Ressources pour l'apprentissage, le développement et l'évaluation des systèmes de dictée vocale en français : corpus de texte, de parole et lexical

Gilles Adda[⋆], **Martine de Calmès**[†], **Lori Lamel**[⋆],
Guy Pérennou[†], **Martin Rajman**[∗◇], **Sophie Rosset**[⋆], **Jérôme Zeiliger**[‡]

⋆ LIMSI-CNRS, Groupe Traitement du langage parlé,
BP 133, 91403 Orsay cedex, FRANCE,
{gadda,lamel,rosset}@limsi.fr

† IRIT , groupe IHM-PT, Univ. Paul Sabatier,
118 route de Narbonne, 31062 Toulouse cedex, FRANCE,
{decalmes,perennou}@irit.fr

∗ Laboratoire d'Intelligence Artificielle, Ecole Polytechnique Fédérale de Lausanne,
IN Ecublens, 1015 Lausanne, SUISSE,
rajman@lia.di.epfl.ch

◇ ENST, Dpt Informatique,
49, rue Vergniaud, 75013 Paris, France,
rajman@inf.enst.fr

‡ ICP - INPG,
46, avenue Felix Viallet, 38031 Grenoble cedex 1 , FRANCE,
zeiliger@icp.grenet.fr

Résumé

Dans cet article, nous décrivons les ressources de textes, de parole et de lexiques qui ont été mis à la disposition des participants de la campagne d'évaluation "Dictée Vocale" faisant partie des Actions de Recherche Concertées "Linguistique, Informatique et Corpus Oraux".

Ces ressources ont été définies par les fournisseurs de corpus et les organisateurs, en accord avec les participants. La mise à disposition de ressources réutilisables de textes, de parole, et de lexiques pour l'apprentissage des systèmes, ainsi que la définition et la construction de corpus de texte et de parole pour le développement et l'évaluation constitue une partie majeure de cette campagne d'évaluation.

Introduction

Dans le cadre des Actions de Recherche Concertées "Linguistique, Informatique et Corpus Oraux", l'AUPELF-UREF a lancé une campagne d'évaluation des systèmes de dictée vocale parole continue à grand vocabulaire pour la langue française.

La mise à disposition de ressources de textes, de parole, et de lexiques pour l'apprentissage des systèmes, ainsi que la définition et la construction de corpus de texte et de parole pour le développement et l'évaluation constitue une partie essentielle de toute évaluation, et en particulier lorsqu'il s'agit, comme s'est le cas pour l'Action en question, d'une première dans un cadre francophone : jusqu'à présent, certains corpus en français étaient disponibles mais aucun ensemble cohérent permettant une évaluation objective de la dictée vocale à grand vocabulaire en français n'avait jamais été mise à la disposition de la communauté francophone en reconnaissance de la parole.

Pour les ressources nécessaires à l'apprentissage, les corpus oraux, ainsi que les corpus de textes ont été fournis par le LIMSI. En ce qui concerne les phases de développement et d'évaluation, les lexiques et les modèles de langage ont été faits au LIMSI, la phonétisation les lexiques a été réalisée à l'IRIT, la partie textuelle des corpus a été définie et saisie à l'ENST, la partie orale a été enregistrée au LIMSI, puis la sélection finale, le formatage, la mise sur CDROM de ces corpus (ainsi que les corpus de textes et les modèles de langage) ont été effectués à l'ICP.

Corpus Oraux d'Apprentissage

Deux corpus oraux enregistrés au LIMSI, sous-corpus du corpus BREF ont été mis à disposition des participants.

BREF ([Lamel L., Gauvain J.-L., Eskénazi M., (1991)]) est un corpus de textes lus, contenant plus de 100 heures de parole produites par 120 locuteurs.

Les textes proviennent de textes journalistiques. Cinq millions de mots du journal *Le Monde* ont été analysés et un ensemble de textes (paragraphes ou phrases) a été sélec-

tionné ([Gauvain J.-L., Lamel L., Eskénazi M., (1990)]). La première étape de l'analyse a consisté à phonétiser chaque phrase en utilisant des règles de transformation graphème-phonème et un dictionnaire d'exceptions. Les propriétés statistiques du texte ont été déterminées en comptant les occurences des phrases, des mots et unités phonétiques. Pour les phrases, nous avons pris en compte leur longueur et le type de phrase. Pour les mots, nous avons évalué le nombre de mots distincts et leurs fréquences. Les unités phonétiques comprennent les syllabes, dissyllabes, phonème, diphones et triphones.

Ces textes ont ete sélectionnés pour maximiser le nombre de contextes phonétiques et par là même, la taille du vocabulaire correspondant à ces textes (plus de 20 000 mots). Contenant 1115 diphones distincts et plus de 17 500 triphones, BREF peut être utilisé pour l'apprentissage de modèles phonétiques indépendant du vocabulaire. Des textes différents, ayant des propriétés distributionnelles comparables, ont été sélectionnés pour l'apprentissage, le développement et l'évaluation des systèmes de reconnaissance. Ces textes comprennent 18 phrases contenant tous les phonèmes français, environ 840 paragraphes, 3300 phrases courtes (12,4 mots/phrase), et 3800 phrases longues (21 mots/phrase).

80 locuteurs ont lu chacun environ 10 000 mots (soit environ 650 phrases), et 40 autres locuteurs ont lu environ la moitié. Les locuteurs, choisis parmi un groupe de plus de 250 sujets de la région parisienne, ont été rémunérés pour leur participation. Tous les sujets ont effectué un test de lecture correspondant à quelques phrases du journal "Le Monde", représentatives du type de matériel à enregistrer. Les locuteurs enregistrés ont été choisis parmi les sujets jugés capables d'accomplir cette tâche.

Les enregistrements, effectués simultanément avec un microphone Shure SM10 et un microphone Crown PCC160, ont été réalisés dans une pièce insonorisée et vérifiés sur le champ pour en assurer le contenu. Les deux canaux ont été échantillonnés a 16kHz et digitalisés sur 16 bits. Au total, 65 femmes et 55 hommes ont été enregistrés, soit 120 locuteurs d'âges compris entre 18 et 73 ans, avec une majorité entre 20 et 40 ans.

Les deux sous-corpus mis à disposition sont :

- un sous corpus correspondant au seul microphone Shure SM10 (13 CDs),
- le sous-corpus BREF-80 correspondant à la prononciation par 80 locuteurs d'environ 67 phrases par locuteurs (2 CDs).

Bien qu'à l'origine, les enregistrements aient été séparés en 3 parties (apprentissage, développement et test), tous les enregistrements peuvent être utilisés pour l'apprentissage.

Corpus de Texte

Grâce à un accord avec le journal "Le Monde", 2 années (87-88) de ce journal ont pu être mises à disposition des participants. Le LIMSI, à l'aide d'outils spécifiques ([Gauvain J.-L., Lamel L., Adda G., Mariani J., (1994)]) a normalisé ce texte afin de le rendre directement utilisable pour la construction de modèles de langage pour la reconnaissance de la parole. En effet, de grandes quantité de textes sont nécessaires pour développer les modèles de langages probabilistes, qui sont les modèles de langage habituellement utilisés dans les systèmes de dictée vocale à très grand vocabulaire.

Les quantités en jeu (40 millions de mots dans le cas du corpus mis à disposition dans l'ARC B1) font que ces procédures de normalisation sont à la fois difficiles et coûteuses.

Ces sources de données contiennent de nombreuses fautes de frappe ou des textes parasites, d'où nous devons à la fois extraire un lexique et une estimation de fréquences de mots et de séquence de mots; aussi, il est nécessaire de nettoyer et de segmenter ces textes en "mots", la définition de ce qu'est un mot dépendant principalement des contraintes inposées à la fois par le système de reconnaissance et le modèle de langage. Nous devons donc segmenter le texte en satisfaisant au mieux au moins deux critères contradictoires: avoir le lexique le plus petit possible (pour diminuer la charge de travail pendant la reconnaissance) et de générer des mots non ambigus (afin d'augmenter la discrimination du modèle). Si nous utilisons uniquement le premier critère, nous transformerons toute majuscule en minuscule (pas de noms propres) et séparerons systématiquement toute ponctuation des mots (pas de mots composés). L'effet en sera une grande ambiguité syntaxique (par exemple pas de distinctions entre **Roman** (Polanski) et **roman**, et **sec.** (secondes) et **sec .**).Si nous ne prenons que le second critère, nous ne segmenterons aucun mot ni ne transformerons aucune majuscule; le lexique contiendra alors de nombreuses ambiguités lexicales: par exemple les mots **C'est** et **c'est**.

Notre position est donc de rechercher le compromis acceptable, étant donné le temps considérable que coûte toute intervention manuelle sur des textes d'une telle taille.

Ces normalisations incluent (entre autres):

- codage des accents et autres signes diacritiques: le codage est ISO-latin1
- une séparation hiérarchique en articles, paragraphes, phrases et mots, le tout en format SGML.
- élimination de symboles non conformes; traitement d'exceptions propres au texte portant sur des fautes de frappes fréquentes (du genre **4o C** au lieu de **4 oC**); prétraitement des chiffres (**10 000** → **10000**).
- traitement des unités relatives (**kg/cm3**, etc..).
- élimination d'erreur de formatage propres aux textes journalistiques; on peut citer (entre autres) :
  - élimination de lettres parasites à la fin des articles.
  - recollement de paragraphes séparés par suite d'erreurs de formatage.
  - élimination de symboles de ponctuation parasites au debut des articles (faute de formatage).

– détection d'abréviations nouvelles, correction d'erreurs de ponctuation.

– traitement des ponctuations dites "non ambiguës", dans la mesure où elles ne conduisent pas à des mots composés, et séparation en phrases. On segmente en phrase après les ponctuations fortes, et après ":", s'il est suivi de '"'. Par contre, les incises sont laissés dans la phrase (par exemple:
" Bonjour ", dit-il , " comment ca va ? " . est une phrase).

– traitement des mots composés (avec - et ') et des majuscule de début de phrase. Ce traitement se fait en 2 passes, et utilise 2 dictionnaires généralistes (BDLEX ([Pérennou G., (1988)]) et DELAF via INTEX ([Silberztein M., (1993)])), contenant des listes de mots composés. Lors de la 1ère passe, on recueille les mots de tous les textes à traiter, et ne comportant pas d'ambiguïtés (par exemple dans : **La voiture de M. Pierre Durand** , le mot **Pierre** n'est pas ambigu, et est conservé. Des poids différents (assimililables à des "vraisemblances") sont affectés aux mots suivant leur origine; un poids plus important est accordé aux mots provenant de l'article en cours de traitement. Lors de la 2ème passe, on construit les différents découpages possibles, et on choisit le meilleur en fonction d'euristiques reposant sur les listes de mots pondérés. On traite également dans cette étape, le problème des phrases ou début de phrase entièrement en majuscules (problème de formatage).

– traduction des chiffres romains en chiffres (exemple : **chapitre XII → chapitre 12**, **Francois Ier → Francois 1er**), puis des chiffres en mots (**1993 → mille neuf cent quatre-vingt-treize**, **les 24e journées de l'enseignement → les vingt-quatrièmes journées de l'enseignement**).

– éclatement des sigles non acronymes à l'exception des plus fréquents (sigles présents dans la liste des 20 000 mots les plus fréquents).

– transformation de la ponctuation en ponctuation verbalisee: (exemple : ","→ **,VIRGULE**), en vue de son élimination pour la construction des modèles de langage.

Lexiques et Modèles de Langage

Ces corpus ont permis au LIMSI de définir 2 listes de mots officielles, une constituée des 20 000 mots les plus fréquents de ces textes (dénommée "liste 20k"), et l'autre constituée des 64 000 mots les plus fréquents (dénommée "liste 64k"); des modèles de langage de types bigrammes et trigrammes avec réestimation par technique dite du "back-off", pour ces 2 listes ont été également obtenus à partir de ces textes. Les lexiques et modèles de langage ont été mis à disposition des participants en même temps que le corpus de développement.

Corpus Lexical

Une phonétisation de la liste officielle de mots pour la catégorie dite P0 (contrainte d'utilisation d'une liste de réference de 20 000 mots) a été réalisée à l'IRIT.

La liste lexicale 20k provient d'un corpus de texte et, de ce fait, comprend des éléments inattendus par rapport aux lexiques classiques, en particulier certains d'entre eux présentent des fautes d'accents, de capitalisation, . . ., d'autres sont des parties de mots composés. Il est apparu qu'il serait difficile d'obtenir un consensus au sein de B1 sur une régularisation de ces éléments de la liste 20k.

L'IRIT a proposé de résoudre cette difficulté au moyen de 2 fichiers :

– un lexique morphosyntaxique phonétisé (LexP), issu de BDLEX ([Pérennou G., (1988)]),

– une table de correspondance (T) associant à chaque mot de la liste officielle 20k une clé d'entrée dans le lexique phonétisé (LexP).

Ceci laisse une latitude à l'usager pour construire son lexique de reconnaissance.

Table de Correspondance

(T) résulte d'une jointure tolérante portant sur l'attribut graphie accentuée entre les mots de la liste 20k et les formes fléchies de BDLEX. Elle permet d'associer à chaque mot de la liste 20k une ou plusieurs formes fléchies selon un code de correspondance : égalité de graphie (=), faute d'accent (A), différence de capitalisation à l'initiale (I), différence de casse (M), tolérance au caractère point dans le cas des sigles (P), tolérance à une faute orthographique ou typographique (O), partie de mot composé (+). La table suivante illustre ces différentes possibilités.

Mot Liste 20K	Code	Clés LexP
abaissement	=	abaissement
abaissé	=	abaissé
A.	=	A. (1)
ad	+	ad_hoc
tandis	+	tandis_que, tandis_qu'
trainer	A	traîner (2)
Africain	I	africain (3)
TOUR	M	tour
judaïme	O	judaïsme
APL	P	A.P.L.

(1) Dans ce cas nous ne cherchons pas à savoir si 'A.' réfère à une lettre ou une abréviation de prénom (André, Antoine...).
(2) Les journaux ignorent souvent l'accent circonflexe ; d'où le traitement de 'trainer' non sans risque cependant car il peut très bien être une partie de 'home trainer'.
(3) LexP ne distingue pas Africain et africain comme le fait le dictionnaire.

Lexique Morphosyntaxique Phonétisé des Formes Fléchies

Les informations disponibles pour une forme fléchie de (LexP) sont : la graphie accentuée, la représentation phonologique, le fonctionnement phonologique de la finale, la catégorie syntaxique, des informations morphosyntaxiques (genre, nombre, temps, mode, . . .) et la graphie accentuée de la forme canonique dont elle est issue.

Une clé d'entrée LexP de (T) pointe sur une ou plusieurs formes fléchies homographes de (LexP). Ainsi la clé 'photographie' permet d'accéder à la forme 'photographie' (nom fém. sing.) et aux formes conjuguées du verbe 'photographier' (1ère pers. prés. ind., 3ère pers. prés. ind, . . .). A la liste 20k, correspondent 30.700 formes fléchies dans LexP.

L'usager peut accéder d'une part aux informations morphosyntaxiques et d'autre part aux prononciations des mots. Pour la représentation des prononciations, nous avons adopté les conventions utilisées dans BDLEX. En outre, on trouve une modélisation des mots étrangers : des groupes de phonèmes suscitant des variantes de prononciation sont notés entre parenthèses, par exemple "Jane" /(dZ)(Ei)n/, "charter" /(tS)aRt(6R)/.

Corpus de Développement et d'Évaluation

La même procédure a été utilisée pour la constitution du corpus de développement et la constitution du corpus d'évaluation.

Corpus Textuel

Le volume de texte original représente 50 articles du Monde, soit environ 900 phrases.

Le but de ce texte était de sélectionner et saisir un échantillon aléatoire de 50 séquences de 15 phrases consécutives.

Le choix du texte a été fait de manière à ce que les séquences sont facilement "lisibles" donc, en particulier :

- être compréhensibles hors contexte : pour cette raison nous avons choisi de sélectionner systématiquement des débuts d'articles de journal (pour lesquels il y a de bonnes chances que les références au contexte – liens anaphoriques, déictiques, . . .– soient limitées);
- contenir une proportion majoritaire de texte "standard" (à ce titre, on a éliminé de la sélection les cours de la bourse, les résultats sportifs, les listes de nominations, etc . . .);
- ne pas contenir une proportion trop importante d'extraits dans une langue étrangère (chansons, romans, etc . . .);
- ne pas contenir de passages pouvant choquer le lecteur (et éventuellement provoquer un refus de lecture); à ce titre ne sont pas retenus des articles contenant des prises de position politiques ou morales trop marquées)

Les articles ont tous été sélectionnés dans les éditions quotidiennes du journal "Le Monde" publiées sur une période de 15 jours, en mai pour le corpus de développement et en novembre pour le corpus d'évaluation.

La méthode effective de sélection des séquences a été la suivante :

- on estime le nombre N d'articles dans les 12 editions du Monde concernées;
- on génère une liste aléatoire de 0 et de 1 avec une probabilité de $\frac{50}{N}$ pour 1 (et donc $1 - \frac{50}{N}$ pour 0).
- en démarrant à partir du premier article de la première édition, on ne considère que les articles correspondant à un 1 dans l'ordre de la séquence aléatoire et respectant les conditions suivantes :
  - l'article contient au moins 15 phrases;
  - les 20 premières phrases (arrondies au paragraphe superieur) sont lisibles (au sens défini ci-dessus);
  - l'article n'a pas déjà été retenu (ce qui permet de parcourir plusieurs fois la séquence des articles si le premier parcours n'a pas fourni 50 candidats valides)

  L'intérêt de cette méthode est que le tri se fait avant saisie et que l'on réduit de ce fait la quantité de texte saisie pour rien.
- pour chaque article sélectionné, on retient les 15 premières phrases en arrondissant au paragraphe supérieur.

Une deuxième sélection est faite pour redescendre aux environs de 600 phrases (taille du corpus de test dans la catégorie la plus élevée pour l'action ILOR_B1) sur la base du nombre de mots par phrase (n): les phrases trop courtes ($n < 2$) ou trop longues ($n > 65$) sont retirées et avec elles tout l'article dont elles font partie de façon à préserver la cohérence interne des articles conservés. En effet il a été décidé de conserver l'unité de structure d'article au niveau de l'enregistrement : ainsi chaque locuteur prononce un ou plusieurs articles en entier, ce qui est garant d'un meilleur naturel et propage les effets de sens. On obtient ainsi un corpus de 609 phrases (33 articles) pour le développement et de 655 phrases (38 articles) pour le test.

On détermine ensuite pour chaque corpus le sous-ensem -ble dit "de 300 phrases" qui est la taille de corpus retenue pour le test en première catégorie (correspondant aux systèmes qui ont des capacités de traitement et un vocabulaire plus limités). La détermination de "l'ensemble des 300" se fait sur la base du taux de mots hors vocabulaire. Le texte est donc nettoyé de ses balises SGML, soumis à un ensemble de pré-traitements (ceux ayant présidé à la constitution de la liste 20K de référence) et le taux de mots hors vocabulaire (MHV) (par rapport à la dite liste 20K) est calculé pour chaque phrase, puis sommé au niveau de chaque paragraphe, ceux-ci étant ensuite classés. En effet,

le principe de cohérence qui vise à maintenir une certaine continuité sémantique est ici appliqué au niveau du paragraphe. Ce sont donc les n premiers paragraphes ayant le taux de MHV le plus bas qui seront sélectionnés, à concurrence de 300 phrases. Par contre il n'y a pas de continuité entre les paragraphes.

Corpus Oral

Pour de l'enregistrement, on s'assure de la répartition équilibrée des articles sur l'ensemble des 20 locuteurs retenus. Les locuteurs ont été recrutés par petite annonce, et rémunérés. Ils ont été choisis de manière à obtenir une bonne répartition par age et sexe.

Chaque locuteur prononce 3 articles entiers: un article présentant globalement un faible taux de MHV, un article à fort taux de MHV, et un article tiré au sort. Un tableau de marche est ainsi fourni, qui spécifie qui prononcera quoi, et assure également que tous les articles seront à peu près équitablement représentés.

L'enregistrement proprement dit a été effectué à l'aide d'un micro Shure SM10, et échantillonné à 16 kHz. Le locuteur se voit présenté un "prompt" orthographique, lit chaque phrase dans l'ordre original de l'article, et doit répéter la phrase lorsque la personne chargée de l'enregistrement détecte une erreur (répétition, oubli, etc, . . .) . Les fichiers produits sont au format NIST (entête SPHERE) pour le signal, et un fichier de description donne pour chaque locuteur les prompts orthographiques des phrases qu'il a enregistré, accompagnés d'un identificateur de phrase qui est une clé unique.

Respectivement 1050 et 1042 fichiers de signal acoustique ont donc été produits pour les corpus de développement et de test, puisque chacun des 20 locuteurs a prononcé 3 articles parmi la trentaine conservée parmi les cinquante saisis. L'ensemble des fichiers a été organisé à chaque fois sur CDROM avec une documentation en rapport (ILORB1_DEV1 et ILORB1_TST1).

La spécification finale du Corpus 300 et du Corpus 600 pour les différentes catégories de test, consiste à identifier nommément chacun des fichiers acoustique qui va faire partie du test. En effet les listes de phrases 300 et 600 ne préjugent pas des locuteurs qui vont effectivement les prononcer, par contre les systèmes de reconnaissance sont sensibles à la variabilité inter- et intra-locuteurs. La liste définitive de 300 fichiers qui spécifie le test 300 est donc une liste qui satisfait deux contraintes: que les 300 phrases retenues y soient présentes, et que l'ensemble des locuteurs y soit représenté autant que possible de façon équilibrée (répartition équitable de locuteurs masculins et féminins), le tout sur la base d'une structure de paragraphe. C'est-à-dire que chaque fois qu'un locuteur est retenu pour avoir prononcé une phrase, c'est l'ensemble des phrases appartenant au même paragraphe qui est retenu pour ce locuteur. Le Corpus 600 étant un sur-ensemble du 300, a été défini par la même méthode en se basant cette fois sur une unité d'article entier prononcé par un même locuteur.

Conclusion

Nous avons mis à disposition dans le cadre de l'ARC B1 "Dictée Vocale", un certain nombre de ressources :

- 2 corpus de parole :
  - BREF, representant la prononciation de plus de 100 heures de parole par 120 locuteurs;
  - le sous-corpus BREF-80, représentant la prononciation de 67 phrases par 80 locuteurs;
- un corpus de texte journalistique de 40M de mots, balisé et normalisé.
- 2 listes de mots de 20 000 (20k) et 64 000 mots (64k), ainsi que les modèles de langage de type bigramme et trigramme pour ces 2 listes;
- une phonétisation de la liste 20k;
- un corpus oral et textuel de développement et d'évaluation, afin de tester les systèmes de dictée avec et sans contrainte sur le taux de mots hors vocabulaire.

L'utilisation du paradigme de l'évaluation, à travers une diffusion plus large de ces ressources uniques pour la langue française, peuvent permettre de faire avancer de manière significative la recherche en reconnaissance de la parole en langue française.

Remerciements

Une partie de ce travail a été effectué dans le cadre des Actions de Recherche Concertées "Linguistique, Informatique et Corpus Oraux", financées par l'AUPELF-UREF.

Références

[Gauvain J.-L., Lamel L., Eskénazi M., (1990)] , "*Design considerations & text selection for BREF, a large french read-speech corpus*", Proceedings ICSLP-90, Kobe, Japan, Nov. 1990.

[Gauvain J.-L., Lamel L., Adda G., Mariani J., (1994)] "*Speech-to-Text conversion in French*", Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 8 no 1 (1994) 99-131.

[Lamel L., Gauvain J.-L., Eskénazi M., (1991)] "*BREF, a Large Vocabulary Spoken Corpus for French*", Proceedings EUROSPEECH-91, Gênes, Italie, septembre 1991.

[Pérennou G., (1988)] "*Le projet BDLEX de base de données lexicales et phonologiques*", Actes des 1ères journées du GRECO-PRC CHM, EC2 éd., Paris, 24-25 novembre 1988.

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