker
Sorbonne Nouvelle University
286 PUBLICATIONS **3,428** CITATIONS

Jean-Luc Gauvain
Computer Science Laboratory for Mechanics and Engineering Sciences
316 PUBLICATIONS **12,982** CITATIONS

Lori Lamel
French National Centre for Scientific Research
423 PUBLICATIONS **14,674** CITATIONS

On the use of speech and text corpora for speech recognition in French

Martine Adda-Decker, Gilles Adda, Jean-Luc Gauvain, Lori Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{madda,gadda,gauvain,lamel}@limsi.fr
http://www.limsi.fr/TLP

Abstract

In this contribution we investigate the importance of different size text and speech corpora for speech recognition in French, measuring the impact of the training data size on the recognition results and the effect of different text normalization choices. Compared to English, the French language has higher lexical variety, which in turn implies lower lexical coverage for fixed size lexica, and poorer language modeling for fixed size training text corpora. Increasing the size of training text corpus is shown to be more effective for error reduction than adding acoustic data. Experimental results indicate that a significant increase in training texts for language modeling should be accompanied by an increase in vocabulary size, at least as long as lexical coverage remains a problem. The impact of text normalization on recognition results is demonstrated by applying different types of normalizations to the reference and hypothesis strings. An analysis of the word error rate is given as a function of word frequency rank. The contribution of the language model is shown to be of particular importance to discriminate homophones in French which cannot be seperated on an acoustic basis. Additional training text data should still allow improvement of language model accuracy and hence yield better recognition results.

1. Introduction

It is generally admitted that increasing speech and text corpora for training, results in more accurate acoustic and language models entailing reductions in the recognition error rate. In this contribution we measure the impact of different size acoustic training corpora and different size text training corpora on our French large vocabulary continuous speech recognizer.

French is known to be an inflected language with a relatively high lexical variety. When developing speech recognition systems this property inhibits the achievement of high lexical coverage for a fixed lexicon size and accurate language modeling given limited size training corpora. The lexical variety as observed in different text sources (mainly newspapers) can be partly reduced by appropriate text normalization (Adda 1997), but the need of additional text corpora for French language model (LM) training has been clearly noted (Adda-Decker 1997). A large proportion of observed lexical variety corresponds to homophones in speech, which can be seperated only by an appropriate language model. Concerning homophones, a comparative study of French and English showed that, given a perfect phonemic transcription, about 20% of words in English newspapers are ambiguous, whereas 75% of the words in French newspaper texts have an ambiguous phonemic transcription (Gauvain 1994). Concerning lexical coverage, the number of words in French must typically be doubled in order to obtain the same word coverage as in English for comparable newspaper text conditions. This difference between French and English mainly stems from the number and gender agreement in French for nouns, adjectives and past participles, and the high number of different verbal forms for a given verb (about 40 forms in French as opposed to at most 5 in English).

When increasing the system's lexicon from 20k to 65k words, additional text corpora are required to estimate LM parameters. Observed gains in recognition performance are then due to both improved lexical coverage and language modeling. The use of different amounts of acoustic training material is discussed. Recognition results are presented and compared on 20k and 65k systems using test sets with and without out of vocabulary word (OOV) control.

The impact of text normalization on recognition results is demonstrated by applying different types of normalizations on the reference and hypothesis strings. The results show the link between a proper tokenization of the text material and recognition results. Observed word error rates are related to word frequency ranks in order to highlight the LM contribution during the recognition decision. These results underline the need for better language models.

In the next Section we provide a short description of the general framework in which our latest French large vocabulary continuous speech recognition system has been evaluated (AUPELF project) and in which a major part of the below described developments have been carried out. In Section 3 we present the different French speech corpora used and corresponding recognition results. Section 4 provides a description of text corpora, language models and related recognition experiments. In Section 5 the problem of text normalization or tokenization is addressed with respect to coverage and recognition result scoring. In Section 6 the main error sources for French speech recognition are described and related to either acoustic or language modeling problems.

2. The Francophone AUPELF Project

A speech recognition evaluation project for French recognizers has been launched in 1996 by the Francophone AUPELF-UREF organisation. Academic sites with French recognition systems could participate in various evaluation categories on read speech from *LeMonde* newspaper. The different categories mainly differed by the allowed lexicon size (20k,65k), and by the use or not of an OOV-controlled test set. Previous experiments in large vocabulary speech recognition in French have been reported in Lamel (1995) and in Young (1997) using a 20k vocabulary (LRE-SQALE project) on test sets with a controlled OOV rate of about 2%. Without artificial limitation the OOV rate tends to be closer to 5 or 6% with 20k systems. For the AUPELF'97 evaluation (Dolmazon 1997) development and evaluation test sets $\mathcal{T}$ of 600 sentences have been selected without prior control of the OOV rate. From these, $\mathcal{T}'$ subsets (containing about 300 sentences) of paragraphs minimizing OOV rates have been selected. All results reported below are obtained on the development test sets. Comparisons between $\mathcal{T}$ and $\mathcal{T}'$ are carried out. More extensive information concerning our AUPELF system and the results obtained can be found in Adda-Decker (1998), and in Adda (1997a,b,c).

3. Speech corpora

We briefly summarize the acoustic modeling approach of our system, before describing the investigated speech corpora and the results obtained.

Acoustic modeling: The recognizer makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling. The acoustic models are sets of context-dependent (CD), position-independent phone models, which include both intra-word and cross-word contexts, selected automatically based on their frequencies in the training data. Each phone model is a 3-state left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). The acoustic parameters consist of 39 cepstral parameters (including first and second order derivatives) derived from a Mel spectrum estimated on a 8kHz bandwidth.

The acoustic models are built in a series of steps. A first set of models is used to segment and label the training data using Viterbi alignment of the text transcription and a lexicon containing one or more pronunciations per word. The chosen phone sequence and segmentation are then used to construct a set of context-independent models, with a maximum of 32 Gaussians per state. Larger context-dependent model sets can then be built in a similar way, using new segmentations obtained with a previous set of acoustic models. As contexts are selected based on their frequencies in the training data, additional training data should result in improved acoustic modeling accuracy.

Corpora description: The corpora contain read newspaper texts from *LeMonde*, selected to cover a high range of phonemic contexts (Lamel 1991). Three corpora have been used in the recognition experiments reported here:

Bref80: 5.3k sentences from 80 speakers, as used in the SQALE experiments.

Bref: 66.6k sentences from 120 speakers (Bref80 $\subset$ Bref).

Bref+Bref2: 85.9k sentences from 420 speakers (Bref$\subset$ Bref+Bref2).

Bref contains a relatively small number of speakers uttering each a large number of sentences (close to 500), whereas Bref2 contains a large number of speakers with about 60 sentences each.

Experimental results: Using Bref80, Bref and Bref+Bref2 corpora about 1800, 5500, 6100 tied-states, gender-dependent triphone models have been estimated respectively. The impact of acoustic training sizes on recognition results is illustrated in Table 1. These results were obtained with a 65k system as described in (Adda 1997b) on the AUPELF'96 development test set (600 sentences from 20 speakers). An important gain is observed from Bref80 to Bref, which can be directly related to the increased number of CD models. Whereas the number of acoustic models is larger for Bref+Bref2 than for Bref, no significant difference in recognition results was measured.

	Bref80	*Bref*	*Bref+Bref2*
#sentences (training)	5.3k	66.6k	85.9k
#CD models	1800	5500	6100
$\mathcal{T}$-word error rate	15.0%	12.9%	12.9%
$\mathcal{T}'$-word error rate	10.8%	8.8%	8.8%

Table 1: Sizes of the different speech corpora for training, the number of context-dependent acoustic HMM models and word error rates on the AUPELF'96 development set $\mathcal{T}$ (600sentences from 20 speakers) and $\mathcal{T}'$ (300 sentences subset with controlled OOV rate).

It is worthwhile to note that a significant increase in the number of training speakers (from 120 to 420 speakers) leaves recognition results roughly unchanged. Similar results could be observed on the AUPELF'97 evaluation set. These experiments suggest that the adopted acoustic model training approach has reached a limit where larger corpora no longer yield better recognition results.

4. Text corpora

The higher lexical variety in French as compared to English entails lower lexical coverage for a given size lexicon (N words) and poorer language modeling as long as the training corpus size remains limited. For statistical word-based language models the needed amount of training material naturally depends on the system's vocabulary size. Better recognition results are achievable only if a vocabulary increase is carried out jointly with a significant increase in text corpora for LM training. Conversely it may be important to increase the lexicon size when enough LM training material is available, as long as lexical coverage remains a problem.

Lexical coverage: Table 2 gives OOV rates for different values of N (ranging from 20k to 65k lexical items) measured on the AUPELF'96 development test set. The N most frequent words have been obtained from a training data set (T_0 *LeMonde*,years 1987-88 (40M words)).

Corpora description: Training texts have been added to the data used in the SQALE evaluation in 1995 where only 40M words from *LeMonde* (years 1992-93) were available. Training corpora used for the AUPELF evaluation included

word list	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
$\mathcal{T}$	6.4	4.3	3.2	2.4	2.0	1.8

Table 2: OOV rates on the dev $\mathcal{T}$ set, for word lists ranging from 20k to 65k words. The word lists consist of the N most frequent words in T_0 training data

over 250M words. These texts come from similar but different sources:
Le Monde: a daily French newspaper,
Le Monde Diplomatique: a monthly political and cultural newspaper,
Agence France Presse: the main French news agency.
We describe the amounts of data as used for the AUPELF development experiments:
LeM: 185M words from *Le Monde* years 87-96[1],
MD: 6M words from *Le Monde Diplomatique*, years 89-96,
AFP: 64M words from *Agence France Presse*, years 94-96.

Language models: Statistical *n*-gram models attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. A backoff mechanism (Katz 1987) is used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. Bigram and trigram language models have been trained using different combinations of the above described corpora. In Table 3 the LM size for fixed cutoff values are shown as a function of training corpus size.

	LeM	*LeM + MD*	*LeM+ MD + AFP*
#words	185 M	191 M	255 M
#bg	11.9 M	12.1 M	13.5 M
#tg	13.9 M	14.3 M	18.1 M
ppx.	137.7	137.3	135.2

Table 3: LM size (number of bigrams and trigrams) and perplexity (ppx.) as a function of different training corpora: *LeM*, *LeM + MD*, *LeM + MD + AFP*. Bigram and trigram cutoffs of 0 and 1 are applied respectively.

When building N-gram language models for French, we use cutoffs of 0/1, whereas in English we typically apply 1/2 cutoffs for bigram/trigram selection. Adding the AFP data yields an increase of about 10% (relative) for the number of bigrams in the LM, whereas for trigrams a 20% (relative) increase is observed. Cutoff values and the increase of LM size when adding training data suggest that still more data are necessary for accurate LM training.

Experimental results The recognition results presented hereafter are based on 20k and 65k recognition systems. For the 20k system, two different language models were trained using either the complete 255M text set (LeM+MD+AFP) or a 40M text subset (T_0). The obtained results are shown in Table 4. The gains observed when significantly increasing the training text material remains rather low: 9% (relative) on the $\mathcal{T}$ development set 600 sentences) and of 16% (relative) on the $\mathcal{T}'$ set (300 sentence subset of $\mathcal{T}$ with controlled OOV rate). A possible conclusion here is that the low lexical coverage prevents the LM from taking advantage from the larger text corpus. This hypothesis is supported by the larger gains observed for the $\mathcal{T}'$ subset where OOV rates are significantly smaller. When moving from a 20k to a 65k system OOV rates are reduced by nearly 3% (absolute) for the $\mathcal{T}$ set and almost 1% (absolute) for the $\mathcal{T}'$ subset, ranging from 1.3% to 0.5%respectively. Comparing the 20k-255M and 65k-255M systems the relative gain is about 40% for both $\mathcal{T}$ and $\mathcal{T}'$ sets, consisting in roughly 9% absolute error reduction for the $\mathcal{T}$ set and almost 6% for the $\mathcal{T}'$ subset. The important gain is due to combined improvements in lexical coverage and language modeling: as the language model is based on 65k different lexical items, better advantage can be taken from the training corpus (255M words). These results illustrate the importance of increasing the system's vocabulary size provided there are enough data for LM training available.

	LM	*OOV*	*Werr*
$\mathcal{T}$	20k-40M	6.4%	23.9%
	20k-255M	6.4%	21.8%
	65k-255M	1.3%	12.9%
	LM	*OOV*	*Werr*
$\mathcal{T}'$	20k-40M	3.6%	17.3%
	20k-255M	3.6%	14.6%
	65k-255M	0.5%	8.8%

Table 4: Recognition results obtained by 20k and 65k systems. 20k-40M and 20k-255M systems use LMs estimated from 40M T_0 data or from the 255M *LeM + MD + AFP* data respectively. $\mathcal{T}$ is the 600 sentences AUPELF development set, $\mathcal{T}'$ corresponds to a subset of 300 sentences with controlled OOV rate.

Looking at recognition results one can observe that many errors are due to short term gender and number disagreements (example: *elle ét<u>aient</u>* (she was) instead of *elle ét<u>ait</u>*). Whereas long term agreement (example: *la femme de trente ans habitu<u>é</u> aux* ... (the thirty year old woman, used to ...) instead of *habitu<u>ée</u>*) cannot be handled by N-gram language modeling, short term errors should not occur with properly trained N-gram models from sufficient data. An algorithmic solution to this problem has been investigated by interpolating the trigram backoff LM with a biclass LM (Jardino 96). This allows for an improved LM contribution when the trigram LM has to back-off to unigrams. The applied interpolation formula is as follows:

$$P_{int}(w) = \lambda P_{tg}(w) + (1 - \lambda) P_{bc}(w)$$

$P_{int}(w)$, $P_{tg}(w)$ and $P_{bc}(w)$ stand for probabilities on word sequence w from interpolated, trigram and biclass LMs respectively. An experimentally optimized value of $\lambda = 0.9$ has been fixed. Biclass-based LM interpolation allowed a perplexity reduction from 135 to 131 and a relative word

[1] *LeM* corresponds to the T_1 corpus for lexical coverage presented in section 5, + 4 months of 1996

error reduction of 1.5% due to a reduction of some short term disagreements.

Results show that for highly inflected languages like French coverage improvements together with a significant increase in training data represent the main reasons in error reduction. Further improvements can be expected by additional data and larger vocabularies, but algorithmic solutions taking better benefit from a fixed sized training corpus are interesting research alternatives.

5. Tokenization

The issue of tokenization evaluation in the natural language processing domain is addressed more extensively in (Habert 1998), in these proceedings. The impact of tokenization, what we usually refer to as text normalization, on lexical coverage and language modeling has been extensively described in (Adda 1997a, Adda 1997c). We briefly recall here the importance of training text corpus selection and normalization to optimize lexical coverage before discussing the impact of normalization on recognition results.

To measure lexical coverage as a function of training text corpus the *LeMonde* newspaper corpus has been divided in different subsets (differing in size and epoch):

T_0 : years 1987-88 (40M words)[2]
T_0' : years 1994-95 (40M words)
T_1 : years 1987-95 (185M words)
T_2 : years 1991-95 (105M words)[3]

In Figure 1 out of vocabulary (OOV) rates are given for 65k word lists derived from these different subsets and for different text normalizations. For T_0', T_1 and T_2 subsets almost identical OOV rates are obtained, showing that corpus size is not critical. To optimize coverage, text epoch is more important than text size: comparing T_0 and T_0' OOV rates a significant reduction (about 25% relative) can be observed when replacing 40M words from years 1987-88 by the same amount from years 1994-95.

The importance of proper tokenization for lexical coverage is also demonstrated in Figure 1. OOV word rates are shown to be reduced by about 50% when going from raw but clean data (N_a text form) to stronger normalized data (N_b, N_c). The N_b normalized text form derives from N_a after processing of ambiguous punctuations, capitalized sentence starts, digits and acronyms. The N_c form differs from the N_b form by additional case-insensitivity, absence of diacritics and systematic decomposition on ambiguous ponctuation marks. The N_b text form has been used for all previously presented recognition experiments.

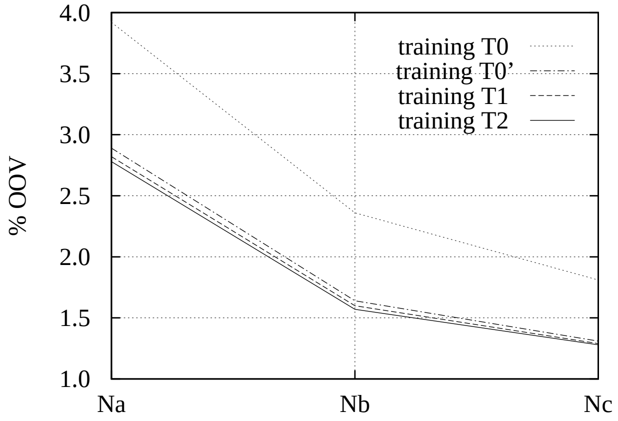

Figure 1: OOV rates on development test data for different normalisation versions N_a, N_b, N_c on T_0, T_0', T_1, T_2 training data using 65k word lists.

In the following we discuss the dependence of speech recognition results on text normalization. The speech recognition evaluation community indirectly faced the problem of tokenization for years during a so-called adjudication phase, where multiple graphemic forms of words and word sequences are discussed and a decision was taken to accept or reject the alternate forms, decided acceptable or rejected, resulting in a list of acceptable rewriting rules. Such redefinitions of lexical items may significantly influence word error rates. In Table 5 recognition results using different text normalizations are shown: reference and hypothesis strings of a given test condition are normalized as follows: starting from a standard scoring without rewriting rules (N_b result) strings are then normalized by adding case-insensitivity, by removing diacritics and by decompounding. Decompounding is seen to be the most effective normalization with respect to error rate reduction (3.5% relative), for the simple reason that the total number of words is larger and that errors are rarely found in more than one of the compounding items. The absolute number of errors remains globally constant, whereas the total number of lexical items increases. Case-insensitivity mainly concerns removing emphatic capitals (for example $\{$*Journal, journal*$\}$ or $\{$*Ministre, ministre*$\}$) which are rather common in French newspaper texts, but emphatic capitals are limited to a relatively small set of words. In (Adda 1997c) the emphatic capital normalisation has been shown to be negligible with respect to lexical coverage. However concerning recognition rates a relative error reduction of 2.5% is obtained by ignoring confusions between a word and its emphatic capitalized counterpart (which are homophones). Removing diacritics which has been shown to be important when optimizing coverage, is less effective here with only about 1% relative word error reduction. After removing diacritic symbols, words which are not homophones have the same orthographic form. However, these were not originally prone to mutual confusion (for example the words $\{$*accusé, accuse*$\}$ correspond to the phonemic forms $\{$*akyze, akyz*$\}$). The major reason of error reduction here is due to one of the rare homophone word pairs $\{$*à, a*$\}$ (in English $\{$*at, has*$\}$), both words being among the 20 most common words in French text corpora. The importance of inflected form substitutions is highlighted by the two last entries in Table 5. Root forms are obtained using the INTEX system (Silberztein 1995). More than 20% relative error reduction is achieved by reducing inflected forms to their root form.

[2] These were baseline resources for all partners in the AUPELF French recognizer evaluation project.

[3] T_2 is significantly smaller than T_1, but contains on average more recent data.

normalization	%Werr
N_b form (standard)	13.62%
N_b + case-insensitive	13.3%
N_b + ci, no diacritics	13.2%
N_b + no compounds	13.1%
N_b + no comp., ci	12.8%
N_c (N_b + no comp., ci, no diac.)	12.7%
N_b + root forms	10.3%
N_c + root forms	9.6%

Table 5: Word error rates as a function of different text normalizations applied on reference and hypotheses strings as produced by the recognizer (N_b form). The two last entries of the table result from reducing inflected forms to root forms.

6. Error Analysis

A major part of observed errors can be attributed to weak language modeling. This assertion is first supported by manual investigations of the recognizer's output. It can also be concluded from the observed error rate reductions obtained by applying different text normalizations and finally from an automatic analysis of word error rates against word frequency ranks.

Looking at recognition errors, gender, number and tense disagreements and other homophone substitutions are frequently observed. About 40% of confusion errors are due to single word homophones (for the most part these are homophone gender and number agreement forms and homophone verb forms), where the LM contribution is solely responsible. About 15% of the substitutions are due to proper names, which are difficult to model both on LM and acoustic levels, as in general they are infrequent in training texts, and foreign proper names often have a large variety of acceptable pronunciations.

Word error rates are usually obtained by averaging error measures on a sentence by sentence basis. This allows sentence error rates to be related to LM perplexity. Sentences with high error rates generally have high perplexity values. To more precisely investigate how word error rates are related to LM accuracy, word error rates can be measured on a word frequency basis, instead of the usual sentence by sentence basis. To do so, the system vocabulary is first partitioned into I word frequency rank regions $]K_{i-1}, K_i]$, which are logarithmically distributed along the word frequency rank axis. Each word w_n of the test set is associated its frequency rank k_n in the system's vocabulary. If $k_n \in]K_{i-1}, K_i]$ then w_n belongs to the i-th frequency rank region (FFR) ($1 \leq i \leq I$). The first FFR contains the 10 most frequent words (in training data): $de, la, l', le,$ à, et, les, des, d', un which are inflected forms of defined and undefined articles, the conjunction *and*, and prepositions *of* and *at*. OOV words are grouped in an OOV subset. Error rates can then be measured for each subset. In Figure 2 we analyze the word error rate as a function of $]K_{i-1}, K_i]$ word frequency rank regions (FFRs). The word occurrence distribution of the test is provided in the same figure as a complementary information. The OOV subset (0.45% of the data) with a 100% error rate is not represented. For each curve the dots correspond to the upper bound of an FFR. Figure 2 illustrates that for ranks $k_n > 5000$ error rates tend to increase drastically, but only 15% of the test are concerned, i.e. not covered by the first 5000 words (the first 7 rank regions). The first FFRs contain very short words (including many monophone homophones) which are acoustically very difficult to identify. The best results are obtained for words in the 5th FFR (rank between 500 and 1200). Here words are well trained and in general polysyllabic, which are acoustically easier to discriminate.

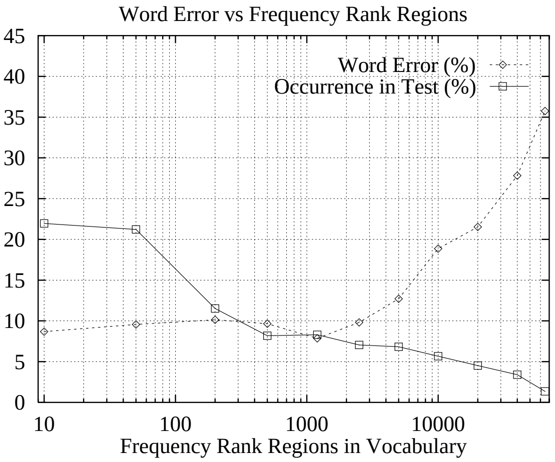

Figure 2: Word error rates and word occurrence rates as a function of frequency rank regions (FRRs) in the 65k system vocabulary. Each point defines the upper limit of an FRR. 11 FFRs have been defined and distributed logarithmically from 1 to 65000.

Concerning acoustics only a small part of errors can be clearly related to acoustic modeling reasons, like missing schwas and liaisons in the pronunciation lexicon, syllabic reduction phenomena, respirations and other noises. Progress in acoustic modeling is nonetheless important, in order to experiment with different weightings between acoustic and LM scores in the decoder.

7. Discussion and Perspectives

Increasing acoustic training data from about 6k sentences to 65k sentences allows for significant reductionsin the word error rate. A larger speech corpus did not further improve recognition rates, indicating that the acoustic modeling approach being used has probably reached its limits. Research in defining new relevant acoustic unit contexts may lead to additional benefit from larger acoustic training data. The increase of the text training corpora from 40M to over 250M words allowed a significant error rate reduction. When extending the 20k system to a 65k system recognition results are improved by 40% (relative) when moving from a 20k word system to a 65k word system. This may be explained by simultaneous improvements in lexical coverage and language modeling. We have shown the importance of increasing the lexicon size if LM training material is available, at least as long as lexical coverage remains a problem.

Small gains are achieved by trigram-biclass LM interpolation avoiding some erroneous short-term number and gender agreements. Taking into account morphological information as proposed by El-Bèze (1990), can be an interesting alternative to achieve better language model predictability and to introduce linguistic knowledge into the statistical models for highly inflected languages.

Recognition errors are mainly due to homophones, mostly arising from gender and number disagreements. The impact of text normalization (tokenization) on recognition results has been discussed. The importance of inflected form substitutions has been shown by reducing inflected forms to their root forms for both the reference and the hypothesis strings. More than 20% relative error reduction is achieved by such a reduction.

Error rates have been shown to increase drastically for less frequent words as these words are less well represented by both the acoustic model and the LM. Improving present LM techniques can be considered as a challenging research direction for French speech recognition during the next years. New application-related text sources will certainly continue to contribute to improve recognition results in the future.

8. Acknowledgment

Part of this work has been carried out within the ARC[4] *Linguistics, Computer Sciences, and Spoken Corpora* supported by the AUPELF-UREF[5].

9. References

Adda-Decker M.; Adda G.; Gauvain J-L; Lamel L. (1998), “Elements pour la mise au point de système de reconnaissance grand vocabulaire en français”, XXIIèmes JEP, Martigny, June 1998.

Adda G.; Adda-Decker M.; Gauvain J-L; Lamel L. (1997a) “Text Normalization and Speech Recognition in French”, *EuroSpeech’97*, Rhodos, Sept. 1997.

Adda G.; Adda-Decker M.; Gauvain J-L; Lamel L. (1997b), “Le système de dictée du LIMSI pour l’évaluation AUPELF’97”, *1ères JST FRANCIL*, Avignon, April 1997.

Adda G.; Adda-Decker M.; (1997c), “Normalisation de textes en français: une étude quantitative pour la reconnaissance de la parole”, *1ères JST FRANCIL*, Avignon, April 1997.

Boula de Mareüil P. (1996) “Pour une approche par règles en transcription graphème-phonème”, Séminaire GDR-PRC CHM Lexique et Communication Parlée, (pp. 203-209). Toulouse’96 France.

Dolmazon J-M et al. (1997), “ARC B1 - Organisation de la 1e campagne AUPELF pour l’évaluation des systèmes de dictée vocale”, *1ères JST FRANCIL*, Avignon, April 1997.

El-Bèze M. (1990), “Choix d’unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole”, PhD thesis, Paris VII, November 8th 1990.

Gauvain J-L; L. Lamel L.; Adda G.; M. Adda-Decker M. (1994), “Speaker-independent continuous speech dictation,” Speech Communication **15**, pp. 21-37, Sept. 1994.

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”, LREC’98 Conference, Granada.

Jardino M. (1996), “Multilingual stochastic n-gram class language models,” *IEEE ICASSP-96*, Atlanta, 1996.

Katz S.M. (1987), “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,” *IEEE Trans. ASSP*, **35**(3), 1987.

Lamel L.; Adda-Decker M.; Gauvain J-L (1995), “Issues in Large Vocabulary, Multilingual Speech Recognition,” *Eurospeech’95*, Madrid, Sept. 1995.

Lamel L.; Gauvain J-L; Eskénazi M. (1991) “BREF, a Large Vocabulary Spoken Corpus for French,” *EuroSpeech’91*, Genoa, Sept. 1991.

Silberztein M. (1995), *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, 1993.

Young S.J et al. (1997), “Multilingual large vocabulary speech recognition: the European SQALE project,” *Computer Speech & Language*, **11**(1), pp. 73-89, Jan. 1997.

[4] Actions de Recherche Concertées, Coordinated Research Actions

[5] AUPELF: Association des Universités Partiellement ou EntiŁrement de Langue Française - UREF: Université des Réseaux d’Expression Française. The AUPELF-UREF is partially sponsored by the French government.