INTERSPEECH 2009 BRIGHTON

# Constraint selection for topic-based MDI adaptation of language models

*Gwénolé Lecorvé*[1], *Guillaume Gravier*[2], *Pascale Sébillot*[1]

[1]IRISA/INSA, Rennes, France
[2]IRISA/CNRS, Rennes, France
{gwenole.lecorve, guillaume.gravier, pascale.sebillot}@irisa.fr

## Abstract

This paper presents an unsupervised topic-based language model adaptation method which specializes the standard minimum information discrimination approach by identifying and combining topic-specific features. By acquiring a topic terminology from a thematically coherent corpus, language model adaptation is restrained to the sole probability re-estimation of $n$-grams ending with some topic-specific words, keeping other probabilities untouched. Experiments are carried out on a large set of spoken documents about various topics. Results show significant perplexity and recognition improvements which outperform results of classical adaptation techniques.
**Index Terms**: language model adaptation, topic terminology, minimum discrimination information, speech recognition

## 1. Introduction

In large vocabulary automatic speech recognition (ASR) systems, unsupervised statistical language model (LM) adaptation is a key problem to face discourse variations in spoken documents of different task domains, *e.g.*, shifts in style, epoch, topic, *etc*. Given a baseline LM trained on a large general-purpose corpus, LM adaptation seeks to re-estimate $n$-gram probabilities so that they better represent a specific task domain, and hopefully lead to better transcriptions. Due to the LM training process, adaptation is commonly done by refering to a domain-specific corpus from which the $n$-gram probability distribution is re-estimated [1]. However, in the field of topic LM adaptation, the corpus-based approaches mainly focus on the way to retrieve adaptation data for a given topic whereas they simply perform the $n$-gram re-estimation step through standard methods, without taking into consideration specificities of the topic adaptation task.

The work presented in this paper is in line with a completely unsupervised topic adaptation approach, detailed in [2], based on the retrieval of topic-specific corpora from the Internet, and which seeks both to avoid the use of *a priori* knowledge and to integrate Natural Language Processing (NLP) techniques. Hence, this paper does not study the adaptation data retrieval step but focuses on the LM re-estimating problem for the specific case of corpus-based topic adaptation, with the even constant will not to rely on *a priori* knowledge.

LM re-estimating techniques used in the topic adaptation field can be split into two main classes. First, $n$-gram probabilities or counts derived from a topic-specific LM or corpus can be directly interpolated with the ones of the baseline general-purpose LM [3]; this approach is not optimal since it attaches the same importance to any $n$-gram, disregarding its relevance to the topic. Second, the targeted adapted $n$-gram distribution can be seen as the solution of an optimization problem in which the baseline LM must maximize or minimize a given measure over a topic-specific corpus, leading to rescale independently the $n$-grams probabilities. Especially, the final distribution can be obtained using Minimum Discrimination Information (MDI) adaptation [4]. This latter approach is particularly interesting since it proposes a flexible adaptation scheme in which constraints on the targeted distribution can be almost freely set according to the adaptation task. In many topic adaptation works using MDI, constraints are derived from $n$-gram probabilities trained on topic-specific corpora. However, since these corpora are rather small to estimate reliable statistics, unigram probabilities are frequently used as the sole information source [5]. To circumvent this problem, [6] proposes to also consider reliable higher order $n$-grams by computing confidence intervals from which inequality constraints are derived. Notwithstanding the good results of these works, they still do not take into consideration specifities of the topic adaptation task since $n$-grams are processed the same way, whatever their relevance w.r.t. the considered topic. Other works precisely seek to encapsulate more topic-specific information using MDI and various probabilistic latent semantic analysis techniques [7]. Probabilities to encounter a word in a given text are computed based on a word-document co-occurrence matrix, *a priori* trained and decomposed into a static number of concepts. Finally, these unigrams are used to constraint the adapted LM. While these approaches propose a more adequate solution for the topic LM adaptation task, they rely on abstract concepts rather than directly on words which makes them difficult to combine with standard NLP techniques, as we are interested in. Furthermore, these techniques use pre-calculated topic knowledge which is excluded from our method, since we want it to be completely unsupervised.

In this paper, we aim at better understanding mechanisms that are useful for topic LM adaptation using the MDI framework. More precisely, this work adresses the problems of extracting and combining appropriate topic-specific features by using NLP techniques, which results in a new completely unsupervised LM adaptation method. While Section 2 recalls MDI main principles, Section 3 presents our topic-based feature selection and gathering method for LM adaptation. Finally, experiments and results are reported in Section 4.

## 2. Minimum discriminant information language model adaptation

The goal of MDI adaptation is to find out a new LM whose probability distribution satisfies some constraints derived from a specific task and whose relative entropy (minimum information) with the baseline LM is minimal. This section first introduces the general principles of MDI adaptation before discussing the way it is used in the frame of corpus-based topic LM adaptation.

Let us consider a baseline distribution $P_B$ over a set $V^n$ of $n$-grams, and $k$ features of an information source, where fea-

tures are observable characteristics of the source. The basic idea of MDI adaptation is to get the adapted distribution $P_A$ which solves a constraint system derived from the features, where each constraint $i$ restrains a mass $K_i$ to be spread over the $n$-grams recognized by the feature $i$ [8]. Depending on the features chosen, the recognition criterion is also defined according to the adaptation task. This can be expressed by the constraints:

$$\langle f_i, P_A \rangle = K_i, \quad \forall i \in [1..k] \tag{1}$$

where

$$\langle f_i, P_A \rangle = \sum_{hw \in V^n} f_i(hw) P_A(hw), \tag{2}$$

and $f_i$ is a feature function over $n$-grams $hw$, defined as:

$$f_i(hw) = \begin{cases} 1 & \text{if } hw \text{ is recognized by feature } i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then, $P_A$ is defined as the solution of (1) which minimizes the Kullback-Leibler divergence with respect to $P_B$:

$$P_A = \arg\min_P D_{KL}(P||P_B). \tag{4}$$

Considering the constraint system (1) to be consistent, $P_A$ can be computed using the Generalized Iterative Scaling (GIS) algorithm [9]. Due to computation complexity, GIS is often iterated only once, leading to the following rough $P_A(hw)$ estimate:

$$P_A(hw) \approx P_B(hw) \prod_{i=1}^{k} \left( \frac{K_i}{\langle f_i, P_B \rangle} \right)^{\frac{f_i(hw)}{N_{hw}}} \tag{5}$$

$$= P_B(hw) \times \alpha(hw) \tag{6}$$

where $N_{hw}$ is the number of constraints recognized by the sequence $hw$, and $\alpha(hw)$ is called *scaling factor*. Finally, to fit LM structure, adapted conditionals can be written as:

$$P_A(w|h) = \frac{P_B(w|h) \times \alpha(hw)}{\sum_{\hat{w}} P_B(\hat{w}|h)\alpha(h\hat{w})}. \tag{7}$$

Since topic-specific adaptation corpora are usually small[1], the generic solution presented above is traditionally implemented in a manner that overcomes data sparseness and reliable probability estimating problems: $n$-grams $hw \in V^n$ sharing a same final token $w$ are gathered into a same feature, and are rescaled according to the sole unigram probability $P(w)$ trained on the topic-specific corpus and supposed as reliable [10]. Given the topic-specific corpus $\mathcal{C}_a$ and its distribution $P_a$, the constraint system (1) can be written as:

$$\langle f_{\hat{w}}, P_A \rangle = P_a(\hat{w}), \quad \forall \hat{w} \in V \tag{8}$$

where $V$ is the ASR system vocabulary and $f_{\hat{w}}$ is defined as:

$$f_{\hat{w}}(hw) = \begin{cases} 1 & \text{if } hw \text{ ends with } \hat{w}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

According to (5), the scaling factor $\alpha(hw)$ reduces then to:

$$\alpha(hw) = \alpha(w) = \frac{P_a(w)}{P_B(w)} \tag{10}$$

In many works [5, 7, 10], $\alpha(w)$ is exponentially smoothed by a coefficient lower than 1, optimized on heldout data. However, in our experiments, we chose to use (10) as it is, since this paper does not seek to perfectly tune a LM adaptation but rather aims at better understanding mechanisms that are useful for topic adaptation.

[1] In this paper, the average size of the topic-specific corpora is of about 700,000 words, which is very low w.r.t. the 400M words used to train the baseline LM.

# 3. Topic-specific constraint building

Contrary to other MDI topic adaptation works, we consider that topic adaptation should not rely on all the ASR system vocabulary words, but rather only on the few ones which most contribute to the topic. These latter words constitute the *topic terminology* and are referred as *terms*. We also consider that, among these terms, some words have close meanings and, thus, play a same role within the topic adaptation. As a consequence, our goal is both to identify and to combine these terms in order to build topic-specific features. Based on these features, only a few $n$-gram probabilities will be re-estimated, keeping other probabilities untouched. After presenting a method to acquire a topic terminology from a topic-specific corpus, this section studies how the topic terminology can be integrated into the MDI framework through the questions of feature selection and of feature gathering. For illustration purposes, we report here results obtained on a development set of spoken documents for which topic-specific corpora have been automatically retrieved, as described in Section 4.

## 3.1. Topic terminology acquisition

Based on a topic-specific corpus, terminology acquisition seeks to highlight words which represent substantial notions of the topic. To do this, beyond methods coming from the terminology domain [11], our approach relies on Information Retrieval (IR) methods [12]. Given the topic-specific corpus $\mathcal{C}_a$, each document $d$ of $\mathcal{C}_a$ is projected into a high dimension space using the TF-IDF criterion leading to a normalized vector $\overrightarrow{v_d}$:

$$\overrightarrow{v_d} = \left( \begin{array}{ccc} \sigma_d(w_1) & \cdots & \sigma_d(w_N) \end{array} \right) \tag{11}$$

where $\sigma_d(w_i)$ is a score depending on the frequency of the word $w_i$ in $d$ and on the number of documents containing $w_i$ in a reference corpus[2]. In practice, when computing these scores, lemmas[3] are considered instead of words since the topic characterization task does not depend on the inflexion information. Then, topic characterization $\overrightarrow{T}$ of the whole corpus is computed as the average of all the normalized document vectors:

$$\overrightarrow{T} = \frac{1}{|\mathcal{C}_a|} \times \sum_{d \in \mathcal{C}_a} \overrightarrow{v_d} \tag{12}$$

in which words with highest scores are the most supposed to be related to the topic. Finally, a topic terminology is defined as the set of the $n$ words with the highest scores in $\overrightarrow{T}$, noted as $T_n$. As an illustation of this method, Table 1 presents the set $T_{30}$ obtained for a corpus dealing with atypical pneumonia[4].

## 3.2. Feature selection

Using such a terminology, one wish not to build constraints for all the $n$-grams but only for those recognized by a term. This brings the constraint set (8) to only consider unigrams of a terminology, as follows:

$$\langle f_{\hat{w}}, P_A \rangle = P_a(\hat{w}), \quad \forall \hat{w} \in T_n \tag{13}$$

where $n$ is an empirically set parameter. As it can be shown from (5), scaling factor of $n$-grams which are not recognized by

[2] 800,000 articles from the French newspaper *Le Monde*, 1987–2003.

[3] A lemma is a canonical form of a word. For example, plural nouns are reduced to their singular form, conjugated verbs are reduced to their infinitive form...

[4] The words *Toronto* and *Canada* are listed here since the spoken document for which the topic-specific corpus has been retrieved is about a new screening method developed by Canadian researchers.

| | | | | | |
|---|---|---|---|---|---|
| #1 | pneumonia | #11 | psychosis | #21 | Canada |
| #2 | atypical | #12 | psychoses | #22 | pneumopathy |
| #3 | SARS | #13 | Toronto | #23 | hospital |
| #4 | WHO | #14 | respiratory | #24 | hospitals |
| #5 | virus | #15 | Hong-Kong | #25 | death |
| #6 | disease | #16 | case | #26 | test |
| #7 | diseases | #17 | cases | #27 | tests |
| #8 | epidemic | #18 | flu | #28 | patient |
| #9 | epidemics | #19 | symptom | #29 | patients |
| #10 | health | #20 | symptoms | #30 | China |

Table 1: *List of the 30 words with the highest scores obtained from a corpus of* 200 *documents about "atypical pneumonia".*

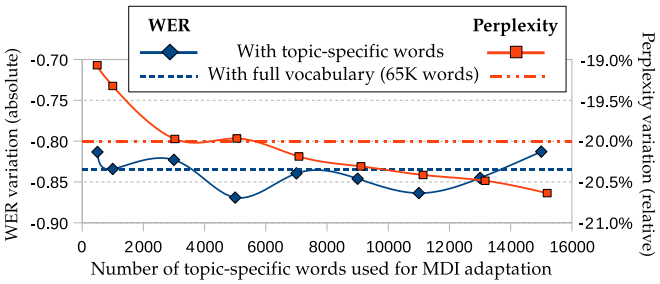


Figure 1: *Influence of the number of terms selected on WER and perplexity.*

any topic-specific word reduces to 1, *i.e.*, their probability is directly reported from the baseline LM except the normalization factor. Figure 1 presents word error rate (WER) and perplexity variations measured on our development set using either topic terminologies of different sizes or using the whole vocabulary. It appears that WER and perplexity gains obtained with topic terminologies are about the same as the one using classical unigram rescaling, even when considering only 500 words for the adaptation. One possible hypothesis to explain these results is that only topic-specific words contribute to perform the topic LM adaptation. Hence, other experiments have been carried out alternately using or excluding topic terminologies of 500 words ($T_{500}$) and $5,000$ words ($T_{5000}$). Perplexity, WER and lemma error rate[5] (LER) of Table 2 clearly show that adaptation has no effect when probabilities relating to topic-specific words are not adapted, which confirms our working hypothesis.

### 3.3. Feature gathering

In addition to selected features, the feature function used is also an important parameter in MDI adaptation since it splits $n$-grams in different classes, each corresponding to a target probability mass. In standard unigram rescaling method, $n$-grams are gathered according to their last word. However, semantic similarities shared by some words within the topic urge on gathering them into a same feature through the use of more appropriate feature functions. Hence, we tested two feature functions: a first one based on lemmas, defined by:

$$f_\ell(hw) = \begin{cases} 1 & \text{if } \ell \text{ is the lemma of } w, \\ 0 & \text{otherwise,} \end{cases} \quad (14)$$

[5]LER is the WER measured on lemmatized lexical words, *i.e.*, nouns, adjectives and non-modal verbs reduced to their lemmatized form.

| | Perplexity | WER | LER |
|---|---|---|---|
| Baseline | 96.9 | 22.1 | 19.4 |
| Linear interpolation | 80.1 (-17%) | 21.4 (-0.7) | 18.6 (-0.8) |
| MDI based on V | 75.5 (-20%) | 21.3 (-0.8) | 18.3 (-1.1) |
| MDI based on $T_{500}$ | 76.7 (-19%) | 21.3 (-0.8) | 18.5 (-0.9) |
| MDI based on $T_{5000}$ | 75.5 (-20%) | 21.2 (-0.9) | 18.4 (-1.0) |
| MDI based on V – $T_{500}$ | 94.8 (-2%) | 21.9 (-0.2) | 19.3 (-0.1) |
| MDI based on V – $T_{5000}$ | 95.4 (-2%) | 22.0 (-0.1) | 19.4 ( 0.0) |

Table 2: *Perplexity, WER and LER measured on the development set without topic LM adaptation (Baseline) and with different adaptation methods. In brackets, average perplexity relative variations and absolute WER/LER variations.*

| | | $T_{500}$ | $T_{5000}$ |
|---|---|---|---|
| PPL | No gathering | 76.7 (-19%) | 75.5 (-20%) |
| PPL | Gathered by lemmas | 77.0 (-19%) | 74.2 (-20%) |
| PPL | All gathered | 89.2 (-7%) | 94.0 (-3%) |
| WER | No gathering | 21.3 (-0.8) | 21.2 (-0.9) |
| WER | Gathered by lemmas | 21.4 (-0.7) | 21.4 (-0.7) |
| WER | All gathered | 21.7 (-0.4) | 21.92 (-0.2) |

Table 3: *Perplexity and WER measured for different feature functions for two topic terminology sizes.*

and a second one which gathers all the words of a given terminology $T$ in one same class:

$$f_T(hw) = \begin{cases} 1 & \text{if } w \text{ is in } T, \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

In practice, this latter feature function leads to consider only one constraint, meaning that all the topic-specific words have the same importance. Perplexity and WER measured using these two feature functions are compared to the ones obtained with the standard unigram method in Table 3. On the one hand, it appears that function $f_T$ is the worst function, probably because it is too rough. On the other hand, results show that gathering $n$-grams based on lemmas is nearly equivalent to the standard gathering based on words: while perplexity improvements are the same, WER is slightly worse when using lemmas. In the light of these preliminary results, we chose to discard the feature function $f_T$ for the backend experiments.

## 4. Experiments and results

Our ASR system is a multipass 65K words system based on two general-purpose LMs: a 3-gram LM to create word graphs from acoustic features and a 4-gram LM to score word graphs. Experiments are carried out on 172 thematically coherent segments from 6 hours of Broadcast News (BN) shows from the French radio BN corpus ESTER [13]. These segments, coming from 3 different broadcasters and all dated from the same period of time, are spread over diversified topics (war in Iraq, national politics, sports, weather, *etc.*) and lengths (from 30 to 2,000 words). This collection is divided into a development set and a test set of respectively 91 and 81 segments. For each segment, a topic-specific corpus is automatically retrieved from the Internet, as detailed in [2], and 3-gram and 4-gram adapted LMs are computed before generating new word graphs and a new transcription. In this section, results are presented for the test set using topic terminologies of $5,000$ words, which led to the best

| | | Perplexity | WER | LER |
|---|---|---|---|---|
| Baseline | | 96.7 | 20.7 | 18.3 |
| Linear interpolation | | 78.8 (-16%) | 20.2 (-0.5) | 17.4 (-0.9) |
| MDI based on | $f_{\hat{w}} + \mathrm{V}$ | **74.7 (-21%)** | 20.1 (-0.6) | 17.3 (-1.0) |
| | $f_{\hat{w}} + \mathrm{T}_{500}$ | 76.5 (-19%) | 20.2 (-0.5) | 17.4 (-0.9) |
| | $f_{\hat{w}} + \mathrm{T}_{5000}$ | 75.5 (-20%) | **20.0 (-0.7)** | **17.2 (-1.1)** |
| | $f_{\ell} + \mathrm{T}_{500}$ | 76.8 (-19%) | 20.2 (-0.5) | 17.5 (-0.8) |
| | $f_{\ell} + \mathrm{T}_{5000}$ | 75.7 (-20%) | 20.1 (-0.6) | 17.2 (-1.1) |

Table 4: *Perplexity, WER and LER measured on the test set without topic LM adaptation (Baseline) and with different LM adaptation methods. In brackets, average perplexity relative variations and absolute WER/LER variations.*

| $t$-TEST | $f_{\hat{w}} + \mathrm{V}$ | $f_{\hat{w}} + \mathrm{T}_{500}$ | $f_{\hat{w}} + \mathrm{T}_{5000}$ |
|---|---|---|---|
| Baseline | $8.8 \times 10^{-5}$ | $3.4 \times 10^{-7}$ | $\mathbf{1.5 \times 10^{-7}}$ |
| $f_{\hat{w}} + \mathrm{V}$ | – | 0.67 | 0.63 |

| WILCOXON | $f_{\hat{w}} + \mathrm{V}$ | $f_{\hat{w}} + \mathrm{T}_{500}$ | $f_{\hat{w}} + \mathrm{T}_{5000}$ |
|---|---|---|---|
| Baseline | $2.5 \times 10^{-4}$ | $1.2 \times 10^{-6}$ | $\mathbf{3.3 \times 10^{-7}}$ |
| $f_{\hat{w}} + \mathrm{V}$ | – | 0.77 | 0.45 |

Table 5: *Statistical significances for the paired $t$-test (top) and for the paired Wilcoxon test (bottom) between WERs measured without topic LM adaptation (Baseline) and with different MDI adaptation methods. Confidence level is $\alpha = 0.05$.*

WER results on the development set, and of $500$ words, which is an extreme case since it represents less than $1\,\%$ of the ASR system vocabulary. Adaptation is processed using a feature function either based on words ($f_{\hat{w}}$) or based on lemmas ($f_{\ell}$).

Perplexity, WER and LER are measured for each segment and are compared to those obtained with our baseline LM, *i.e.*, without topic adaptation, and using linear interpolation with an interpolation factor set to $0.8$ after optimizing on the development set, and with MDI unigram rescaling (Table 4). First, baseline results are much better than those obtained on the development set, especially the WER absolute difference is of $1.4$. Hence, one can notice a slight overall decrease of WER improvements, whereas perplexity and LER improvements are of the same order as previously. Then, as for the development set, terminology-based unigram rescaling results are comparable to those of standard unigram rescaling. This is confirmed by low significance values between these two methods in Table 5. Nonetheless, WER gains w.r.t. baseline results appear to be much more significant when using topic terminologies. This is the case both when using $T_{5000}$ and $f_{\hat{w}}$, which leads to the best recognition improvements, and when using a small topic terminology ($T_{500}$), though it results in slightly worse improvements than those of standard unigram rescaling. This latter point shows that only a few words contribute to establish the topic of a document and, as a consequence, that topic adaptation must not be performed for all the $n$-grams. This is all the more reasonable since adaptation data is sparse and not always reliable. Finally, recognition rates still tend to be slighly worse when $n$-grams are gathered based on lemmas instead of words. This leads us to conclude that our assumption about the weak role of inflections for topic adaptation is probably too categorical, be it for the feature gathering step as well as for the topic terminology acquisition task. To bridge over the slight recognition differences, it could be interesting to integrate importance factors inside the $n$-gram classes, *e.g.*, by considering non-binary feature functions.

# 5. Conclusion and future work

In this paper, we have presented a new LM adaptation method which specializes the general MDI adaptation framework for the specific case of topic adaptation using IR and NLP techniques. Especially, the standard unigram rescaling method has been refined, without relying on *a priori* knowledge, by automatically extracting topic terminologies from which different constraint building methods have been proposed. Experiments lead to significant perplexity and recognition gains which outperform gains of standard LM adaptation techniques. These results are all the more interesting since it appears that topic adaptation can be properly performed with very few topic-specific words whereas it cannot when these latter words are discarded.

Future work should explore three main directions. First, as discussed in Section 3.3, feature function used in MDI should better fit the topic adaptation task. Especially, relationships between $n$-grams should integrate more linguistic knowledge, *e.g.*, lemmas, semantical similarities, *etc.*, to adjust biases implied by adaptation data sparseness. Second, one should apply the feature selection scheme to higher order $n$-grams to characterize topic-specific phrases. However, this arises even more the sparseness problem. Finally, a topic-specific corpus may deal with different thematic aspects which should probably be treated separately. For example, when considering a document about a movie dealing with the conjugal duties, the topic-specific corpus may contain documents about cinema while others would get onto family life problems. For a better adaptation method, one should seek to identify the different thematic aspects of an adaptation corpus before processing them separately.

# 6. References


[1] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communications*, 2004.

[2] G. Lecorvé, G. Gravier, and P. Sébillot, "An unsupervised web-based topic language model adaptation method," in *Proc. ICASSP*, 2008.

[3] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Using information retrieval methods for language model adaptation," in *Proc. Eurospeech*, 2001.

[4] S. Della Pietra, V. Della Pietra, R. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," in *Proc. ICASSP*, 1992.

[5] M. Federico, "Efficient language model adaptation through mdi estimation," in *Proc. Eurospeech*, 1999.

[6] C.-H. Chueh and J.-T. Chien, "Reliable feature selection for language model adaptation," in *Proc. ICASSP*, 2008.

[7] Y.-C. Tam and T. Schultz, "Unsupervised language model adaptation using latent semantic marginals," in *Proc. Interspeech*, 2006.

[8] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language*, 1996.

[9] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals of Mathematical Statistics*, 1972.

[10] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. Eurospeech*, 1997.

[11] P. Drouin, "Term extraction using non-technical corpora as a point of leverage," *Terminology*, 2003.

[12] A. Kilgarriff, "Comparing corpora," *International Journal of Corpus Linguistics*, 2001.

[13] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. Eurospeech*, 2005.