

La campagne d'évaluation RIMES pour la reconnaissance de courriers manuscrits

<http://www.int-evry.fr/rimes>

Emmanuèle Grosicki¹ – Edouard Geoffrois¹ – Matthieu Carré² – Emmanuel Augustin³ – Françoise Prêteux²

¹ CEP Arcueil

16 bis, avenue de la Côte d'Or, 94114 Arcueil cedex

² INT/ARTEMIS Unit

9, rue Charles Fourier, 91011 Evry Cedex

³ Société A2IA

40 bis, rue Fabert, 75007 Paris

emmanuele.grosicki,edouard.geoffrois@etca.fr,

Matthieu.Carre,Francoise.Preteux@int-evry.fr, emmanuel.augustin@a2ia.com

Résumé : *Cet article présente la campagne d'évaluation RIMES (Reconnaissance et Indexation de données Manuscrites et de fac similES) du programme Techno-Vision à travers :*

- son originalité,
- ses objectifs,
- les différentes ressources mises en place : la construction d'une grande base de courriers annotée et les différentes tâches proposées à l'évaluation avec leur métrique.

Mots-clés : RIMES, Campagne d'évaluation, Tâches, Métriques, Base de données.

1 Introduction

Le projet RIMES s'inscrit dans le cadre du programme Techno-Vision lancé en 2004 par le Ministère de la Recherche et le Ministère de la Défense pour créer une dynamique de l'évaluation de technologies de vision par ordinateur et du traitement d'image. Les domaines concernés sont la vision pour la robotique, la vidéo-surveillance, l'indexation d'images et de séquences d'images, l'analyse de documents complexes, la stéréovision, la biométrie, le traitement d'images aériennes et satellites et l'imagerie médicale. Dix projets ont ainsi été mis en place. Le projet RIMES est l'un d'entre eux et couvre les problématiques dédiées à l'analyse automatique de documents. Il vise à permettre l'évaluation objective de systèmes de reconnaissance et d'indexation de courriers manuscrits en français, tels que ceux envoyés par des individus à des entreprises. Il s'agit d'un projet très novateur pour ce domaine car, à l'heure actuelle, très peu de campagnes d'ampleur ont vu le jour. On peut citer les campagnes réalisées par le NIST qui se sont limitées aux tâches de reconnaissance de caractères ou de mots isolés [1] et les quelques campagnes sur des images un peu plus complexes organisées en marge de conférences telles que GREC ou IC-DAR qui ne disposaient que de moyens limités.

Concernant les bases de données, une seule base de do-

cuments manuscrits complets est librement disponible [2], mais elle est monospace et de petite taille (environ 4500 mots sur 25 pages). A notre connaissance, aucune base de fac-similés n'est publiquement disponible, celles existantes étant de petite taille ou encore partiellement annotées, comme par exemple celle du projet Majordome [3].

Le projet RIMES a donc comme ambition de créer une base de données conséquente de documents complexes, de mettre en place le protocole de la campagne d'évaluation utilisant ces données (tâches, métriques,...), de conduire ladite campagne et de mettre la base résultante à la disposition de la communauté scientifique à l'issue de la campagne.

2 La base de données

Comme il a été rappelé en introduction, peu de grandes bases de données de documents complexes incluant de l'écriture manuscrite sont à ce jour disponibles. C'est pourquoi, un des enjeux du projet RIMES consiste à mettre en place une base conséquente de documents peu contraints comme ceux envoyés par des individus à des entreprises. N'étant pas possible pour des raisons juridiques évidentes d'utiliser de véritables courriers, il a été choisi de faire écrire des courriers factices par des scripteurs volontaires sur la base de scénarii réalistes pré-établis. On vise ainsi la création de 8000 courriers comportant deux ou trois pages : Lettre Manuscrite (L), Questionnaire Technique (Q), et éventuellement page de garde de Fax (F) (cf figure 1). Ces courriers sont écrits par 1600 scripteurs différents sur la base de 9 grandes classes de scénarii pouvant comporter chacune plusieurs variantes :

- changement de données personnelles,
- demande d'information,
- difficulté de paiement,
- ouverture (ex : compte),
- fermeture,
- modification de contrat/commande,
- réclamation,

- relance de courrier sans réponse,
- gestion de sinistre.

Il est clairement souhaité dans le projet RIMES d'obtenir une base de courriers qui soient les plus réalistes possible. Pour cela, aucune consigne n'est donnée aux scripteurs sur le type de vocabulaire à utiliser ou sur la disposition des informations dans la lettre ou le fax. Il est seulement demandé d'écrire lisiblement sur du papier blanc avec une encre foncée et d'éviter les ratures et/ou ajouts. Cette indication permet d'obtenir une base de documents "propres" et ne constitue pas à notre sens un obstacle à l'obtention d'une base de courriers réalistes.

La numérisation de la base est, quant à elle, faite par un scanner professionnel uniforme sur toute la base. D'autres numérisations sont également possibles sur certains courriers avec des scanners personnels et/ou fax personnels. Un site pour la collecte des courriers auprès de scripteurs volontaires est mis en place <http://www.scribio.org> qui gère notamment la rétribution des scripteurs pour l'écriture de 5 courriers par des chèques-cadeaux.

3 Description des tâches et des métriques associées

Le projet couvre l'essentiel des problématiques de l'analyse automatique de document, que ce soit dans le domaine de la reconnaissance de l'écriture manuscrite ou dans celui de l'analyse de la structure des documents, et les combine autour d'une tâche complète et réaliste, l'analyse de courriers et de fax pour en déterminer les informations essentielles (thème, identité de l'expéditeur, ...).

L'évaluation du projet RIMES porte ainsi sur les 5 thèmes suivants :

- structuration de document (**S**),
- reconnaissance de caractères ou de mots manuscrits (**R**),
- reconnaissance de scripteurs (**Sp**),
- reconnaissance de logos (**L**),
- extraction d'informations (**E**).

A partir de ces 5 thèmes, des tâches de difficultés variables ont été définies portant sur les données d'entrée suivantes : **C** pour images de Caractères isolés, **M** pour images de Mots isolés, **B** pour images de Blocs, **L** pour images de Lettres Manuscrites, **Q** pour images de Questionnaires, **F** pour images de Fax, **G** pour images de loGo.

Pour chaque tâche, une métrique est définie pour l'évaluation avec comme objectif qu'elle soit à la fois simple, générale pour prendre en compte plusieurs sorties possibles des systèmes automatiques et précise pour permettre une évaluation fine des différents systèmes.

La notion de référence ou vérité-terrain (V.T.) renvoie aux fichiers XML (cf section 6) contenant l'ensemble des informations "exactes" relatives à une image.

Mme Séverine CHAUVIN
44 rue neuve
70 250 Plancher les Mirois
Référence 0017

Traçabilité de
Plancher les Mirois

P. les Mirois le 2 février 2005

Madame, Monsieur,

Je me suis aperçu que le compte non liquidé vous allez prélever la base d'habitation à changer. Donc je vous joins mon nouveau RIB. Je vous en souhaite bonne réception et vous prie de m'envoyer l'expression de mes salutations respectueuses.

P.S. mon nouveau RIB
à la Société Générale
de Plancher les Mirois

S. Chauvin

Questionnaire

Code d'envoi

Identité fictive

Nom Prénom

Adresse

Code postal Ville

Tél.

Scénario

Catégorie

Objet

Complément

Référence client éventuelle (.....)

Interlocuteur

Adresse

Code postal Ville

Rédaction

Temps passé à rédiger ce courrier minutes.

Type de stylo utilisé : bille plume feutre

Contribution & Rétribution

Cocher les éléments de votre contribution pour le scénario traité :

(+6) J'envoie ma contribution par voie postale à la date du

(+26) En plus de la lettre manuscrite et du présent questionnaire (obligatoires), ma contribution contient une page de garde de fax.

(+16) Avant mon envoi postal, j'ai faxé ma contribution le

(+16) J'ai envoyé une version numérisée de ma contribution le par courriel. La résolution de mon fichier image est dpi/ppp, et son format est PNG GIF TIFF. La marque et le modèle de mon scanner sont

MINISTÈRE DE LA DÉFENSE

TÉLÉCOPIÉ Non protégé Diffusion restreinte

N° Date :

DESTINATAIRES : N° de télécopieur :

Expéditeur :
DGA
DIRECTION GÉNÉRALE POUR L'ARMEMENT
DIRECTION DES CONTRÔLES OPÉRATIENS ET FINANCIERS
Centre technique d'Alsace

Caractères d'urgence : immédiat urgent routine

OBJET : réclamation

ZONE DE TEXTE :

La M. Stover a joint copie de la lettre Stover à votre service de réclamation

Cordialement
Dp

ELG

Centre technique d'Alsace - 1526, avenue Prieur de la Cité d'Or 6414 Annuel Cedex
Téléphone : 33 (0) 42 31 92 39 - Télécopie : 33 (0) 42 31 92 38

FIG. 1 – Exemple de courrier : une lettre, un questionnaire et éventuellement une page de garde de fax.

3.1 Thème 1 : structuration de document (S)

3.1.1 Définition des tâches

Il s'agit de détourner dans des lettres et des fax, les écritures correspondant à certains champs utiles pour un traitement automatique comme les coordonnées expéditeur, destinataire, la date, le lieu, l'objet du courrier, corps de texte, les logos ...

Deux tâches sont ainsi proposées : une portant sur les lettres (tâche **SL**) et une sur les fax (tâche **SF**). Les intitulés des champs à localiser sont différents pour les fax et pour les lettres. Pour les lettres, ils sont au nombre de 8 (Coordonnées expéditeur, Coordonnées destinataire, Date/Lieu, Objet, Ouverture, Corps de texte, Signature, PS/PJ) alors que pour les fax, ils sont au nombre de 3 et correspondent uniquement à la nature de l'écriture (manuscrite, dactylographiée) et aux logos. Cette différence s'explique par le fait que les conventions d'usage sont beaucoup plus fortes pour les lettres que pour les fax. Et par conséquent, l'identification de la nature même des champs nécessiterait dans le cas des fax la mise en place d'un module d'interprétation, même partiel, des écritures ; ce qui relève d'autres tâches.

Concernant la délimitation des zones d'écritures associées aux champs, il a été choisi pour des raisons de simplicité d'utiliser des boîtes englobantes (rectangle dont les côtés sont parallèles à la page). Dans la plupart des cas, une seule boîte englobante est suffisante par champ. Toutefois, dans certains cas, comme par exemple lorsque les écritures sont penchées, l'utilisation d'une seule boîte peut conduire à des intersections inter-champs gênants pour le traitement automatique des courriers. La solution alors choisie consiste dans les V.T. à utiliser lorsque c'est nécessaire plusieurs boîtes englobantes pour un même champ. La lettre de la figure 2 illustre ce point pour les champs "Objet", "Date/Lieu", "Corps de texte".

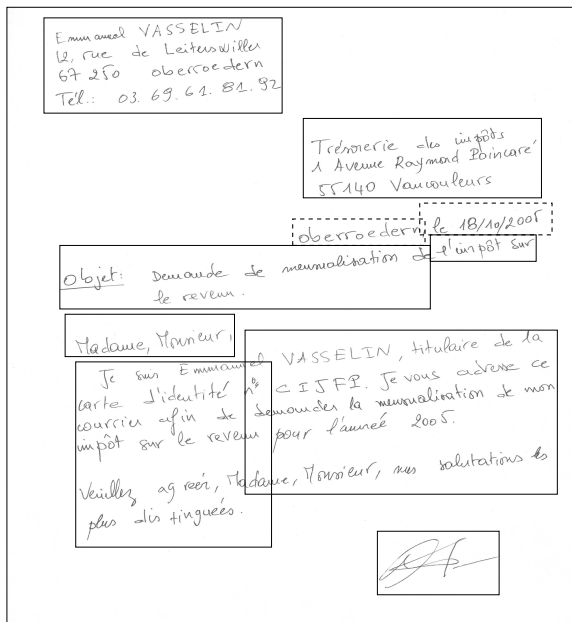


FIG. 2 – Utilisation de plusieurs boîtes englobantes pour éviter les recouvrements inter-champs.

3.2 Métrique choisie

Il existera toujours une différence entre les boîtes englobantes des vérités-terrain et celles des sorties des systèmes automatiques, différence pouvant être accentuée dans le cas de l'utilisation de plusieurs boîtes englobantes pour un même champ. Pour comparer les sorties automatiques avec la vérité-terrain, on utilise classiquement un taux de recouvrement de boîtes. Dans notre cadre, cette métrique présenterait l'inconvénient de tenir compte dans le taux d'erreur des différences causées par des pixels blancs du fond. Pour atténuer cet inconvénient, le taux classique de recouvrement de boîtes est pondéré par les niveaux de gris des pixels [4].

Le principe de la métrique est donnée ci-dessous :

A partir des coordonnées des boîtes englobantes, on associe à chaque pixel de l'image une étiquette :

- Si ce dernier appartient à une boîte englobante, son étiquette est celle de la boîte,
- Sinon, par défaut, son étiquette est 'Fond'.

La métrique choisie consiste alors à comparer pour chaque pixel l'étiquette donnée par la vérité-terrain (référence) avec celles données par les différents fichiers des participants (hypothèses). Les pixels dont l'étiquette de la référence diffère de celle de l'hypothèse sont comptabilisés comme des erreurs. Le taux d'erreur proposé correspond à la somme des niveaux de gris des pixels mal classés, divisée par la somme des niveaux de gris de tous les pixels de l'image.

On voit bien que la métrique choisie ne limite pas le nombre de boîtes englobantes par un même champ mais pénalise les recouvrements de boîtes englobantes de types différents (un pixel ne peut pas appartenir à deux boîtes englobantes portant des étiquettes différentes). La figure 3 illustre la métrique sur un exemple simple de structuration de lettre.

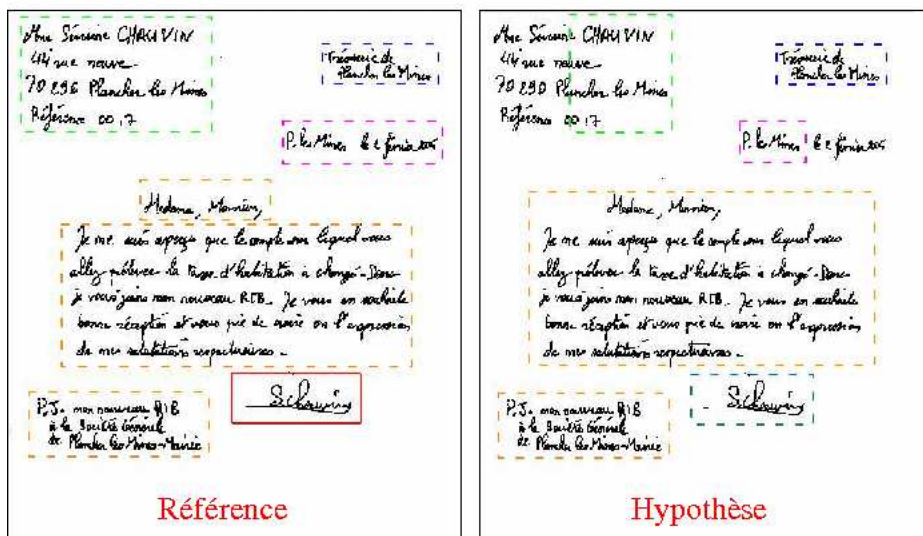
3.3 Thème 2 : reconnaissance d'écriture manuscrite (R)

3.3.1 Définition des tâches

Cette tâche vise à évaluer les algorithmes de reconnaissance d'écriture manuscrite. Elle concerne les trois types de documents (lettres, questionnaires et pages de gardes de fax). Elle correspond aussi bien à la reconnaissance de caractères, de mots isolés que de mots en contexte, *i.e.* dans le document complet. Les transcriptions des vérités-terrain sont fidèles à ce qui est écrit dans les lettres (fautes d'orthographe et de grammaire incluses). Toutefois, pour ne pas pénaliser les systèmes qui corrigeraient les fautes, un dictionnaire d'équivalence est mis en place pour prendre en compte au moment du scoring les mots avec plusieurs orthographes comme j'essaie/j'essaye, événement/évènement ou encore ultrason/ultra son et ceux mal orthographiés par les scribes.

5 tâches sont proposées de difficultés croissantes pour ce thème :

- **RC** (reconnaissance de caractères) : la reconnaissance s'effectue sur des imagerie de caractères (chiffres et lettres) extraites des questionnaires.
- **RM** (reconnaissance de mots) : la reconnaissance s'effectue sur des imagerie de mots extraites des lettres



• Nombre de pixels mal classés (pondérés par leur niveau de gris) : 3247

Taux d'erreur = 17%

FIG. 3 – Exemple de scoring pour la tâche SL

manuscrites.

- **RB** (reconnaissance de blocs) : la reconnaissance s'effectue sur des blocs extraits des lettres et des fax (exemple : corps de la lettre, coordonnées expéditeur,...).
- **RL** (reconnaissance de lettre) : pour des raisons de simplicité et pour éviter toute confusion dans l'ordre de lecture des différentes écritures, cette tâche est limitée à la transcription des champs : corps de texte et coordonnées expéditeur (lorsque ces dernières apparaissent dans un unique bloc où l'ordre des mots n'est pas ambigu).
- **RQ** (reconnaissance de questionnaire) : cette tâche correspond à la transcription de tous les champs du questionnaire.

3.3.2 Métriques choisies

Pour toutes les tâches, l'évaluation se fait au niveau mot et au niveau caractère.

Il faut distinguer le cas des tâches **RC** et **RM** pour lesquelles, le taux d'erreur correspond juste au nombre de caractères ou de mots, bien reconnus sur le nombre total de caractères ou de mots, et le cas des tâches (**RB**, **RL** et **RQ**) pour lesquelles un alignement entre la référence (vérité-terrain) et les sorties est nécessaire et le taux d'erreur correspond à la somme du taux de caractères ou de mots substitués, insérés ou supprimés. L'alignement est réalisé grâce à l'outil ScLite du NIST, très largement utilisé en traitement de la parole [5] et diffusé librement sur le site web du NIST (<http://www.nist.gov/speech>). La figure 4 donne un exemple d'alignement réalisé par cet outil avec le score obtenu.

Notons que dans le cas de la tâche **RM**, les systèmes automatiques renvoyant le plus souvent pour chaque transcription une liste de mots avec un taux de confiance associé à chacun, le taux d'erreur est également calculé, pour chaque transcription, sur les N meilleurs taux de confiance (N étant traditionnellement choisi égal à 10).

3.4 Thème 3 : Reconnaissance de Scripteurs (Sp)

3.4.1 Définition des tâches

Ce thème regroupe trois tâches : une tâche de vérification **SpVM** et deux tâches d'identification avec rejet de scripteurs **SpIM** et **SpIB**.

Plus précisément :

- La tâche **SpVM** concerne la vérification de scripteurs sur des imagerie de mots. Il s'agit de confirmer qu'une certaine imagerie de mot a été écrite par un scripteur donné. Pour cela, le participant dispose d'une autre réalisation de ce mot écrite par le scripteur en question ainsi que de l'ensemble des imagerie de mots de la base d'apprentissage provenant de tous les scripteurs présents dans cette dernière.
- Les tâches **SpIM** et **SpIB** concernent l'identification de scripteurs avec rejet sur des imagerie respectivement de mots et de blocs corps de texte. Il s'agit d'identifier le scripteur de l'image parmi les scripteurs de la base d'apprentissage, sachant que certains scripteurs de la base de test ne font pas partie de la base

d'apprentissage. Ces tâches visent l'évaluation des systèmes automatiques utilisant uniquement comme approche la forme de l'écriture manuscrite pour différencier les scripteurs.

3.4.2 Métrique choisie

La métrique consiste simplement en un taux d'erreur scripteur : nombre de scripteurs mal reconnus ou mal authentifiés sur le nombre total de scripteurs de la base de test.

3.5 Thème 4 : Reconnaissance de Logos L

3.5.1 Définition de la tâche

Ce thème est marginalement traité dans la campagne RIMES ; ceci s'explique par le fait qu'un autre projet du programme techno-vision EPEIRES y est plus précisément dédié.

Une seule tâche est proposée (**LG**) sur ce thème qui consiste à identifier le logo présent sur des imagerie de logos de la base de test parmi ceux de la base d'apprentissage. On entend par logo une représentation graphique d'une marque commerciale ou d'un organisme (sociétés, associations, ministères,...).

3.5.2 Métrique choisie

La métrique consiste simplement en un taux d'erreur logo : nombre de logos mal reconnus sur le nombre total de logos de la base de test.

3.6 Thème 5 : Extraction d'informations (E)

3.6.1 Définition des tâches

Il s'agit d'extraire de façon automatique des lettres un certain nombre d'informations, sans contrainte sur la technique à employer. Deux tâches sont ainsi proposées sur ce thème :

- La tâche **ESn** qui consiste à déterminer la classe du scénario de la lettre parmi les 9 classes possibles,
- et la tâche **ESp** qui consiste à déterminer l'identité du scripteur de la lettre.

3.6.2 Métrique choisie

La métrique consiste simplement en un taux d'erreur respectivement sur les scénarii et les scripteurs (nombre de scénarii, respectivement scripteurs, mal reconnus sur le nombre total de lettres).

4 Organisations des tests

La campagne RIMES comporte 2 phases : une phase de tests à blanc permettant d'ajuster les métriques ainsi que les paramètres des systèmes automatiques et une phase de test officiel. Les 8000 courriers sont diffusés aux participants en 4 lots de 2000. Le premier lot sert à l'apprentissage des systèmes. Les deuxième et troisième lots sont utilisés pour des tests à blanc et le dernier lot pour le test officiel. A l'issue de chacun des tests à blanc, les vérités-terrain des lots correspondants sont distribuées et peuvent être utilisées pour l'apprentissage des systèmes automatiques.

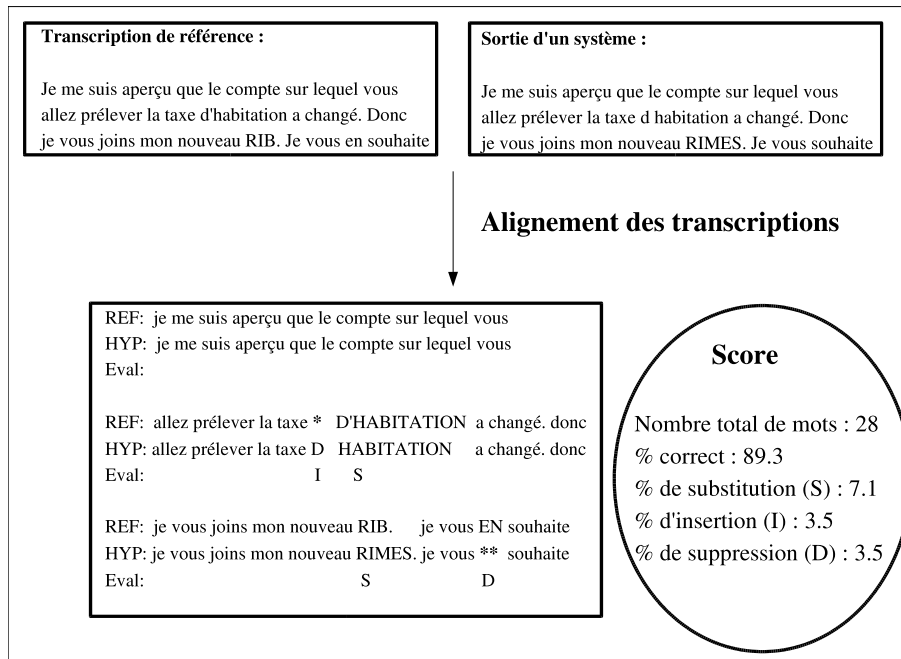


FIG. 4 – Alignement et scoring au niveau mot obtenus avec ScLite sur deux blocs de texte.

5 Conclusion

La campagne d'évaluation RIMES est la première campagne d'envergure couvrant l'ensemble des tâches relatives à la reconnaissance de l'écriture manuscrite et de la structuration de documents complexes. Elle propose ainsi autour d'une application de traitement de courriers en français tels que ceux envoyés par des individus à des entreprises 13 tâches de difficultés variables. Elle permet aussi de constituer une base de données conséquente annotée (8000 courriers de 2 à 3 pages) qui sera mise à la disposition de la communauté scientifique à l'issue de la campagne.

6 Annotation des courriers

Tous les courriers de la base sont annotés manuellement de manière à faire apparaître :

- Les informations utiles pour un traitement automatique du courrier comme la classe du scénario, le nom de l'expéditeur, du destinataire, la date, l'objet de la correspondance... Ces informations sont placées dans des balises d'en-tête sans indication de leur position dans les courriers.
- La transcription des différents champs présents dans les courriers avec leur type et les coordonnées de leurs différentes boîtes englobantes.

Les annotations de chaque image sont données dans un fichier xml.

7 Bibliographie

Références

- [1] NIST 1995, <http://www.nist.gov/srd/nistsd19.htm>.
- [2] Andrew W. Senior and Anthony J. Robinson, PAMI vol 20(3), pp. 309-321, mars 1998.

- [3] L. Likforman-Sulem, G. Chollet, P. Vaillant, N. Az-zabou, R. Blouet, S. Renouard et D. Mostefa, "Reconnaissance de noms propres et vérification d'identité dans un système de messagerie", convention MinEFI n° 01.2.93.0268, rapport final (100 pages), janvier 2004.
- [4] B.A. Yanikoglu et L. Vincent, "Pink panther : a complete environment for ground-truthing and benchmarking document page segmentation", Pattern Recognition, Vol. 31, N°9, pp. 1191-1204, 1998.
- [5] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait et K. Choukri, "ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français", avril 2004.