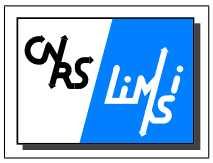
Approches empiriques et modélisation statistique de la parole

Adda Gilles



Mémoire d'Habilitation à Diriger des Recherches

Approches empiriques et modélisation statistique de la parole

Gilles ADDA

Soutenue le 14 novembre 2011, avec le Jury :

Pierre Zweigenbaum	Président	Directeur de Recherche CNRS	LIMSI/ILES (Orsay)
Régine André-Obrecht	rapporteur	Professeure	Université Paul Sabatier IRIT (Toulouse)
Mark Liberman	rapporteur	Professeur	University of Pennsylvania (Philadelphia, USA)
Holger Schwenk	rapporteur	Professeur	Université du Maine LIUM (Le Mans)
Jean-Luc Gauvain	parrain	Directeur de Recherche CNRS	LIMSI/TLP (Orsay)
Joseph Mariani	examinateur	Directeur de Recherche CNRS	LIMSI/TLP et IMMI (Orsay)
Jean-Paul Haton	examinateur	Professeur émérite	Université Henri Poincaré LORIA (Nancy)
Edouard Geoffrois	invité	Ingénieur en chef de l'armement	DGA (Montrouge)

Table des matières

Introduction Générale

Une Habilitation à Diriger des Recherches est souvent un catalogue des travaux passés, ce qui est logique étant donnée la fonction officielle de ce diplôme :

> *L'habilitation à diriger des recherches sanctionne la reconnaissance du haut niveau scientifique du candidat, du caractère original de sa démarche dans un domaine de la science, de son aptitude à maîtriser une stratégie de recherche dans un domaine scientifique ou technologique suffisamment large et de sa capacité à encadrer de jeunes chercheurs.*

Pour certaines universités, un curriculum vitae étendu avec un tiré à part de quelques articles suffit à démontrer ces différentes capacités ; dans d'autres, il faut écrire une nouvelle thèse, renouant ainsi avec la tradition de la « thèse d'état »que l'habilitation a remplacée. En suivant mon penchant naturel, je n'aurais pas eu l'impudence d'essayer d'écrire un document spécifique ayant pour but de prouver que j'ai bien toutes ses qualités : j'aurais largement préféré laissé les rapporteurs en décider à la seule lecture de mon CV plutôt que de leur imposer une lecture forcément plus longue et fatalement fastidieuse (par moments). Mais l'usage local en Informatique à l'Université Paris-Sud a adopté une position intermédiaire : un document « original »mais de taille réduite, d'où le document présent.

J'ai choisi de faire un document en deux parties.

La première partie, intitulée pompeusement « Un parcours en modélisation statistique du langage et son application aux systèmes multilingues de traitement de la langue » relatera succinctement mes travaux de recherches, en une présentation diachronique selon quelques grandes rubriques. Une présentation à la fois succincte et diachronique, étant donné que j'écris ce document après 28 ans de recherches, me paraissait à la fois plus charitable pour les lecteurs et peut-être plus intéressante car apportant un éclairage historique sur les avancées en traitement de la parole.

La deuxième partie se rapportera plus spécifiquement au titre du document « Approches empiriques et modélisation statistique de la parole ». Ce titre a le bon goût de pouvoir chapeauter l'ensemble de mes travaux qui sont de nature fondamentalement expérimentale, dans le cadre de la modélisation statistique de la parole et de ses applications, mais également la deuxième partie intitulée « Développement d'une science empirique en traitement de la parole ». En effet, j'ai choisi de succomber au péché d'orgueil et d'aborder un problème non strictement technique ou scientifique qui m'intéresse

depuis quelques années, le statut épistémologique des sciences du langage, et en particulier de l'étude de la parole : quel est le statut de la connaissance que nous produisons, comment la qualifier par rapport à d'autres sciences (sciences dites « dures » comme la physique et le traitement du signal ou « molles » comme la linguistique) ? est-il possible d'autonomiser les sciences du langage en une véritable science, en essayant de trouver à la fois quel est son observable et le moyen d'améliorer la manière de l'observer, et d'en tirer des connaissances généralisables ? J'ai parlé de péché d'orgueil car ni le statut scientifique ni la compétence épistémologique ne me permettent de parler de ces sujets en usant de l'argument d'autorité. Mon but ici est d'apporter mon verre d'eau au ruisseau d'un mouvement que l'on voit sourdre dans les différentes communautés des sciences du langage, et en particulier dans un certain nombre de disciplines dont l'objet est l'étude de la parole, à partir du travail de pionnier de quelques personnes. Ce mouvement tend à considérer que nous sommes arrivés à une maturité des instruments automatiques à notre disposition qui peut nous permettre de passer un cap scientifique, et de rapprocher la communauté du traitement automatique et les communautés des sciences humaines afin qu'elles s'enrichissent mutuellement, voire qu'elles collaborent véritablement autour des mêmes objets et des mêmes instruments. J'espère en particulier que ce document permettra un débat sur certaines propositions concrètes exprimées à la fin du document en ce qui concerne la structuration de la production scientifique et le développement de centres instrumentaux sur le modèle du CERN.

Première partie

Un parcours en modélisation statistique du langage et son application aux systèmes multilingues de traitement de la langue

Chapitre 1

Introduction

Mes travaux, quoique ayant principalement porté sur la modélisation du langage, a de fait touché à beaucoup de composantes d'un système de transcription ; de plus j'ai participé au développement de nombreuses applications, industrielles ou non, ainsi qu'à de nombreuses évaluations, du côté des participants, mais également du côté de l'organisation. Structurer ces travaux est pour moi une gageure, j'ai donc choisi de les présenter en thèmes, et dans chaque thème de les présenter dans une perspective historique, ce que me permettent mes vingt-huit années de recherche en traitement automatique de la parole. J'ai également ajouté à chaque thème les citations de ma bibliographie personnelle se référant à ce thème.

Une présentation diachronique de ces travaux permettra de remonter par exemple aux prémisses de l'utilisation de n-grammes [Debili77] ; les corpus étaient alors très rares, en particulier pour la langue française, et les premiers développements les utilisant paraîtront bien élémentaires au vu des tailles des corpus et des modèles de langage que l'on développe de nos jours. Ensuite, les systèmes de transcription ont gagné en puissance et en précision, et se sont développés à la fois en plusieurs langues, mais également vers de nouvelles applications passant de la dictée vocale à la transcription de conversations ou de réunions en passant par la transcription d'émissions radio-télédiffusées. La modélisation du langage s'est alors développée largement, et le zoo des modèles de langage s'est peuplé de modèles de plus en plus compliqués, (modèle cache, modèles triggers, ...) et également bénéficiant de méthodes de lissage et d'adaptation de plus en plus pertinentes.

Nous explorerons dans cette première partie la préhistoire de la modélisation aux systèmes actuels, et aborderons les différents travaux que j'ai

menés, en normalisation et en modélisation à partir de classes au début, puis au passage aux n-grammes de mots en même temps que nos premières participations aux évaluations ARPA. Ensuite viendra l'extension vers d'autres langues et d'autres applications, ce qui passera vers le développement de méthodes d'adaptations et en particulier d'adaptation non ou faiblement supervisée, mais également de développement de lexiques de prononciation.

Enfin je mettrai en perspective les travaux que j'ai pu mener en recherche, avec les différents systèmes qui ont été développés mais également quelques applications industrielles dans lesquelles ces travaux ont été intégrés. Les systèmes de traitement automatique de la parole ont été utilisés récemment afin d'acquérir plus de connaissances en linguistique et en particulier en phonétique ; je montrerai un aperçu de ces expériences qui se situent dans le cadre plus général (abordé dans la seconde partie du document) de l'utilisation des outils de traitement automatique comme instruments pour explorer les corpus afin d'en extraire de la connaissance.

Je présenterai enfin les travaux que j'ai menés dans le cadre de l'évaluation comparative, que ce soit par le développement de corpus, la mise en place d'évaluations mais aussi les nombreuses participations à des évaluations.

Le but de cette première partie n'est pas d'être exhaustif en présentant en détail l'ensemble de mes travaux en traitement de la parole durant la période 1983-2011 ; cette revue, qui restera (je pense définitivement) à écrire, dépasse les limites, en temps et en place, d'une HDR. Le but est ici de décrire succinctement, en les mettant dans une perspective historique, les différents travaux que j'ai menés pendant cette période.

Chapitre 2

Modélisation statistique du langage et systèmes de transcription

2.1 Introduction

Pour la plupart des lecteurs, ce qui va suivre dans cette introduction va être une évidence maintes fois répétée dans toutes les thèses utilisant une modélisation statistique du langage ; pour ceux-là, je les invite à sauter cette section introductive. Pour les autres, j'ai bien peur que cette courte introduction ne soit pas suffisante pour saisir les implications théoriques et pratiques que sous-tend la modélisation probabiliste de la parole [Jelinek76, Baker75], fondée sur la théorie de l'information. Que les lecteurs intéressés se reportent plutôt à (par exemple) [Mariani09, Jelinek97, Rabiner93] pour avoir une vue plus précise et complète. Le but de cette introduction est fixer certains termes pour des lecteurs candides en ce domaine.

Dans la modélisation statistique de la parole, il s'agit de déterminer la meilleure suite de mots $\hat{m}$ étant donnée l'observation acoustique x, ce qui peut s'écrire, en appliquant la formule de Bayes :

$$\hat{m} = \underset{m}{argmax} P(m/x) = \underset{m}{argmax} P\left(x/m\right) P\left(m\right)$$

Le décodeur, c'est-à-dire l'opérateur $argmax$, reposant sur le principe de programmation dynamique [Bellman57], doit évaluer la probabilité de toutes les suites de mots m possibles pour ce signal, $P\left(m/x\right)$. Grâce à la for-

mule de Bayes, le problème se transforme en une optimisation à deux termes $P(x/m)\,P(m)$ pour lesquels des modèles peuvent être estimés à partir d'un corpus d'apprentissage. Le premier terme $P(x/m)$ permet d'évaluer la vraisemblance d'observer le signal acoustique x, en faisant l'hypothèse que la suite de mots m a été prononcée. Cette probabilité est estimée grâce aux modèles acoustiques et au modèle lexical.

Le deuxième terme $P(m)$ permet d'estimer la probabilité a priori de la séquence de mots m, grâce au modèle linguistique. La modélisation du langage est une brique fondamentale de la modélisation statistique de la parole, et elle constitue le cœur des travaux que j'ai menés.

Un troisième terme peut se rajouter pour tenir compte des variantes de prononciation Φ détaillées dans un dictionnaire de prononciation :

$$\hat{m} = \underset{m}{argmax} P(x/\Phi)\,P(\Phi/m)\,P(m)$$

On peut représenter le processus statistique de reconnaissance de la parole par un schéma (voir figure 2.1), qui visualise également les processus conduisant à l'élaboration des modèles ; nous pouvons voir que les processus de modélisation ne sont pas déconnectés : le corpus de parole qui servira à estimer les modèles acoustiques participera également (à travers sa transcription) à l'estimation du modèle linguistique, mais également (après alignement entre les variantes de prononciation et le signal) à l'estimation des probabilités de prononciation dans le modèle lexical. On peut voir également le rôle central et primordial de la normalisation : elle agira non seulement sur le modèle de langage, mais également sur le modèle acoustique et sur le modèle lexical.

La modélisation linguistique permet d'inclure dans un système les contraintes résultant des régularités du langage [Rosenfeld00]. La modélisation la plus courante dans les systèmes actuels est dite n-gramme, car elle repose sur des statistiques de n unités consécutives qui capturent certaines contraintes syntaxiques et sémantiques du langage. En écrivant la probabilité d'une séquence de mots $(m_1, m_2, \cdots, m_k)$ comme :

$$P(m_1, m_2, \cdots, m_k) = P(m_1) \times \prod_{i=2}^{k} P(m_i | m_1, \cdots, m_{i-1})$$

En utilisant l'hypothèse markovienne que seuls les $n-1$ mots précédents apportent une information, on peut réduire l'historique du mot m_i aux $n-1$

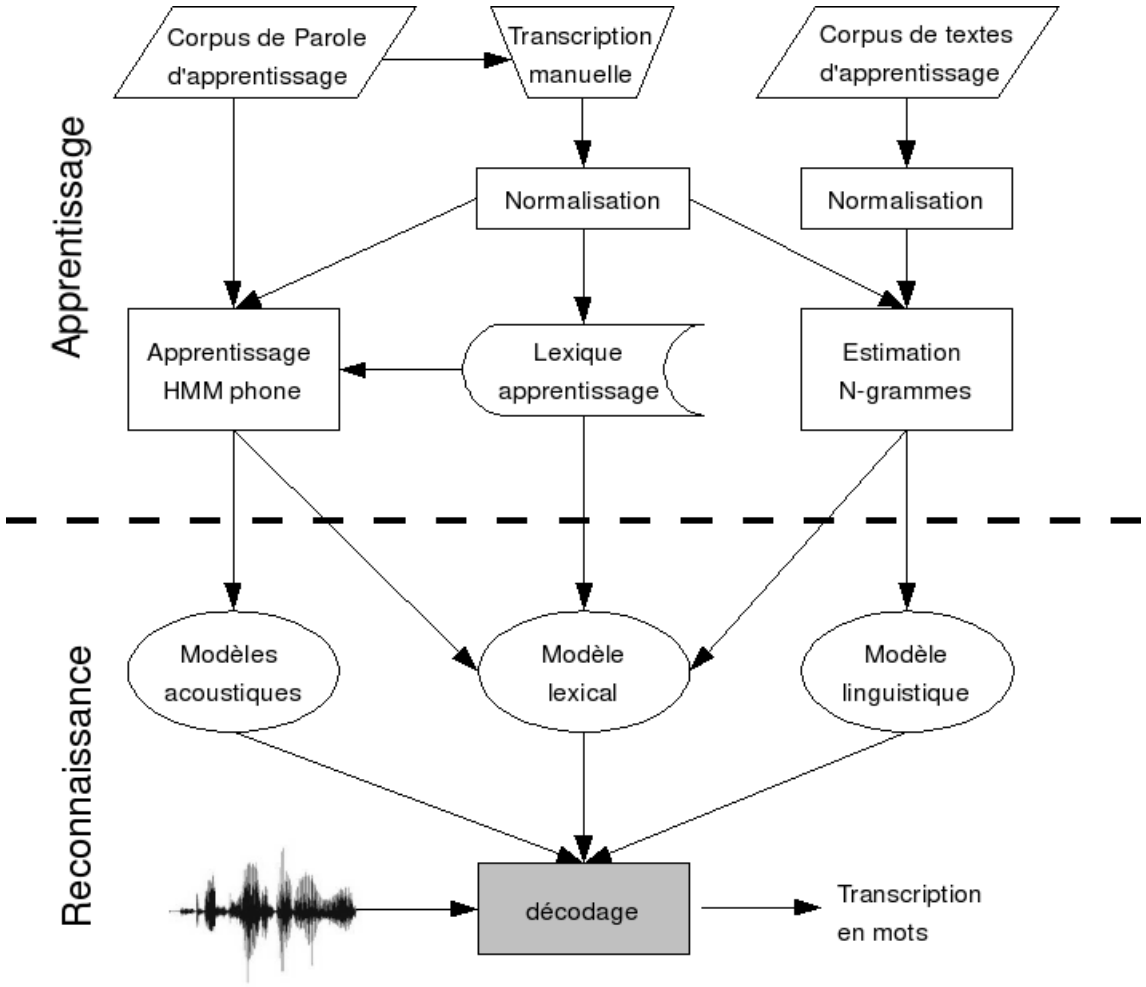


FIGURE 2.1 – Schéma du processus de décodage de la parole (**Reconnaissance**), montrant les différents modèles mis en œuvre, ainsi que leur élaboration (**Apprentissage**)

mots précédents. Cette probabilité peut alors être approchée par :

$$P(m_1, m_2, \cdots, m_k) \approx P(m_1) \times \prod_{i=2}^{k} P(m_i | m_{i-n+1}, \cdots, m_{i-1})$$

Dans les systèmes de l'état de l'art, sont utilisées des portées de 2 à 4 mots. L'approximation n-gramme réduit considérablement les données à collecter pour évaluer la probabilité d'une séquence quelconque de mots [Brown92, Rosenfeld94].

Lorsque les données d'apprentissage sont insuffisantes, les statistiques des n-grammes peu fréquents (ou non observés) sont « lissées », en utilisant un mécanisme de redistribution et de repli vers des statistiques d'ordre inférieur [Katz87]. Dans les différentes techniques de lissage, l'idée de base est de réserver une certaine quantité de probabilité provenant des estimations de fréquence relative des événements vus, pour la redistribuer aux événements non vus. Les techniques de lissage diffèrent selon la façon dont on prélève la probabilité réservée et sa quantité, et la façon de la redistribuer. On peut trouver dans [Chen98] une description des techniques classiques de lissage.

En général, les modèles de langage pour la reconnaissance de la parole sont évalués en fonction de leur impact sur le taux d'erreurs de mots lors de la reconnaissance. On utilise cependant lors de leur développement, comme mesure intermédiaire, leur capacité de prédiction des mots d'un texte, en particulier la *perplexité* [Jelinek90] ; elle est généralement un bon indicateur de la qualité du modèle, mais sa corrélation avec le taux de reconnaissance n'est pas avérée car d'autres facteurs influant sur la qualité de la reconnaissance, comme par exemple la similarité acoustique des mots, ne sont pas pris en compte.

La perplexité d'un modèle de langage, calculée sur un corpus est définie par $PP = P'\left(C_1^L\right)^{-\frac{1}{L}}$, où $P'(C_1^L)$ représente la probabilité estimée à l'aide du modèle de langage sur un corpus C_1^L contenant L mots.

2.2 Normalisation

La normalisation est une étape très importante de la modélisation du langage, souvent négligée car elle implique des processus de nettoyage de textes qui sont souvent considérés comme du bricolage (et qui le sont dans la plupart des cas). Nous essayerons ici de montrer comment inclure la normalisation

dans le processus de reconnaissance. Normaliser un texte [Adda00C] c'est définir explicitement ou implicitement l'unité retenue dans la modélisation du langage [1], puis de conditionner les textes afin d'estimer au mieux les distributions de ces unités dans les textes. Cette étape doit permettre à la fois la réduction de la variation lexicale (et donc l'augmentation de la couverture lexicale pour une même taille de lexique, voir section 2.4) et la conservation d'une précision suffisante au modèle de langage. Ces deux exigences peuvent être contradictoires.

Si nous ne suivons que la voie de la réduction de la variation lexicale, la stratégie retenue sera donc par exemple de ne pas admettre de distinction de casse (pas de distinction majuscule-minuscule), de segmenter au maximum selon toute marque non alphabétique (pas de mots composés) ; le résultat sera une ambiguïté syntaxique (par exemple pas de distinction entre *Roman* (Polanski) et un *roman*, pas de distinction entre *sec.* (abréviation de « secondes ») et au *sec* .. Si nous suivons la voie de la réduction de l'ambiguïté syntaxique, nous garderons toutes les formes différentes, sans segmentation, voire nous amalgamerons entre elles certaines formes séparées (locutions) en ce qu'elles ont un comportement atypique. Le résultat sera la présence d'ambiguïtés lexicales, comme par exemple la présence dans le lexique des deux formes *C'est* et *c'est*.

Il s'agit donc de faire un compromis entre les deux exigences, tout en prenant les stratégies qui sont efficaces pour l'une sans influer sur l'autre.

Des études nous ont montré l'importance de la normalisation, à la fois en terme de couverture du vocabulaire de reconnaissance et de perplexité du modèle de langage ; ces études [Adda97B, Adda97D, AddaDecker98A] ont permis de nous guider de manière explicite dans le choix de la définition du mot pour le français, afin d'avoir à la fois une bonne couverture lexicale et une bonne précision au niveau des modèles de langage.

Toute modélisation du langage implique l'utilisation de textes (textes écrits ou transcriptions de l'oral) et l'emploi d'une normalisation, c'est-à-dire donc une définition de ce qu'est un mot, et une stratégie de compromis entre les deux exigences évoquées précédemment. Lorsque le choix est implicite (parce que le choix de la normalisation s'est fait a priori, ou parce les textes ont été normalisé par un outil d'une provenance quelconque), il conduit à des résultats sous-optimaux, que ce soit en reconnaissance [Adda00C] ou en traduction [Dechelotte07A]. Si cette normalisation suit les indications telles

1. c'est-à-dire répondre à la question *qu'est-ce qu'un mot ?*

que trouvées dans les articles de référence, l'impact d'une normalisation plus « optimale » est réduit ; mais au-delà d'une certaine qualité de reconnaissance, il est des cas où la normalisation doit se coupler intimement avec les tentatives de meilleure modélisation ; par exemple, si l'on désire pouvoir ajouter des variantes de prononciations réduites[AddaDecker00B, Lamel96A] dans le cas de parole spontanée, il est nécessaire que la définition des mots soit compatible avec ce choix ; c'est pourquoi nous avons choisi de définir des mots composés permettant ces réductions, par exemple *il_y_a* en français ou *a_little_bit* en anglais.

Bibliographie :[Adda87B] [Adda97D] [Adda97B] [AddaDecker98A] [Habert-et-al98] [Adda00B] [Adda00C] [AddaDecker00B] [AddaDecker00A] [Rosset07B] [Dechelotte07A]

2.3 Modèles de langage

2.3.1 modélisation n-gramme, n-grammes de mots

Paradoxalement, la modélisation classique n-gramme de mots, qui est pourtant la brique essentielle, incontournable d'un système de transcription, n'est pas celle qui génère le plus de littérature. Les techniques sont bien éprouvées, et quelques articles de référence permettent de choisir la méthode de lissage, d'optimisation, appropriée à la tâche. Mais cette « évidence » des modèles n-gramme ne s'est mise en place que progressivement, et seulement au cours des années 90.

Dès le début de mes travaux en collaboration avec mes collègues du LIMSI (C. Fluhr, A. Andreewsky, J. Mariani, F. Néel), nous avons envisagé de traiter le problème de la reconnaissance de la parole avec une modélisation linguistique utilisant une modélisation n-gramme[Adda84B, Bellilty85], ici sous sa déclinaison « matrice de précédence » [Andreewsky78, Debili77]. Nous avons pris comme sous-problème spécifique la transcription d'une chaîne ou d'un treillis phonétique en phrase orthographiée, en utilisant le problème spécifique de la traduction sténotype-graphème [Adda84A] (voir section 3.4). Les circonstances ont fait que cette modélisation linguistique de grands vocabulaires ouverts (au sens où il est possible et prévu qu'un mot inconnu du vocabulaire puisse être présent dans l'application) a été mise en sommeil, les applications s'étant orientées temporairement vers des vocabulaire réduits (de taille inférieure à 1000 mots), et des syntaxes contraintes (comme le

DARPA Resource Management [Price88, Lamel92]), ou de la reconnaissance de mots isolés [Nait92]. L'utilisation d'une modélisation n-gramme dans le cadre d'un système à grand vocabulaire est réapparu lors de la participation du LIMSI aux évaluations DARPA sur les données *Wall Street Journal*, avec un vocabulaire de 5k et 20k mots et un bigramme appris avec un texte d'apprentissage de 33M de mots [Gauvain93A, Gauvain93B], puis 37M [Gauvain94F], puis étendu au français dans des conditions identiques (lecture du journal *Le Monde*, avec utilisation du corpus BREF [Lamel91]) avec un texte d'apprentissage de 4M de mots [Gauvain93B, Gauvain93C], puis de 38M de mots [Gauvain94A].

A la suite de ces premières modélisations, la taille des textes d'apprentissage a augmenté drastiquement puisque pour la première évaluation sur les données radiophoniques (HUB4) en anglais [Gauvain97C], nous disposions de 300 millions de mots, et de 200 millions de mots pour la première évaluation de dictée vocale en français [Adda97D]. Au cours de l'évaluation HUB4 a été introduite une originalité dans la fabrication des modèles de langage, consistant à modifier les textes d'apprentissage afin qu'ils ressemblent plus aux transcriptions détaillées de la tâche [Gauvain97C, Gauvain98A, Adda99A].

Au cours des années, la taille des textes a continué à augmenter et à se diversifier : en 1998, pour les évaluations HUB4, 540M de mots ont été utilisés en provenance de différentes sources. La constitution des modèles se fait alors par interpolation linéaire de modèles individuels appris sur chaque source. Les coefficients d'interpolation sont appris grâce à l'algorithme EM *(Expectation-Maximization)* [Dempster77], en optimisant la perplexité sur un corpus de développement. Cependant, cela oblige à choisir, la plupart du temps de manière heuristique, une séparation des textes (en fonction de la provenance, de la date) qui soit judicieuse : une segmentation trop fine génère des modèles individuels appris sur trop peu de données, et une segmentation trop grossière ne permet pas de coller au mieux aux données de la tâche [Adda99A]. Autre particularité, ces modèles interpolés sont mergés en un seul modèle, plus facile à intégrer au système de reconnaissance [Gauvain99A, Adda99A]. Ce mode de fonctionnement a perduré jusqu'à maintenant.

Parmi les nouveautés intégrées dans les modèles de langage n-gramme du LIMSI, celle qui a apporté les gains les plus stables et les plus conséquents ces dernières années est l'introduction des modèles de langage neuronaux, sous la forme des « Continuous space language models » [Schwenk07]. L'idée de base consiste à traiter la tâche d'estimation des n-grammes dans un espace continu. Pour cela, chaque mot est projeté dans un espace vectoriel de dimension 40

à 100, la probabilité d'un mot en contexte étant ensuite estimée à partir de cette représentation.

Bibliographie :[Adda84A] [Adda84B] [Bellilty85] [Adda87B] [Gauvain93B] [Gauvain93C] [Gauvain94A] [Gauvain94F] [Adda99A]

2.3.2 n-grammes de classes

Au début des travaux sur la modélisation du langage, et en particulier pour le français, les textes étaient rares. Aussi, les différentes équipes françaises utilisant à l'époque une modélisation statistique[Bellilty85, Derouault85] se sont tournées vers une utilisation de classes de mots (ou de parties du discours) permettant d'apprendre avec plus de fiabilité des modèles n-grammes (ou encore les *matrices de précédence* selon la terminologie utilisée par [Andreewsky78, Debili77] ou [Adda87B]). Par exemple dans ma thèse [Adda87B], étaient utilisées des matrices de précédence binaire ou ternaires de 180 catégories syntaxiques, correspondant donc à des 2 et 3-grammes de classes, apprises sur uniquement 35k mots de texte, ce qui s'est avéré trop peu ; à l'époque je le rappelle, nous ne disposions pas de textes, qui donc étaient récupérés tant bien que mal, voire tapés à la main. Dans la thèse de Anne-Marie Derouault [Derouault85] (et dans l'ensemble des travaux menés par l'équipe d'IBM France à l'époque), était utilisé un ensemble de 92 parties du discours et un modèle triclasse avec back-off sur un modèle biclasse ; ce modèle était appris sur 50k mots étiquetés manuellement, complété par un apprentissage sur 1,2M de mots étiquetés automatiquement. Ces modèles n-classes étaient utilisés pour la transcription phonème-graphème, cette tâche étant considéré comme une sous-tâche de la transcription de la parole avec une transcription phonétique quasi-parfaite.

Peu de temps après, nous avons fait la liaison avec l'acoustique, dans le cadre d'une petite application de dictée vocale en mots isolés[Gauvain94E], où le texte utilisé était un livre d'apprentissage du français aux étrangers (« Le mot et l'idée ») contenant 40k mots entièrement entrés manuellement par Jean-Luc Gauvain et moi-même [2]. Le modèle de langage utilisé était également un bigramme de 160 classes, appris sur le texte en question, et avait conduit à un taux d'erreurs de 7,5%.

Cette approche utilisant uniquement des n-grammes de classes, a ensuite

2. Cette anecdote pour souligner le manque de ressources à l'époque ; dans le même temps nous avions entré manuellement le dictionnaire Julliand [Julliand70]

évolué vers l'utilisation conjointe de n-grammes de mots et de classes déterminées automatiquement [Adda99A]. Ces classes automatiques [Jardino93A, Jardino94A, Jardino94B] étaient déterminées à l'aide d'un algorithme utilisant du recuit simulé, en utilisant comme mesure la perplexité. Nous avions pu disposer de textes de plus grande taille, puisque nous avons utilisé environ 2M de mots en provenance de journal « Le Monde », ce qui permettait, en utilisant plus de 1000 classes, de baisser la perplexité sur un test. Cette méthode a donc ensuite été utilisée en reconnaissance [Adda99A] pour la langue anglaise, en utilisant un corpus de bien plus grande ampleur ($> 500M$ de mots) avec un gain relatif de 1 à 4% relatifs. Les classes automatiques ont eu tendance ensuite à s'imposer, mais aucune des deux approches n'est pleinement satisfaisante. Dans l'utilisation de classes déterminées à partir de catégories morphosyntaxiques, on introduit un a priori, la détermination des classes, qui est le plus souvent sous-optimale par rapport à un corpus et une tâche précise. Dans les classes automatiques, on a deux défauts :

- un mot n'appartient qu'à une seule classe, ce que l'on sait être sous-optimal, un grand nombre de mots, en particulier les mots grammaticaux, appartenant à plusieurs classes bien distinctes (par exemple **le** qui est à la fois déterminant et pronom) ;
- la méthode utilisée pour classer des mots repose toujours sur une mesure fondée sur l'entropie ou la perplexité, et tend à optimiser cette mesure. La conséquence est que les mots de faibles occurrences, en particulier les singletons, ont très peu d'influence sur cette mesure, ce qui conduit à leur mauvais classement ; et bien sûr en ce qui concerne les mots n'étant pas apparu dans le corpus d'apprentissage, on est réduit la plupart du temps à les mettre dans une même classe « poubelle ».

Pour conclure, l'utilisation de n-grammes de classes, qu'elles soient déterminées automatiquement ou issues de parties du discours fondées linguistiquement, semblent apporter un bénéfice réduit ; ce bénéfice peut venir de la structure de la langue, comme dans le cas du français et la présence d'homophones en genre et nombre mais partageant la même partie du discours (par exemple parti, parties, partis, parties), ou dans le cas où la quantité de données est insuffisante, au moins pour couvrir certains phénomènes (comme par exemple pour représenter les mots inconnus d'un certain type). Le bénéfice vient également d'un lissage légèrement différent, qui semble, dans certain cas, être complémentaire des lissages plus classiques introduits dans les n-grammes.

Bibliographie :[Adda87B] [Jardino93A] [Jardino93B] [Jardino94A] [Jardino94B]

[Adda99A]

2.3.3 Adaptation

Les modèles n-grammes de mots sont des modèles statiques : ils sont appris avant toute reconnaissance, à partir de textes et de transcriptions préexistants. Très rapidement, des études ont essayé d'améliorer ces modèles en les adaptant [Bellegarda04], soit en utilisant une faible quantité de données préexistantes, soit en utilisant la sortie de reconnaissance pour adapter au contenu lexical ou au contenu thématique.

C'est dans cette dernière catégorie que nous avons commencé nos travaux en adaptation des modèles de langage. Nous nous sommes placés dans le cadre où les données d'adaptation sont extraites des données utilisées pour apprendre le modèle n-gramme ; il s'agit donc ici non pas d'apprendre de nouvelles connaissances, mais de pondérer différemment celles-ci en fonction de ce que l'on cherche à reconnaître. A cette fin, nous avons mené une première expérience sur le Mandarin [Chen00] qui a conduit à de premiers gains. Nous avons commencé par effectuer un segmentation en « histoires » de longueur inférieure à 3000 mots, puis affecté à chaque histoire un ou plusieurs thèmes, en fonction d'une liste de mots-clés (3000) et d'une liste de thèmes (8 catégories générales). Ensuite, un modèle de langage est entraîné pour chaque thème, et les coefficients du modèle interpolé sont appris sur le jeu de développement. L'étape suivante est d'utiliser la sortie du système de reconnaissance, pour extraire les mots clés d'une sortie initiale (non adaptée) du système, pour ensuite s'en servir en utilisant des méthodes de recherche d'information et choisir des histoires pertinentes afin de construire un texte qui servira pour apprendre un (ou des s'il y a plusieurs thèmes) modèles, les poids étant ajustés sur la sortie de reconnaissance. Cette technique originale a permis d'obtenir des gains très significatifs pour le mandarin [Chen01A, Chen01B] et l'anglais [Chen03]. Nous avons également exploré la recherche de la technique d'adaptation était la plus efficace [Chen04] dans le cadre décrit ci-dessus ; nous avons donc comparé l'interpolation linéaire (un modèle est construit à partir des données d'adaptation et interpolé avec le modèle général), l'adaptation MAP (Maximum A Posteriori [Federico96]) qui est une autre forme d'interpolation, l'utilisation de mélanges de modèles, où l'on utilise un ensemble de textes correspondant à chaque thèmes, les poids d'interpolation étant déterminés sur les données d'adaptation, l'utilisation dynamique des mélanges de modèles, où les thèmes sont déterminés

dynamiquement pendant la reconnaissance, et enfin l'adaptation MDI (Minimum Discrimination Information [Federico99]). L'ensemble de ces techniques obtiennent à peu près les mêmes résultats, que ce soit sur l'anglais ou le mandarin, avec cependant un léger avantage à l'adaptation MDI. Enfin, un gain significatif peut être obtenu si l'on utilise des modèles de langage entraînés par histoires, et non pas par émission.

Bibliographie : [Chen00] [Chen01A] [Chen01B] [Chen03] [Chen04]

2.4 Modèle lexical : Vocabulaire de reconnaissance et lexique de prononciation

Le lexique constitue le point de rencontre entre le modèle de langage et les modèles acoustiques. Si l'on se place du côté du modèle de langage, on parlera de vocabulaire de reconnaissance, et les problèmes principaux seront la détermination de la couverture du vocabulaire en prenant en compte la définition du mot donnée par la normalisation (voir section 2.2), les sources possibles permettant de déterminer les mots à inclure, les contraintes de l'application qui donnent en particulier la taille maximale utile pour ce vocabulaire, et surtout le taux de mots hors vocabulaire qui sera le paramètre principal à optimiser lors de la détermination de celui-ci. Si l'on se place du côté du lexique de prononciation, le problème principal est d'avoir de bonnes prononciations, et des variantes pertinentes, en fonction de la tâche.

J'avais abordé le problème de la détermination du vocabulaire de reconnaissance et des variantes de prononciation admissibles lors de la construction du prototype de transcription sténotypes-graphèmes, au cours de ma thèse [Adda87B] ; la traduction sténotypes-graphèmes nous a permis de regarder les problèmes spécifiques à la reconnaissance de grands vocabulaires (ici 270 000 mots) et en particulier comment prévoir des variantes (principalement dans le cas de la sténotypie, les variantes dues à la présence ou non de la marque de voisement). Mais je me suis rapidement aperçu que la chaîne sténotypique n'était pas exempte d'erreurs, et que l'ambiguïté propre à la langue accentuée par l'ambiguïté propre à la chaîne sténotypique aboutissaient à un problème lors de l'accès à un lexique de grande taille. A la suite de certaines études [Pisoni85], Nous avons donc imaginé [Adda86] la possibilité d'un accès différent au lexique de prononciation, avec une étape de préclassification, à base de traits robustes [Crystal77], fondé sur des résul-

tats en psycholinguistique. L'idée sous-jacente à cette méthode est que lors de l'accès par un humain à son lexique, il y a affinement phonétique, suivant en cela la théorie des cohortes de Marlsen-Wilson [Pisoni85]. Le but est de réduire l'ambiguïté de l'accès au lexique, afin de ne pas choisir un mot sur un modèle phonétique inadéquat [Adda87]. Ce type d'accès au lexique, assez en vogue à la fin des années 80, a finalement disparu lorsque l'on s'est aperçu, en particulier avec l'apparition des modèles HMM de phonèmes en contexte, que l'on était capable de gérer une grande partie de l'ambiguïté phonétique directement lors de l'apprentissage des modèles.

Un exemple particulier de l'interaction entre accès au lexique et modèles phonétiques est l'utilisation de marques prosodiques, comme les tons en Mandarin ou le stress lexical en anglais, lors de la modélisation acoustique. Une expérience menée sur un des premiers systèmes de transcription de parole continue au LIMSI [AddaDecker92], a montré que, si l'utilisation de tels paramètres acoustiques pouvait montrer une certaine efficacité dans certaines circonstances, celle-ci était très limitée.

Dès cette époque, la construction des vocabulaire 20k, puis 65k, s'est faite en choisissant les mots en fonction de leur fréquence dans différentes sources, de manière à minimiser le taux de mots hors vocabulaire sur un jeu de données de développement [Lamel96A, Gauvain98A]. Pour ce qui concerne le lexique de prononciation, nous disposions d'une traduction graphème-phonème de qualité pour le français, GRAPHON [Prouts80], ce qui n'était pas le cas pour l'anglais. Nous avons donc mis au point un système afin de minimiser le temps de vérification des prononciations pour l'anglais, utilisant un ensemble de lexiques hétérogènes contenant des prononciations, joint avec une procédure d'inférence à partir des mots présents dans les différents lexiques ; cette procédure tenait compte de la confiance qu'on peut leur accorder, et contenait une recherche automatique d'une prononciation appropriée en ajoutant ou retirant des affixes et des suffixes, ou en accolant des mots présents dans un lexique. A l'aide de ce logiciel, nous avons pu prouver l'intérêt d'ajouter des variantes à un lexique de prononciation, pour peu qu'elles aient été vérifiées [Lamel96A, Lamel96C].

Le passage à la langue allemande qui, par son mode de composition de mots, de dérivation, et de casse, a des problèmes spécifiques de couverture (typiquement $> 5\%$ de mots hors vocabulaire pour un lexique 65k), nous a conduit à tester plusieurs manières de définir l'unité lexicale. Nous avons testé plusieurs décompositions lexicales ; le but de ces décompositions étant d'obtenir une bonne couverture tout en maintenant une bonne précision des

modèles de langage, l'utilisation des outils classiques de décomposition morphologique s'est avérée peu efficace. Nous avons donc introduit d'autres décompositions [AddaDecker00B, AddaDecker00A], y compris en regroupant les différentes inflexions d'une même racine.

Bibliographie :[Adda86] [Adda87] [Adda87B] [AddaDecker92] [Lamel96C] [Lamel96A] [AddaDecker00B] [AddaDecker00A]

2.5 Modèles acoustiques : adaptation non ou faiblement supervisée

Les méthodes statistiques utilisées pour développer les modèles en parole utilisent des corpus, et la corrélation entre la taille des corpus utilisés et les performances est un fait acquis [3]. Si le développement des modèles de langage s'est fait en récupérant de très grandes quantités de textes disponibles, la constitution de corpus de parole nécessite une phase de transcription coûteuse qui en limite le développement.

Afin de pouvoir accéder à de réellement très grandes quantités de données d'apprentissage pour le modèle acoustique, nous avons exploré la possibilité d'effectuer cette transcription *automatiquement*. Un premier test a porté sur l'utilisation de données supplémentaires de manière à aboutir à une transcription faiblement aidée, via à un apprentissage faiblement supervisé [Lamel00, Lamel01A]. Ces données supplémentaires étaient des sous-titres qui constituent une transcription approximative, produite à faible coût. Ces sous-titres étaient utilisés pour complémenter le modèle de langage, et ainsi guider la transcription des données d'apprentissage (500 heures utilisées dans l'évaluation SDR). Les résultats ont montré que cette approche permettait d'obtenir des résultats comparables (avec 10% de dégradation) à l'utilisation de données transcrites manuellement, dans cette configuration, et ce quelque soit le type de supervision choisie.

L'existence de sous-titres, ou plus généralement de textes très corrélés aux émissions à transcrire, est une contrainte forte. Nous avons donc également testé la procédure où le système est initialisé avec uniquement 10 minutes de données acoustiques transcrites manuellement, et un modèle de langage stan-

3. Certains auteurs [Moore03] font état par extrapolation des résultats passés d'une saturation des performances même pour des augmentation très significative ; ce point est débattu dans la section 6.2.1.

dard ; cette méthode ne nécessite aucun travail manuel, en particulier sur le dictionnaire de prononciation. Nous avons observé également une convergence mais avec un taux d'erreur (pour une quantité égale de données d'entraînement acoustique) beaucoup plus important qu'avec la méthode faiblement supervisée.

Bibliographie :[Lamel00] [Lamel01A] [Lamel02A] [Lamel02B] [Lamel03]

Chapitre 3

Applications

3.1 Transcription

Les systèmes de transcription multilingues de l'état de l'art sont une marque de fabrique du Groupe TLP du LIMSI, grâce aux travaux d'une petite équipe de personnes aux premiers rangs desquels se trouvent Jean-Luc Gauvain et Lori Lamel ; la mise au point de ces systèmes s'est faite conjointement avec la participation à des évaluations internationales, avec de très bons résultats. Cette histoire a commencé pour moi en 1992, avec la participation à la mise au point d'un système de dictée vocale en parole continue, d'abord en anglais [Gauvain93B, Gauvain93C], puis en français [Gauvain93C, Gauvain93B]. Ces systèmes pionniers nous ont permis d'effectuer des recherches en transcription multilingue, en comparant les performances sur des tâches identiques en français et en anglais avec un système de l'état-de-l'art [Gauvain94C, Gauvain94E, Gauvain94B, Gauvain94A, Lamel95]. Le système de transcription du LIMSI s'est ensuite développé au fil des années, utilisant les corpus supplémentaires à notre disposition [Gauvain96A, Lamel96D], ajoutant l'allemand [AddaDecker96A] au français et à l'anglais comme langue traitée, nous permettant toujours de comparer les performances théoriques et pratiques des systèmes dans un contexte multilingue [Lamel96B, AddaDecker96B].

A la suite de l'évaluation sur la transcription d'émissions radio-télédiffusées en anglais-américain (HUB4), de nombreuses techniques ont dû être mises en place [Gauvain97C]. En premier la segmentation des données en conditions homogène [Gauvain98B], l'utilisation d'une quantité beaucoup plus importante de données pour la construction du lexique 64k et du modèle trigramme

(environ 300M de mots). Il a fallu de plus gérer un ensemble de phénomènes comme les respirations et les pauses remplies (*filler words*) ; j'ai choisi de les inclure explicitement dans le modèle de langage, en utilisant un modèle génératif appris sur les transcriptions détaillées afin de modifier les textes d'apprentissage préalablement à la fabrication des n-grammes [Gauvain97C]. De plus, nous avons introduit des mots composés comme *going to* afin de permettre d'inclure des variantes de prononciations réduites. L'ensemble de ces techniques, introduites alors, a été conservé dans les systèmes actuels.

En 1998, nous avons introduit l'utilisation de 4-grammes obtenus par interpolation de 4 modèles individuels appris sur un ensemble de 540M de mots [Gauvain99A], avec de plus l'utilisation d'un bigramme de classes [Jardino94A]. L'utilisation de toutes ces techniques sur la langue française a permis également de nous rendre compte que le comportement, s'il restait globalement le même sur les deux langues française et anglaise, comportait quelques différences [AddaDecker99] : couverture plus faible pour le français, perplexité du modèle de langage moins élevée sur le français, et nature des erreurs (plus d'erreurs dues au phénomène d'homophones hétérographes, typique du français).

Le passage à des systèmes permettant des traitements en un temps réduit s'est fait lorsque la maturité des systèmes a été avérée. Ainsi, en 1999, nous avons introduit un système 10xRT (*Real Time*) avec une baisse de performance inférieure à 10%.

En 2000, nous avons ajouté à notre répertoire notre première langue non-européenne, le mandarin [Chen00], avec une évaluation non plus en terme de taux d'erreur de mots, mais de taux d'erreur sur les caractères. En 2004, les langues pour lesquelles nous disposions d'un système de transcription des émissions radio-télédiffusées étaient le français, l'anglais, l'allemand, le mandarin, le portugais, l'espagnol et l'arabe [Lamel04]. En particulier, le portage des techniques développées dans le cadre du projet EARS sur la langue française au cours de l'évaluation ESTER [Gravier04], nous a permis d'explorer en détail l'influence de la taille des lexiques, de la taille des données d'apprentissage acoustique, et de l'utilisation des modèles de classes sur les résultats en reconnaissance [Gauvain05A].

Nous avons ensuite abordé à partir de 2002 un type de parole tout à fait différent de la parole radio-télédiffusée : la parole conversationnelle téléphonique[Gauvain03C, Gauvain04A, Gauvain04B]. Nous avons en effet participé aux évaluations ARPA dites « HUB5 », en portant notre système de transcription HUB4, et en appliquant certaines transformations. Nous avons

ainsi changé l'utilisation de la normalisation du conduit vocal (*Vocal Tract Length normalization, VTLN*), l'utilisation de probabilité de prononciation, une adaptation MLLR utilisant plusieurs classes, un décodage par réseau de consensus et enfin l'utilisation d'un modèle de langage neuronal. Ces techniques avaient été testées auparavant dans le système HUB4, mais n'avaient pas été intégrées car n'apportant pas un gain significatif et stable ; on peut remarquer cependant qu'elles sont également dorénavant parties intégrantes des systèmes HUB4. En 2004, les langues pour lesquelles nous disposions d'un système de transcription des conversations téléphoniques étaient le français, l'anglais et l'arabe [Lamel04].

A la suite des progrès sur les données radio-télédiffusées, puis sur les données conversationnelles téléphoniques, un certain nombre de projets a porté sur la transcription d'autres données, pour des applications plus complexes. Parmi ces nouvelles applications, la transcription de réunions ou de séminaires (dans le cadre du projet CHIL) qui permet de mettre en œuvre le concept de « Rich Transcription », c'est-à-dire « qui dit quoi, où et quand » ; à cette fin, de nombreux enregistrements ont eu lieu dans des pièces aménagées, avec différentes sortes de microphones. Une première expérience en utilisant des données d'apprentissage non spécifiques [Lamel05B, Lamel05A], testées sur des séminaires a permis d'obtenir des taux d'erreur autour de 25% pour les données avec un microphone de proximité, et beaucoup plus important ($> 60\%$) avec des micro distants [Lamel06A, Lamel07A]. Autre application, la transcription des débats à l'Assemblée Européenne (données EPPS, *European Parliament plenary sessions*), qui permet de développer l'application de traduction Parole-Parole (projet européen TC-STAR) [Lamel06B, Lamel07B] ; au contraire des données de réunions, les données EPPS ont prouvé être faciles, car très préparées, avec des taux d'erreur de 8% pour l'anglais et de 7% pour l'espagnol.

Bibliographie :[Bellilty85] [AddaDecker92] [Gauvain93C] [Gauvain93B] [Gauvain94F] [Gauvain94E] [Gauvain94D] [Gauvain94C] [Gauvain94B] [Gauvain94A] [Lamel95] [Lamel96D] [Lamel96B] [AddaDecker96A] [AddaDecker96B] [Adda97E] [Gauvain97C] [Gauvain97A] [Gauvain98B] [AddaDecker98B] [AddaDecker99] [Gauvain00C] [Gauvain00A] [Chen00] [Gauvain01A] [Gauvain02A] [Gauvain03D] [Gauvain03C] [Gauvain03B] [Lamel04] [Gauvain04B] [Gauvain04A] [Lamel05B] [Lamel05A] [Gauvain05B] [Gauvain05A]

3.2 Indexation de documents audio

L'une des application majeure de la transcription de la parole concerne l'accès par le contenu des documents multimédia contenant de l'audio, que ce soient des documents radio-télédiffusés ou des documents disponibles sur le WEB [1]. A la fin des années 90, de multiples travaux et évaluations ont pris pour tâche de réunir en un seul système optimisé la transcription de la parole et la recherche d'information. En prenant pour cadre l'évaluation *spoken Data Retrieval, SDR* [Garofolo00], nous avons construit un système à partir de notre système de transcription HUB4 [Gauvain99A], avec un système de recherche d'informations fondée sur Okapi qui nous a permis de participer a montré la faisabilité de tels systèmes d'indexation [Gauvain99C, Gauvain00B, Gauvain00A]. Cela nous a permis de montrer également qu'une recherche fondée sur une définition partagée de la notion de mot permettait une meilleure optimisation [Gauvain99D], fait qui sera avéré par la suite pour les autres systèmes de traitement de la langue (traduction, systèmes Q&A) pour lesquels nous avons introduit une composante parole. Ce système SDR s'est ensuite affiné par l'utilisation d'une segmentation automatique en thème, et l'utilisation de l'expansion de requêtes [Gauvain00D]. Dans le cadre du projet européen Olive, le système de recherche d'informations sur les données radio-télédiffusées a été porté au français et à l'allemand [Gauvain01A], puis au portugais et au mandarin [Gauvain01C] avec des résultats comparables. Suite aux succès des évaluations SDR, la tâche s'est complexifiée en intégrant d'autres sortes d'applications, par exemple la tâche « Rich transcription », ou l'utilisation de la vidéo, de la musique, et a permis de créer un communauté autour de la problématique de l'indexation par le contenu, par exemple à travers les conférences CBMI (Content Based Multimedia Indexing) [Iquierdo07].

Bibliographie :[Gauvain99C] [Gauvain99D] [Barras00] [Gauvain00A] [Gauvain00B] [Gauvain00D] [Gauvain01A] [Gauvain01B]

1. l'accès aux documents multimédia sur le Web est l'objet spécifique du projet Quaero`http://www.quaero.org/`

3.3 Autres systèmes : Systèmes de Réponses à des Questions (Q&A) et Traduction statistique

Dans la section précédente, nous avons vu l'extension de systèmes de traitement du langage écrit à la modalité parole. D'autre part, le principe de modélisation statistique de la langue a été appliqué à des systèmes traitant d'autres modalités que la parole. En particulier, 2 systèmes de traitement de la langue ont bénéficié à la fois de la modélisation statistique et de l'extension au traitement de la parole, il s'agit de la traduction et des systèmes de réponses à des questions (Q&A).

Les systèmes de réponses à des questions sont une évolution des systèmes de recherche d'informations, où, plutôt qu'un document, on donne la réponse « précise » à la question posée. Comme pour les applications SDR (voir section 3.2), les systèmes Q&A ont évolué vers la réponse à des questions non plus seulement dans des données écrites, mais également sur des données de parole [Rosset07A, Rosset07B, Rosset08]. Comme pour les évaluations SDR, nous avons pu constater que l'homogénéisation de la normalisation des données, c'est-à-dire d'utiliser une définition la plus proche possible de ce qu'on appelle un mot dans le système de transcription et dans le système Q&A apportait un gain absolu en terme de rappel-précision, mais également en ce qui concerne la facilité de développement des systèmes [Rosset09, Bernard09].

En ce qui concerne la traduction statistique, le lien avec la parole se fait encore plus intimement que dans le cas des systèmes Q&A, les systèmes de traduction utilisant un modèle de langage. Une étude assez précise [Dechelotte07B, Dechelotte07A] a permis de mesurer l'impact de certaines normalisations communes lorsque l'on normalise les textes pour fabriquer les modèles de langage pour la traduction, dans un schéma expérimental proche de ce qui avait été fait pour la reconnaissance du français [Adda97D]. Nous avons en particulier remarqué que la normalisation gagnait en efficacité lorsqu'elle permettait de rapprocher le nombre d'unités lexicales des deux côtés de la traduction, et qu'il était important de tenir compte de la ponctuation (par exemple en l'ajoutant dans les sorties de reconnaissance) [Dechelotte08].

Bibliographie :[Rosset07A] [Rosset07B] [Dechelotte07A] [Dechelotte07B] [Rosset08] [Dechelotte08] [Rosset09] [Bernard09] [Quintard10]

3.4 Systèmes industriels

Dans tout mon parcours de recherche, j'ai en permanence participé à la fois à des développements expérimentaux et à la mise au point de systèmes industriels, c'est-à-dire des systèmes ayant pour finalité d'être réellement utilisé dans une application par de vrais utilisateurs. Notre domaine a souffert dans le passé de critiques, à la fois sur le caractère scientifique des travaux (voir la critique de J.R. Pierce 5) et sur la possibilité de voir de réelles applications, utiles commercialement mais surtout ayant un impact social positif (idem, mais aussi [Moore05a] et [Levinson95]). Le travail avec les industriels n'est donc pas seulement (comme le croit certains) un moyen d'avoir de l'argent, directement ou indirectement en montant des projets financés par l'état. C'est également le moyen de vérifier que l'objet de notre recherche peut avoir un impact sociétal réel, et de maîtriser la nature de celui-ci.

J'ai donc commencé dès le début de ma thèse a travaillé sur des systèmes industriels ; tout d'abord parce que celle-ci s'est déroulée au LIMSI, financé sur un contrat industriel avec la société Sténotype Grandjean commercialisant les systèmes de sténotypie et assurant la formation des sténotypistes, au sein d'une PME en tant qu'ingénieur. Cette thèse avait pour sujet de mettre au point un logiciel et un matériel permettant de produire du texte français à partir d'une saisie de sténotypes. Cette application avait pour but d'améliorer en général la productivité des sténotypistes [Adda84A, Adda87B], et en particulier de fournir en temps réel un sous-titre à destination des malentendants, en collaboration avec le CCETT à Rennes [Adda83B, Adda83A, Neel85A, Neel85B, Adda88]. Mais également par le biais de la sténotypie, qui constitue une prise phonétique rapide de la parole, cela nous a permis de nous attaquer au problème du passage d'une transcription phonétique à un texte en français, en vocabulaire ouvert [Adda84A, Bellilty85, Adda87B]. Le contrat a débouché sur un prototype opérationnel, qui a été démontré pendant la présentation au CCETT du prototype [Adda88], avec sous-titrage en temps réel des présentations orales des participants. Hélas pour la survie industrielle de ce prototype, l'équipe parole d'IBM France avait développé en parallèle le logiciel TASF+ [Merialdo85] pour les mêmes raisons de proximité avec le problème de la transcription de la parole [Derouault85, Derouault86] ; seul le système IBM a ensuite perduré.

A la suite de ce projet, la modélisation n-gramme au LIMSI s'est mise quelques temps en sommeil (voir section 2.3.1), avec le développement d'un projet ambitieux, Polyglot [Vittorelli90], au sein duquel les applications in-

dustrielles se concentraient sur la mise en œuvre du système Olivetti de dictée vocale en mots isolés à l'ensemble des autres langues européennes, et en particulier au français [Nait92]. Là aussi, la mise sur le marché du système Dragon, puis IBM n'a pas permis à cette solution d'exister industriellement.

Le LIMSI a ensuite participé à de nombreux projets européens, ainsi qu'à de très nombreuses évaluations, permettant le portage de son système dans un certain nombre d'applications. Le portage d'un système de reconnaissance construit sur un domaine vers un autre domaine a inspiré une étude [Lamel01B] portant sur la capacité à utiliser un système donné pour plusieurs tâches, en adaptant ce système par exemple à l'aide d'un apprentissage faiblement supervisé (voir section 2.5) ; en adaptant un système généraliste (le système de transcription développé dans les évaluations HUB4, voir section 4.3) des résultats comparables voire meilleurs que ceux des systèmes dédiés ont pu être ainsi obtenus.

Dernière application industrielle en date, la fouille de données dans les conversations téléphoniques agent-client, ce que l'on regroupe sous le vocable « Speech Analytics ». Le développement de cette application, CallSurf, implique de nombreuses étapes de traitement allant donc de la transcription et de la structuration des conversations téléphoniques, aux traitements permettant d'extraire les entités nommées et les termes spécifiques à l'application ; ensuite, les outils de fouilles de textes permettent de faire ressortir qualitativement et quantitativement les thèmes exprimés au cours de ces conversations, ou de pointer l'ensemble des conversations où se sont exprimés ces thèmes [GarnierRizet08, Adda11A, Adda11B].

Bibliographie :[Adda83A] [Adda83B] [Adda84A] [Neel85A] [Neel85B] [Adda88] [Vittorelli90] [Nait92][Lamel01B] [GarnierRizet08] [Adda11A] [Adda11B]

3.5 Systèmes automatiques comme instrument d'exploration

Dans notre laboratoire (le LIMSI) nous avons eu la chance que des chercheurs ouverts, en provenance de plusieurs cultures (principalement linguistique, phonétique, statistique, informatique) ont pu se côtoyer puis travailler ensemble. De ce travail est né (comme à d'autres endroits) l'idée que les systèmes automatiques étaient de formidables instruments pour explorer les cor-

pus de grandes taille et découvrir de nouvelles propriétés (voir section 6.2.2).

La première étude en ce sens a porté sur la structure de la syllabe en français dans un corpus de 30 heures d'interviews politiques [AddaDecker02]. On a ainsi pu recenser quantitativement un certain nombre de phénomènes de variations de prononciations, par exemple en ce qui concerne le schwa et la liaison, et observer les phénomènes de restructuration, dues à des omissions de voyelles ; ces variantes sont connues pour poser problème aux systèmes de reconnaissance automatique. Pour ces études, nous avons utilisé un système d'alignement, avec des dictionnaires de prononciation contenant l'ensemble des variantes de syllabes.

D'autre études ont porté sur l'exploration du corpus PFC [Durand03], corpus qui vise à proposer des échantillons représentatifs des parlers normatifs et vernaculaires d'un grand nombre de variétés de l'espace francophone. En particulier, dans [AddaDecker06], 12 points d'enquête, en France, Suisse et Belgique ont été utilisés, et les résultats sur les durées de segments phonémique, en parole lue et spontanée, sur les fréquence des phonèmes en général et en particulier des voyelles en contexte, ont pu être obtenus par segmentation et alignement phonétique automatique.

Nous avons abordé le problème des disfluences, qui sont des phénomènes typiques de la parole spontanée, et qui sont souvent cités comme cause pour les mauvaises performances des systèmes de reconnaissance sur ce type de données. Le but de ce travail était à la fois de développer une méthodologie d'annotation des données spontanées, d'étudier quantitativement et qualitativement ces phénomènes, et enfin d'essayer d'améliorer les systèmes automatiques. Nous avons utilisé un corpus d'interviews politiques de 10 heures, issues d'une émission politique du début des années 1990, « L'Heure de Vérité », où une personnalité politique ou de la société civile, était interrogée à tour de rôle par 3 journalistes. L'ensemble des débats est surveillé par un unique maître de cérémonie, qui dirige les échanges, surveille le temps, etc. Ce type d'interview, où le journaliste intervieweur est souvent pris par le temps, et l'interviewé est obligé de réagir rapidement, amène des digressions, des interruptions, qui favorisent l'apparition de disfluences. Pour ces 10h de parole, nous disposons de 2 types de transcriptions, une transcription exacte, complète, résultat de notre transcription manuelle à l'aide de Transcriber [Barras01], et une transcription approximative mais *bona fide*, destinée à la presse, et qui en particulier ne contiennent plus de disfluences. Une première étude [AddaDecker03, AddaDecker04] a été faite sur un sous-ensemble de transcriptions exactes (10% de chaque émission), afin de mettre au point

les annotations choisies, d'étudier la répartition des disfluences ainsi que les résultats de la transcription automatique. Cette étude a montré que seules 50% des différences entre la transcription exacte et la transcription approchée étaient dues aux disfluences, et que si celles-ci induisaient un taux d'erreur plus important que d'autres mots, la cause principale d'erreur était le phénomène de réduction des prononciations, autre phénomène typique de la parole spontanée. A partir des mêmes données nous avons également regardé si nous pouvions utiliser les transcriptions approchées pour aider à la production des transcriptions exactes [Barras04], par transfert des time codes grâce à la transcription automatique : en effet, seules les zones où la transcription approximative et la transcription automatique sont en désaccord (20% du total) nécessitent une relecture approfondie. Nous avons pu ainsi produire la totalité des transcriptions exactes des émissions, ce qui nous a permis de les annoter précisément en disfluences et en marqueurs de discours, et donc de fournir une étude qualitative et quantitative de ces phénomènes [BoulaDeMareuil05]. Nous avons ensuite abordé un autre type de phénomène spécifique à la parole spontanée interactive, la parole superposée. Ce type de parole est encore peu étudié, principalement parce que peu de systèmes sont capables de la traiter voire même de la détecter. Si la parole superposée représente un fraction quasi-négligeable des données *Broadcast News*, et qu'elle ne constitue pas un réel problème pour les conversations téléphoniques dès lors que l'on peut accéder au canal de chaque locuteur, il n'en est pas de même pour les données *Broadcast Conversation* dont fait partie le corpus de « L'Heure de Vérité », qui contiennent une fraction non négligeable de parole superposée (environ 5% des données). Pour un système dont la charge est de transcrire l'ensemble des données radio-télédiffusées, il est nécessaire d'étudier ce phénomène, pour mieux le connaître, le qualifier, puis le traiter. Nous avons tout d'abord développé un guide d'annotation utilisant 4 types de parole superposée : *régulateur (backchannel), complémentaire (complementary), prise de parole (turn stealing), anticipation (anticipated turn taking)* [Adda07]. L'annotation des 10 heures du corpus nous a permis d'étudier à la fois le phénomène de la parole superposée, et son interaction avec le phénomène des disfluences, en particulier en fonction du rôle du locuteur [AddaDecker08B].

Bibliographie :[AddaDecker02] [AddaDecker03] [Barras04] [AddaDecker04] [AddaDecker05] [BoulaDeMareuil05] [AddaDecker06] [Adda07] [AddaDecker08A] [AddaDecker08B] [Snoeren10] [AddaDecker11]

Chapitre 4

Évaluation comparative

4.1 Développement de corpus

Le développement de corpus est essentiel pour la mise en place de l'évaluation comparative. En parallèle avec nos premières participations aux évaluations ARPA où nous avons pu bénéficier de la mise à disposition des premières grandes bases de données (texte et parole) en langue américaine, le LIMSI a développé un premier corpus de parole de référence pour la langue française, BREF [Lamel91]. Ce corpus a servi de base pour la première évaluation de dictée vocale en français (voir section 4.2) ; au cours de cette évaluation dite AUPELF [AddaDecker97], j'ai de plus participé à la distribution des textes d'apprentissage (40 millions de mots issus du journal *Le Monde*), des lexiques (20k et 64k) et des modèles de langage (bigramme et trigramme) qui ont pu être utilisés par les participants [Adda97C, Adda00B].

L'idée qu'il est nécessaire d'avoir des corpus afin d'apprendre et d'évaluer des systèmes automatiques a atteint à cette époque une étape significative, avec la mise en place de la première conférence « on Language Resources and Evaluation », LREC en 1998. Les industriels ont alors également cherché à développer des corpus qui pourraient s'avérer stratégiques : au LIMSI nous avons alors construit des corpus pour l'identification du locuteur (Convention de recherche France Telecom sur la reconnaissance du locuteur 1994-1997), ou l'identification des langues (Convention de recherche CNET sur l'identification de la langue (IDEAL) 1994-1997) [Lamel98], ce dernier corpus comprenant un grand nombre d'appels en différentes langues, en provenance du pays d'origine, ou d'autre pays, totalisant 70 heures de parole transcrites.

Nous avons ensuite participé à la définition de nombreux corpus, en interaction avec les nombreux projets et évaluations auxquels le LIMSI a participé. Nous avons en particulier développer des corpus de données audiovisuelles dans un grand nombre de langues européennes, dans le cadre du projet Quaero.

La problématique de constitution de corpus, de taille toujours plus importante, pour un nombre de langues et pour des types de parole toujours plus variés, pose actuellement des problèmes nouveaux, en particulier pour les langues peu dotées (voir par exemple notre travail pour la constitution de ressources pour la langue luxembourgeoise [AddaDecker08A, Snoeren10]). Parmi les méthodes actuellement en vogue, mon attention a été retenue par les méthodes de crowdsourcing, et en particulier Amazon Mechanical Turk ; dans ces méthodes, dites de microworking, le travail est séparé en tâches élémentaires, chaque tâche étant postée sur le Web, et effectuée par des anonymes pour une rétribution très modeste. Si j'ai étudié assez en détail ce mode de production [Adda10B, Adda10A, Fort11, Sagot11], c'est parce qu'il est à la fois porteur d'un immense potentiel, mais également de graves problèmes, notamment par rapport à l'éthique et du droit du travail.

Bibliographie :[Adda97C] [AddaDecker98A] [Lamel98] [Adda00B] [Baude06] [AddaDecker08A] [Adda10A] [Adda10B] [Fort11] [Sagot11]

4.2 Mise en place d'évaluations

Les échos du succès du paradigme de l'évaluation sont arrivés rapidement en Europe au début des années 90, mais peu de laboratoires du vieux continent ont été tentés au début par la participation directe. Parmi ceux-ci, le LIMSI a participé dès le début, encouragé en cela par Joseph Mariani qui a développé ce même paradigme en France et en Europe. Le début de l'aventure s'est trouvé être, en France, la mise en place d'une évaluation sur les annotateurs automatiques en parties du discours (POS), et plus généralement les analyseurs morphosyntaxiques, l'action Grace [Adda95, Paroubek97]. Cette action a débuté en 1994 sous les auspices des départements des Sciences Humaines et des Sciences pour l'Ingénieur du CNRS, et de l'AUPELF-UREF [1]. Cette action a d'abord rassemblé un certain nombre d'organisateurs, moi pour le LIMSI, Patrick Paroubek, à l'époque à l'INaLF, et Martin Rajman, à l'époque à l'ENST. Le programme, ambitieux, devait mesurer la performance

1. `http://www.refer.mg/general/agence.htm`

des analyseurs morpho-syntaxiques en absolu, et dans un certain nombre de tâches pour lesquelles l'annotation en parties du discours (POS) est utile. Le programme a dû être revu à la baisse, suite à l'arrêt des financements. Cependant, le comité d'organisation, complété par Joseph Mariani, et Josette Lecomte de l'INaLF a réussi à mettre en place les corpus, leur annotation [Lecomte97], et la procédure d'évaluation assez compliquée : chaque système dispose d'un jeu de POS spécifique pour lequel il a fallu définir une procédure de projection sur un ensemble commun d'étiquettes morphosyntaxiques ; d'autre part, la plupart des systèmes disposaient de leur propre système de normalisation, donc d'une définition d'unité lexicale qui n'était pas toujours compatible, et il a fallu définir une procédure de normalisation sur laquelle toutes les normalisations pouvaient se projeter. La campagne hors financement a rassemblé 13 participants internationaux (sur les 21 institutions inscrites au départ) [Adda98, Adda99C, Adda99B], et malgré quelques imperfections dues au manque de financement, à la nouveauté de la tâche, et à sa difficulté, a permis de dégager un communauté, de mettre en place des outils de mesure acceptés par elle [Adda00D], et a permis de produire un corpus annoté en POS de 1 million de mots, MULTITAG [Paroubek00].

En parallèle de l'évaluation GRACE, j'ai participé à la mise en place de la première évaluation en dictée vocale en français, lors de l'action AUPELF. Au cours de cette évaluation [AddaDecker97], le corpus BREF contenant plus de 100 heures de lecture de journal par 120 locuteurs a été distribué, ainsi que des textes d'apprentissage, des lexiques et des modèles de langage [Adda97C]. Cette première évaluation a renforcé la cohérence de la communauté en traitement automatique de la parole en français, avec la participation de tous les principaux laboratoires en France. Elle a permis de mettre en œuvre, au sein des laboratoires français, la mise en place du paradigme de l'évaluation qui passe obligatoirement par l'existence d'un système complet de transcription.

Bibliographie :[Adda95] [Adda96] [Lecomte97] [AddaDecker97] [Paroubek97] [Adda98] [Adda99B] [Adda99C] [Adda00D]

4.3 Participation à des évaluations

A la suite de la participation du LIMSI aux évaluations ARPA sur les données Resource Management [Lamel92], lorsque ces évaluations ont évolué vers la lecture d'articles de journaux financiers (données *Wall Street Journal*), j'ai incorporé aux systèmes du LIMSI, dans le cadre de la participation du

LIMSI aux évaluations DARPA en 1992 sur les données *Wall Street Journal*, avec un vocabulaire de 5k et 20k mots [Gauvain93A], les premiers modèles n-grammes appris sur des textes réels (en l'occurrence texte d'apprentissage de 33M de mots). Cette participation s'est continuée en 1993, sur la même tâche [Gauvain94G, Gauvain94H], puis sur la tâche HUB3 impliquant plusieurs prises de sons [Gauvain96B, Adda96]. Au cours de ces évaluations, nous avons pu passer à des trigrammes, en utilisant plus de textes, et avec un lexique de 64k mots.

L'évaluation des systèmes a ensuite changé de nature au cours de l'année 1996, lorsque fût abordé le problème de la transcription d'émissions de radio (évaluation dite HUB4) : c'est la première fois qu'une évaluation portait sur de la parole réelle, au sens où elle n'avait pas été recueillie uniquement aux fins d'évaluation. Alors que beaucoup de laboratoires étaient pessimistes quant à la faisabilité d'une telle évaluation, prévoyant des taux de reconnaissance catastrophiques, de premiers tests on montré que cette hypothèse était fausse. Au cours de notre première participation [Gauvain97B], nous avons pu mettre en place un grand nombre de techniques nouvelles, et obtenir un très bon résultat, confirmé les années suivantes [Gauvain98A, Gauvain99A, Gauvain99B, Gauvain00E]. En 1999, nous avons introduit un système 10xRT [Gauvain00E, Gauvain00F], avec une baisse de performance inférieure à 10% par rapport au système 50xRT.

En 1997, nous avons participé à la première évaluation en dictée vocale en français [Adda97A, Adda97E, Adda97D, Adda00A], où nous avons pu évaluer nos résultats en confrontation avec d'autres approches, sur le français ; ici, comme lors des évaluations ARPA, le schéma de reconnaissance désormais classique de reconnaissance statistique de la parole a pu montré sa supériorité sur d'autres schémas, plus exotique, où par exemple la reconnaissance se fait en 2 étapes, avec d'abord construction d'un treillis phonétique puis l'application des connaissances linguistiques sur ce treillis. En particulier le système du LIMSI a obtenu lors de cette évaluation les meilleurs résultats.

Le succès des évaluations successives sur la transcription de la parole, a amené à pouvoir se poser la question des applications possibles. Parmi celles-ci, l'une des plus prometteuses était et reste la recherche d'information dans les documents audio ou multimédia. Ceci nous a amené à participer aux évaluations SDR [Gauvain99C, Gauvain00D] (*Spoken Data Retrieval*) faisant partie des évaluations TREC [2] en 1999, où nous avons couplé un

2. `http://trec.nist.gov/`

système d'information à notre système de transcription en utilisant dans les deux systèmes la même notion de mot fournie par la normalisation. Les résultats obtenus dans ces évaluations ont montré que pour des taux d'erreurs de 20%, la recherche d'information sur la transcription automatique n'était pas dégradée par rapport à la recherche d'information sur la transcription manuelle.

Au cours des années suivantes (2002-2004) nous avons collaboré avec BBN au sein du programme EARS (*Effective, Affordable, reusable Speech-to-text*) sur la transcription des données *Broadcast News, BN* [Gauvain02B, Gauvain03A], pour l'anglais, l'arabe et le chinois [Schwartz04, Nguyen04] et sur la transcription des données conversationnelles de type *Switchboard* (*Conversational Telephone Speech, CTS*) en anglais [Prasad04, Prasad05]. Ceci nous a permis de mettre en œuvre les techniques de portabilité que nous avions étudiées auparavant [Lamel01B, Gauvain02B, Gauvain03D, Gauvain03E]. Les systèmes des deux sites étaient utilisés conjointement, par l'utilisation du système ROVER [Fiscus97], ce qui a permis d'obtenir des résultats significativement meilleurs que le meilleur des 2 systèmes [Prasad04, Nguyen04]. Au total la collaboration a pu aboutir à une réduction du taux d'erreurs relative de 47% sur les données BN et de 51% sur les données CTS [Matsoukas06].

En 2004 nous avons porté notre système de transcription des données radio-télédiffusées, qui avait été développé sur l'anglais, à la langue française, lors de l'évaluation ESTER [Gravier04]. Cette évaluation nous a permis de démontrer des résultats comparables à ceux obtenus sur la langue anglaise [Gauvain05B, Gauvain05A], obtenant par ailleurs les meilleurs résultats de cette évaluation.

Le groupe TLP a continué à participer aux évaluations nationales comme ESTER2 [Galliano09] en 2008, ou les évaluations internationales dans le cadre de GALE[3], ou dans le cadre du projet Quaero[4] [Geoffrois08b]. Ces évaluations font maintenant partie du quotidien des chercheurs en traitement automatique de la parole, ou en traduction automatique. Malgré la lourdeur de l'investissement humain que ces évaluations impliquent, le bénéfice qu'on en retire, en terme de visibilité de nos travaux, de reproductibilité (voir section 6.1.5) justifie pleinement cet investissement.

Bibliographie :[Gauvain93A] [Gauvain94H] [Gauvain94G] [Gauvain96A]

3. `http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_(GALE).aspx`

4. `http://www.quaero.org/`

[Gauvain96B] [Gauvain97B] [Adda97E] [Gauvain98A] [Gauvain99A] [Gauvain99B] [Adda00A] [Gauvain00E] [Gauvain00F] [Gauvain02B] [Gauvain03A] [Gauvain03E] [Schwartz04] [Prasad04] [Nguyen04] [Prasad05] [Lamel06A] [Lamel06B] [Matsoukas06] [Lamel07A] [Lamel07B]

Deuxième partie

Développement d'une science empirique en traitement de la parole

Chapitre 5

Introduction

La reconnaissance de la parole fait partie de ces tâches humaines pour lesquelles on a pensé depuis très longtemps qu'un ordinateur pourrait les simuler, voire les reproduire. Il s'agit d'une tâche complexe, traditionnellement classifiée dans la classe des problèmes de « Reconnaissance des Formes », et qui a donc grandement bénéficié (comme les autres problèmes de cette grande famille) de l'évolution en vitesse et en capacité de mémoire des ordinateurs. Cependant, nous pouvons voir que les progrès observés ne reposent pas uniquement sur l'augmentation de la puissance des machines et de la quantité de données traitée. Ces progrès reposent également sur un changement de paradigme, survenu à la fin des années 1970 aux États-Unis.

Ce changement est apparu conjointement aux critiques formulées par J.R. Pierce en 1969, dans une lettre à l'éditeur publiée dans le JASA [1] :

> **Whither Speech Recognition ?**
>
> [...] General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish.
>
> [...] It would be too simple to say that work in speech recognition is carried out simply because one can get money for it.

1. J.R. Pierce était une personnalité importante, ingénieur et et collègue de Shannon, Shockley, Bardeen, Darlington, *Executive Director* de ATT Bell Labs ; il a été à la tête du « Automatic Processing Committee of Defense Dept. » en 1966, avec une opinion aussi négative de la traduction automatique que de la reconnaissance de la parole qui a conduit le gouvernement à arrêter de financer la traduction automatique, et a mis une grande pression sur les projets en reconnaissance de la parole.

> That is a necessary but no sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.
>
> [...] It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect.
>
> [...] Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

Il n'est pas indifférent de rappeler qu'à cette époque, les approches statistiques et de manière générale utilisant des méthodes d'apprentissage permettant de modéliser les données, étaient déconsidérées depuis longtemps ; cette déconsidération provenait de la position de Chomsky que l'utilisation de statistiques, et de manière générale d'une approche empirique, n'était pas légitime, pas « scientifique », en ce qui concernait le langage. Cette position avait influencé fortement tout le domaine de l'Intelligence Artificielle, qui était l'approche dominante en traitement de la parole et du langage.

A la suite de cette attaque, les résultats du projet ARPA-SUR (1971-1976) ont démontré à la fois l'intérêt des approches dites « ingénieur », la nécessité d'avoir une évaluation scientifiquement fondée, et l'intérêt de s'attaquer en priorité à des problèmes de faible difficulté pour permettre une progression.

Sont alors apparus conjointement, et ce n'est pas un hasard, deux faits majeurs :

- le développement de l'approche statistique, fondée sur la modélisation statistique de la parole (voir section 2), issue de la théorie de l'information ;
- l'introduction du principe de l'*évaluation comparative* des systèmes [Doddington81, Pallett82, Pallett85].

Si les deux approches (Intelligence Artificielle et Modélisation Statistique)

ont cohabité jusqu'à la fin des années 80, avec l'émergence de modèles issus de l'intelligence artificielle, comme les systèmes experts [Bonnet86], ou l'approche « Tableau Noir » [Laasri88], les résultats des évaluations ARPA sur les données *Resource Management* [Price88], puis *Wall Street Journal* [Pallett93], ont rendu rapidement caduque toute autre modélisation autre que statistique.

L'introduction de ces deux principes énoncés ci-dessus structurent la recherche en traitement de la parole depuis 40 ans, au point où, plus qu'une approche dominante, ils constituent un couple de pratiques quasi-hégémonique dans la production scientifique actuelle en traitement de la parole.

Le but de cette deuxième partie de mon document est de reprendre ces deux principes, de décrire l'avancée majeure qu'ils ont engendrée et de tenter une explication de son efficacité. Je reprendrai ensuite certaines critiques qui lui ont été faites et qui font écho à la critique de J.R. Pierce. En s'appuyant sur ces analyses, je proposerai des solutions pragmatiques pour faire vivre cette approche, et la faire évoluer vers une réelle science empirique de la parole.

Chapitre 6

Évaluation et corpus

6.1 Aperçu historique de l'évaluation en transcription automatique

Je vais, avant de présenter certaines vue polémiques, essayer de résumer assez objectivement ce qu'est le paradigme de l'évaluation, avec un focus en particulier sur la tâche de transcription, historiquement et épistémologiquement au cœur du paradigme de l'évaluation, et de son évolution.

6.1.1 Le paradigme de l'évaluation en reconnaissance de la parole

L'amélioration des technologies vocales s'est faite parallèlement à une maturation du marché et une prise de conscience de leur potentiel par les acteurs économiques. L'évaluation des systèmes a pris une part importante dans ces différentes avancées, en donnant aux chercheurs un cadre théorique et pratique qui a permis d'accélérer la mise au point et la diffusion des méthodes efficaces, en abordant des domaines a priori prometteurs mais pour lesquels les investissements nécessaires auraient pu rebuter les volontés, et enfin en permettant aux investisseurs et aux industriels d'avoir une vision assez claire des technologies efficaces et de leurs performances.

Nous abordons ici brièvement ce que nous appellerons l'*évaluation comparative* qui caractérise l'évaluation des technologies, et que l'on peut opposer à l'évaluation qualitative de satisfaction d'utilisateurs (typique de l'évaluation d'un produit) et à l'évaluation absolue qui vise à donner une mesure de la

performance absolue d'un système par rapport à une tâche.

Le principe de l'évaluation comparative est utilisé aux États-Unis depuis le début des années 1980 [Doddington81], puis dans les années suivantes avec succès dans le cadre des évaluations ARPA dans le domaine de l'ingénierie des langues, et en particulier en ce qui concerne la reconnaissance de la parole. Ce principe a été proposé par le DARPA [1] et implémenté par le NIST [2] comme un moyen de promouvoir le développement de la recherche et de la technologie dans le domaine de l'ingénierie des langues. Bien que l'évaluation des systèmes de reconnaissance ait mobilisé l'attention et les efforts de beaucoup de chercheurs dans les années 1980 [Pallett82] et antérieures ailleurs qu'aux États-Unis (voir par exemple [Chollet82, Taylor80, Moore77]), la mise en place du paradigme et son utilisation comme outil de développement scientifique ont commencé avec les premières évaluations ARPA aux États-Unis, à partir de 1987.

La mise en place d'une évaluation comparative nécessite plusieurs éléments indispensables :

– *Une tâche*. Une tâche d'évaluation doit être à la fois proche des applications potentielles, afin que les retombées industrielles soient envisageables rapidement, mais suffisamment générique pour que le plus grand nombre d'intérêts particuliers des acteurs industriels ou académiques intervenant dans le domaine, soient représentés. Une hypothèse est associée à ce présupposé : si la tâche est représentative d'une problématique réelle du domaine, alors une différence *quantitative* significative entre les résultats obtenus en utilisant, au sein d'un système, deux méthodes différentes, traduit une différence *qualitative* entre les deux méthodes, sur l'ensemble des applications potentielles (généricité). Pour ces propriétés, on la nommera **tâche de contrôle**. Par exemple, pour la transcription d'émissions radio-télédiffusées, ARPA et NIST ont défini une tâche de contrôle consistant à transcrire une dizaine d'heures de radio et de télévision, enregistrées pendant une période précise, et pour laquelle il était interdit d'utiliser des données de cette période temporelle ou la postdatant ; cette tâche est générique de toutes les applications (alerte, recherche d'information, sous-titrage,...) impliquant les données radio-télédiffusées d'information. Il est possible que certains domaines ne se prêtent pas au développe-

1. *Defense Advanced Research Agency*
2. *National Institute of Standards and Technology*

ment de tâches de contrôle, ou encore que les seules tâches imaginables soient trop loin des applications réelles ; dans ce cas l'évaluation ne sera pas pertinente. Un exemple de cette situation sont les applications de dialogue homme-machine : s'il a été possible d'organiser des évaluations portant sur des sous-tâches [Walker00] (par exemple la compréhension [Bonneau-Maynard06]), il est très difficile de définir une tâche générique permettant de contrôler de manière pertinente l'ensemble des applications de dialogue.

– *Une métrique.* Il est nécessaire de comparer les systèmes objectivement et utilement. Ainsi la métrique devra être définie de façon à ce qu'elle soit la plus corrélée possible avec ce qu'on suppose être la qualité de la ou des applications potentielles. La qualité de la métrique est cruciale ; une métrique mal définie débouchera presque sûrement vers l'échec d'une campagne d'évaluation [3], car ses résultats ne seront pas interprétables.
Dans le domaine de la parole, nous avons bénéficié d'une mesure qui est à la fois simple à calculer, parfaitement intuitive et qui, bien qu'imparfaite, est très corrélée à de nombreuses applications. Il s'agit du **taux d'erreurs de mot** (*Word Error Rate - WER*). On définit le taux d'erreurs de mots *TEM* à partir d'une référence [4] constituée d'une suite de N mots, par rapport à une hypothèse de M mots contenant S substitutions, I insertions et E élisions ($M = N + I - E$) par :

$$TEM = \frac{S + I + E}{N} \times 100$$

Le nombre de substitutions, insertions, élisions est donné par alignement de l'hypothèse sur la référence. Pour tenir compte des variations jugées non pertinentes pour l'évaluation, on peut normaliser (c'est-à-dire redéfinir ce qu'on appelle un *mot*) l'hypothèse et la référence, et inclure également une liste d'équivalence de variantes orthographiques.

– *Des corpus.* A de nombreux égards on peut dire que le paradigme de l'évaluation est fondé sur le recueil, le développement, l'annotation et

3. Détecter l'échec d'une campagne d'évaluation n'est pas toujours aisé, car les organisateurs auront souvent intérêt à le dissimuler en partie ; cependant, une campagne dont les résultats ne sont repris que par les organisateurs ou les participants, qui plus est si elle n'a eu lieu qu'une seule fois, n'a pas les caractéristiques d'une évaluation réussie.

4. La référence est produite par un ou plusieurs humains qui utilisent toutes les connaissances à leur disposition pour constituer la meilleure transcription possible.

la distribution de corpus de grande taille. Ces corpus peuvent être de toutes natures, mais devront contribuer à apprendre et à évaluer des modèles spécifiques sur la tâche de contrôle, à l'aide de la métrique. Nous reviendrons dans la section 6.1.2 sur une description des corpus dans le cadre de l'évaluation comparative.

- *Un ou plusieurs sponsors*. Le développement des différents corpus, des logiciels de mesure, l'infrastructure à mettre en place pour distribuer les données, puis pour recueillir les sorties des systèmes et les évaluer, voire le financement des équipes de recherche afin qu'elles participent aux évaluations, demande de gros moyens humains et financiers. Ces moyens sont le plus souvent fournis par des agences gouvernementales, nationales ou internationales, afin d'assurer une certaines impartialité, en particulier pour les participants industriels.
- *Des laboratoires participants*. Le concept n'est différent d'une simple compétition que si les différents participants peuvent en retirer de la connaissance sur leur système, mais aussi sur les améliorations potentielles. Il n'est également utile que si ces progrès sont utiles à l'ensemble de la communauté. Si peu de laboratoires participent, ou si les meilleurs laboratoires ne participent pas, il existera une incertitude sur le fait que les méthodes utilisées dans l'évaluation sont les meilleures (à un instant donné) possibles.
- *Un calendrier*. Il s'agit de mettre d'accord les laboratoires participants et les organisateurs sur un planning qui comporte 4 phases : les deux premières coïncident souvent temporellement, *apprentissage*, au cours de laquelle les systèmes interagiront avec l'organisateur pour fixer la règle d'évaluation (protocole) et utiliseront les données d'apprentissage pour construire leur système ; *développement* au cours de laquelle les participants régleront les systèmes sur un corpus de la tâche, et valideront la procédure de mesure des performances ; la troisième phase est l'*évaluation*, où les participants reçoivent de manière synchrone les données d'évaluation et doivent rendre leurs résultats en un temps limité, celui-ci dépendant de la tâche, puis où l'organisateur produit les résultats de chaque participant en fonction des métriques choisies. La quatrième phase se déroule après l'évaluation proprement dite, et peut comporter une étape *d'adjudication* où les participants peuvent amender les résultats, par exemple en remettant en cause la référence, lors d'un dialogue avec les organisateurs, et la *confrontation des résultats*, souvent lors d'un workshop, avec une description détaillée des systèmes,

qui permettra, par une présentation par chaque participant des méthodes mises en œuvre, et du questionnement par les organisateurs et les autres participants, de faire émerger les méthodes intéressantes.

6.1.2 Les corpus

Les organisateurs fournissent des corpus qui sont en tout ou partie annotés, afin de fournir une référence (transcription manuelle dans le cas de la transcription de la parole) qui servira, selon le type de corpus, à apprendre, régler ou évaluer les systèmes. L'évaluation comparative repose sur la disponibilité de 3 types de corpus, qui ne doivent pas admettre d'intersection, afin de ne pas introduire de biais :

- *apprentissage* : des corpus de grande taille sont rassemblés (au besoin créés s'ils n'existent pas) et distribués aux acteurs du domaine qui ont manifesté le désir de participer à l'évaluation envisagée. Ce corpus est nécessaire pour l'estimation des paramètres des systèmes à apprentissage (de type markoviens ou neuronaux). Les corpus de parole doivent être transcrits au niveau du mot (voire du phone), afin de permettre l'apprentissage des modèles acoustiques. Il est également nécessaire de disposer de corpus de textes de transcriptions ou de textes écrits (journaux, livres, ...) afin d'apprendre les modèles de langage. Deux qualités fondamentales sont nécessaires au corpus d'apprentissage afin de permettre une bonne estimation : avoir des caractéristiques aussi proches que possible de celles de la tâche, et être de taille suffisante ; en pratique on essaiera d'avoir la taille la plus grande possible, étant données les contraintes matérielles, car on a pu observer que les systèmes à apprentissage réclamaient des corpus toujours plus grands.
- *développement* : un corpus est utilisé afin d'optimiser les systèmes. L'hypothèse sous-jacente à cette optimisation est que les caractéristiques de ce corpus de développement sont suffisamment semblables aux caractéristiques du corpus d'évaluation pour que l'optimisation sur le premier conduise à des caractéristiques proches de l'optimum sur le deuxième. Bien entendu, pour éviter tout biais, il est nécessaire que le corpus de développement soit distinct à la fois du corpus d'apprentissage et du corpus d'évaluation : un recouvrement avec une partie du corpus d'apprentissage reviendra à choisir de manière exagérée les composants ou modèles contenant cette partie, et les résultats sur le jeu de développement seront abusivement bons et non optimaux sur le jeu

d'évaluation ; un recouvrement avec une partie du corpus d'évaluation reviendra à un apprentissage indirect sur le corpus d'évaluation, et donc à des résultats d'évaluation abusivement bons (et non reproductibles). La contraposée de l'hypothèse de proximité entre le corpus d'évaluation et le corpus de développement est donc que si certains composants du système sont peu stables, ou si le corpus de développement admet une différence significative avec le corpus d'évaluation, il est possible d'observer des différences significatives mais peu pertinentes, de mesure entre les deux corpus.

– *évaluation* : ce corpus ne doit jamais intervenir dans une quelconque optimisation, afin d'éviter tout apprentissage qui introduirait un biais dans l'évaluation. Il est important, pour éviter tout biais dans les résultats, que la référence pour ce corpus soit la plus exempte possible de bruit ; à cette fin, une phase dite d'adjudication prend des fois place après la phase d'évaluation, afin de permettre aux participants et à l'évaluateur de corriger les erreurs présentes dans la référence. C'est à partir des mesures effectuées sur le résultat du traitement de ce corpus, en utilisant la ou les métriques définies pour l'évaluation, que sont définies les performances des systèmes.

6.1.3 Évolution des performances et des tâches

Les premières évaluations (de 1987 à 1992) ont été effectuées sur la tâche « *Resource Management* » (RM) avec un vocabulaire de 1000 mots [Pallett92], puis sur le corpus « Wall Street Journal » (WSJ) en limitant le vocabulaire [Pallett93] puis sans limite [Pallett95]. Parallèlement, une évaluation sur la reconnaissance de la parole conversationnelle par téléphone s'est déroulée [Chase98], ainsi qu'une évaluation sur la compréhension dans le cadre de la tâche d'information sur des vols aériens ATIS [Hemphil90]. À partir de 1995, a été proposée la tâche « Broadcast News » (BN) [Pallett97], puis dans les années 2000 une série de tâches à la fois plus proches de la réalité, et plus difficiles. La figure 6.1 permet de visualiser les différentes tâches et les progrès réalisés au cours des évaluations ARPA et NIST. On peut voir que les améliorations ont été constantes, mais que certaines tâches ne sont plus présentes après 2004 ; pour ces tâches, il a été jugé que le taux d'erreurs était « satisfaisant » ; pour ces tâches, les améliorations se mesurent au travers de tâches plus complexes, par exemple la traduction parole-parole : dans ce cas, on ne juge plus les améliorations à l'aide du taux d'erreurs,

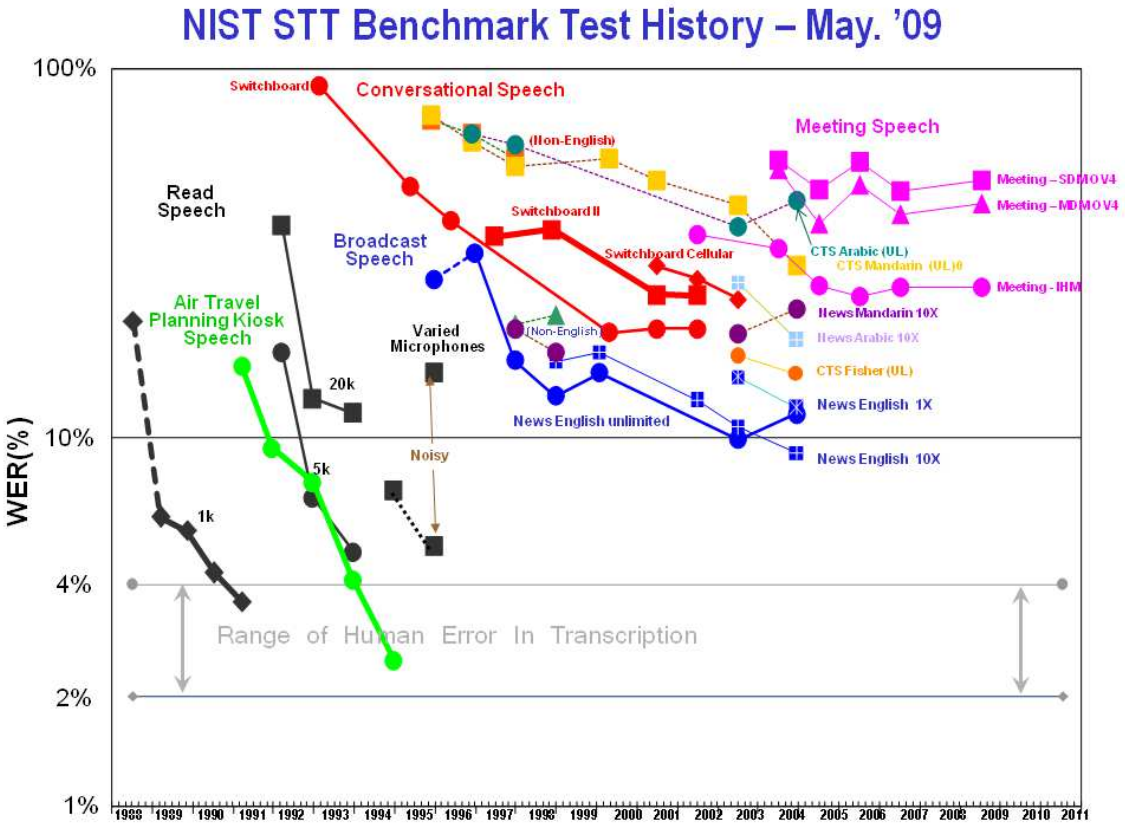


FIGURE 6.1 – Évolution des taux d'erreurs suivant les tâches et les années lors des différentes campagnes d'évaluation organisées par ARPA et NIST (http ://www.itl.nist.gov/)

mais à l'aide des mesures utilisées pour la traduction (BLEU [Papineni02], HTER [Snover06]). On peut voir également que pour certaines tâches complexes, comme la transcription de réunions, réduire le taux d'erreurs s'avère très difficile, et que pour d'autres tâches comme par exemple la parole conversationnelle, des contraintes ont été introduites (par exemple retirer la parole non-native, puis contraindre à de la parole entre personnes ne se connaissant pas a priori) pour faire baisser significativement les taux d'erreurs.

6.1.4 Les acteurs de la mise en place du paradigme de l'évaluation

Aux États-Unis, la mise en place du paradigme de l'évaluation comparative s'est faite à la suite du programme ARPA-SUR, avec des acteurs majeurs comme ARPA, le NIST (Dave Pallett) et le LDC (Mark Liberman). L'effort d'ouverture des évaluations ARPA aux acteurs européens, a porté ses fruits par l'organisation de campagnes d'évaluations en Europe par exemple SQALE [Young97], AUPELF [Dolmazon97], suivies par de multiples évaluations dans de très nombreux pays. Parmi les acteurs majeurs qui ont structuré l'évaluation en Europe, nous pouvons citer Joseph Mariani, qui a œuvré aussi bien au niveau national qu'européen, et qui a mis en place avec Khalid Choukri l'agence ELRA, ainsi que la conférence LREC (Language Resources and Evaluation Conference). Plus récemment, d'autres acteurs sont intervenus, comme par exemple la DGA en France (Edouard Geoffrois), équivalent de la DARPA aux États-Unis, qui a organisé en tant que sponsor plusieurs évaluations en France, dans un grand nombre de domaines : ESTER [Gravier04], ESTER2 [Galliano09], Techno-Vision/RIMES, REPERE...

6.1.5 Le statut actuel de l'évaluation comparative

L'évaluation comparative telle qu'introduite à la section 6.1.1 a été discuté par de nombreux auteurs ; dernièrement, quelques acteurs du domaine ont ajouté des précisions sur les avantages et les difficultés de l'évaluation comparative (par exemple pour les plus récents [Geoffrois08, Geoffrois09, Liberman10b, Pianta10, Gonzalo10]). En particulier, [Geoffrois08] spécifie pourquoi il est nécessaire d'organiser des campagnes d'évaluation en traitement de la langue :

– L'évaluation porte sur un ensemble que l'on peut qualifier d'infini, le langage ; cela implique que les résultats sont très dépendants des don-

nées d'évaluation et que donc ceux-ci doivent être diffusés pour permettre la reproductibilité des résultats.
- La référence est produite par un humain ; aucune évaluation ne peut être complètement automatique.
- Les systèmes évalués sont des système qui apprennent, ce qui implique que la connaissance des données d'évaluation va influencer le système.

A partir de ces 3 contraintes, découle la nécessité d'avoir un tiers qui se charge de l'organisation, que l'évaluation soit synchrone pour tous les systèmes. Dans [Geoffrois09], il est ajouté que la caractéristique des systèmes est qu'ils agissent sur des données non structurées, pour lesquelles il n'est pas possible d'avoir un accès direct, analytique à la « fonction » qui transforme les entrées du système en sorties observables ; il faut donc construire un modèle paramétrique du monde, obtenu par apprentissage, qui sera évalué et amélioré par l'utilisation du paradigme de l'évaluation.

[Pianta10] met l'accent sur la différence entre l'approche américaine, où la plupart des évaluations sont décidées par des agences gouvernementales (NIST, ARPA), et l'approche « européenne » où l'évaluation est organisée par la communauté elle-même ou au sein de projets européens précis, ce qui implique beaucoup de défauts, parmi lesquels on peut mettre en avant le manque de stratégie à long terme, de continuité [5]. Il ajoute que, par le fait même que nous avons en Europe un grand nombre de langues, il devrait être encore plus nécessaire de factoriser les efforts, et d'avoir une coordination au niveau européen.

[Liberman10a, Liberman10b], en se fondant sur une approche historique de l'évaluation, montre que l'évaluation a changé fondamentalement la manière dont on peut faire de la science, et que l' « ingénierie » qui a permis l'accès à de grandes bases de données est équivalente à l'invention du télescope et du microscope « linguistique ». Cet instrument permet au linguiste d'explorer et de répondre à des questions, ce qui est le point de départ à une science empirique. Il rejoint l'analyse de [Habert06], dans « Portrait de linguiste(s) à l'instrument. » (voir section 6.2.2), dans lequel B. Habert plaide que les outils automatiques de traitement des corpus (ici textuels) sont vus par les linguistes comme des instruments qui permettent d'augmenter la connaissance.

5. En France, nous avons eu la chance, grâce aux agences nationales, et en premier lieu ces dernières années la DGA, de pouvoir organiser régulièrement des évaluations depuis AUPELF [Dolmazon97] à la nouvelle évaluation ETAPE, qui aura lieu en 2012.

Nombreux sont les auteurs qui ont décrits les caractéristiques et les bénéfices de l'évaluation comparative ; on se reportera en particulier aux auteurs précités pour un résumé de ceux-ci. Cependant, certains aspects de l'évaluation des systèmes sont peu abordés, voire niés. Dans la vision classique de l'évaluation comparative, le but est de dégager les meilleures méthodes, ce qu'on peut résumer par la *certification des méthodes* : lors d'une évaluation, les meilleurs systèmes sont décrits précisément, les méthodes qui semblent intéressantes sont testées par les autres laboratoires lors de l'évaluation suivante. Dans le schéma classique de l'évaluation comparative, il y a donc *itération*. Or, pour de nombreuses évaluations qui ont été jugées utiles par la communauté on observe une activité sporadique, voire une seule occurrence. Au delà de la certification de méthodes, une évaluation est :

- un *calibrage* de système en tant qu'instrument complexe (et non de ses composantes) par l'emploi d'une mesure objective et de la comparaison avec d'autres systèmes. Les systèmes évalués sont des systèmes très complexes ; si l'évaluation permet certaines fois de certifier des méthodes, ce qui est objectivement le résultat des workshops post-évaluation et des conférences et projets qui suivent, c'est que certains systèmes, sur certaines tâches montrent des capacités reconnues comme l' « état de l'art ». Ce résultat de base sera repris pour juger l'ensemble des systèmes sur la même tâche, pendant toute la période entre 2 évaluations sur la même tâche. Pour certaines tâches bien établies, ce calibrage débouche sur le prototypage de systèmes réutilisables au niveau industriel ou par des institutionnels.
- la *certification* des équipes qui sont à même de mettre au point un système performant et de mener à bien une évaluation, ce qui est une tâche lourde, nécessitant une équipe compétente. La communauté a ainsi une vision des acteurs « fiables », scientifiquement et expérimentalement parlant, et ce sceau de fiabilité servira à la fois lors de la soumission d'articles dans les conférences internationales et dans la soumission de projets. Cette « confiance » que l'on peut accorder à un système complet (c'est-à-dire laboratoire + système + équipe) se transposera à l'ensemble des résultats du laboratoire pendant un certain temps (au moins jusqu'à la prochaine évaluation), même sur des tâches non directement évaluées ; en ce sens il s'agit de certification d'une équipe.
- Le *partage d'intérêt* entre institutionnels et laboratoires : les institutionnels qui financent les évaluations acquièrent ainsi une vision de l'état de l'art et peuvent orienter la recherche sur un sujet donné ; les labo-

ratoires y trouvent des moyens, directs lors de l'évaluation ou indirects avec les projets qu'ils comptent obtenir en ayant montré de bons résultats lors de l'évaluation, et le cadre expérimental rigoureux qui permet la mise en place d'une réelle démarche scientifique.

6.2 Historique des critiques de l'approche dominante et les tentatives d'évolution

6.2.1 Les critiques

Nous avons parlé de la critique de J.R. Pierce (voir section 5), qui en 1969 dans une lettre au JASA, a attaqué frontalement la reconnaissance de la parole, en tant que discipline, mais plus encore les chercheurs dans ce domaine à cette époque. J.R. Pierce critique 2 aspects :

- le fait que la discipline apparaît comme pouvant attirer des moyens, mais qu'elle ne se posait pas la question de ce qu'elle était exactement, quelle était sa finalité (« It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish. »)
- le fait que les chercheurs ne procédaient pas selon un schéma scientifique (« Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. »)

L'utilisation de la modélisation statistique, et de l'évaluation comparative de systèmes a permis, de faire rentrer notre discipline dans un démarche scientifique, et donc de lever l'hypothèque posée par la 2ème critique de Pierce. Peu d'auteurs se sont attaqués à la 1ère hypothèque, en essayant de définir formellement le « pourquoi » de la recherche en reconnaissance, et la plupart du temps pour en arriver à des critiques. Je vais détailler certaines critiques qui ont retenu mon attention [6].

Une critique a été formulée par Stephen E. Levinson[Levinson95] en 1994 ; dans ce papier, « Speech Recognition Technology : a Critique », S.E. Levinson critique la voie optimiste d'un progrès incrémental dans le domaine. Il doute que des applications réellement profitables soient mises sur le marché, et remet en cause le terme « *paradigm shift* » utilisé par [Makhoul95] parce

6. cette partie n'a pas vocation à être exhaustive en ce qui concerne les nombreuses critiques formulées par des chercheurs.

qu'il ne mène pas à une révolution scientifique. Levinson fait le lien avec les découvertes empiriques de Copernic, qui deviennent un « paradigm shift » selon lui, lors de la théorisation de Newton. Cette vision épistémologique d'une science qui n'avancerait que grâce à la seule jambe de la théorie, celle de l'empirisme ne pouvant être qualifié de scientifique est commune. Mais elle n'est pas, loin de là, une vue dominante en épistémologie : la science marche sur les deux jambes de la théorie et de l'empirisme, l'une prenant la place de l'autre quand cela devient nécessaire/possible.

Comme cela a été soulevé dans une tribune [Norvig11] publiée par Peter Norvig sur les vues de Chomsky en ce qui concerne la science et la recherche en langue, la science actuelle, même dans les sciences dites « dures » (biologie, physique) est constituée à 95% de science empirique. Dans ces débats autour des méthodes empiriques en traitement du langage, et en particulier de la reconnaissance automatique de la parole, je pense qu'on se méprend sur le fait que ces méthodes ne sont pas « explicatives », et donc pas scientifiques (au sens d'une théorie réfutable). En effet, le domaine du traitement de la parole est passé d'une époque (pré-datant 1985) où le traitement de la parole était un sous-domaine de l'intelligence artificielle et où le but était une modélisation directe de l'humain, à l'époque présente où l'approche empirique fondée sur une modélisation statistique occupe tout l'espace ; devant la faillite (pratique) des approches issues de l'Intelligence Artificielle, on s'est rabattu sur le moyen de modéliser les observations de ce que produit l'humain. Ce but est beaucoup plus modeste, mais n'est pas incompatible avec le fait de produire *à terme* un modèle explicatif du langage. Il impose juste que ce modèle explicatif soit au moins aussi performant pour expliquer les observations passées (contenues dans les corpus) que la modélisation statistique actuelle du langage.

Une autre critique est apparue en 1996, présentée par Hervé Bourlard [Bourlard96]. Cet article critiquait la façon d'utiliser l'évaluation comparative, au prétexte que cette méthode amenait à ne favoriser que les techniques fournissant des bénéfices à court terme, et donc ne conduisait qu'à explorer des méthodes simples ; l'argument est que les méthodes pouvant être plus optimales que les méthodes actuelles sont « forcément » complexes et donc difficiles à mettre en œuvre, et doivent « forcément » amener à obtenir des résultats moins intéressants à court terme. Cet article a été fort critiqué en son temps, en particulier parce que les méthodes potentiellement intéressantes qu'il présentait, n'était finalement pas si intéressantes. Sur le fond, le contenu du papier n'apportait pas une démonstration qu'explorer les méthodes amenant de moins bons résultats aurait à long terme un effet bénéfique : l'espace

des solutions apportant une dégradation aux systèmes de transcription est particulièrement vaste, et sans indicateur objectif comment savoir si l'exploration d'une méthode apportant un gain négatif doit être poursuivie, à part la « croyance » en cette méthode. Une tentative de reprise de cette hypothèse de dépassement du minimum local, par l'exploration de techniques dégradant les résultats a été apportée par la proposition d'un journal par H. Hermansky et N. Morgan (cosignataires de l'article de Bourlard), le« Journal of Negative results in Speech and Audio Sciences » [Hermansky04]. Ce journal, n'a pas eu beaucoup de succès pour des raisons assez simples :

- beaucoup de causes peuvent expliquer un résultat négatif,
- en particulier publier un résultat négatif, sauf pour quelques rares personnes d'un renom très important, peut amener à une mise en cause de la manière dont il a fait ses recherches.
- inversement si la personne en question est crédible, un résultat négatif peut au contraire amener à ne pas donner de fonds pour développer cette méthode, c'est-à-dire l'effet inverse de ce que voulait les promoteurs de ce journal.

La critique d'une approche d'évaluation comparative recherchant naturellement les améliorations locales, au sein d'évaluations encadrées, rejoint une critique récurrente de l'évaluation comparative : elle bride la créativité, en ne s'occupant que de sujets bien balisés. D'autre part certaines tâches ne sont pas simples à évaluer (voire impossible à évaluer), ce qui conduit à deux phénomènes indésirables : le développement d'évaluations de taille réduite sur des tâches mal définies et/ou sans métrique adéquate, ou l'abandon de certaines voies de recherche parce qu'elles ne conduisent pas à court terme à des évaluations objectives.

Roger K. Moore a également critiqué l'approche dominante [Moore05a, Moore05b] ; dans ces articles, il est fait état d'une limitation des applications en parole, fondée sur des contraintes intrinsèques de l'approche de modélisation probabilistes utilisant l'apprentissage à partir de données : en effet, par extrapolation des courbes des tailles des bases de données en fonction des performances des systèmes, R.K. Moore conclut à la nécessité d'augmenter de manière démesurée les corpus afin d'amener les systèmes à des performances comparables à celles de l'être humain. La conclusion est alors qu'un changement radical de paradigme est nécessaire : il suggère l'introduction d'une approche complètement différente, la « Cognitive Informatics ». Cette critique doit être pondérée : la connaissance des performances comparées des humains et des machines reposent sur des études qui maintenant datent

de quelques années (1997 pour [Lippmann97]) ; pour de nombreuses applications, de nouvelles études perceptives permettraient d'avoir une meilleure vision de l'état comparatif des systèmes actuels et des humains. De plus, pour de nombreuses applications, les systèmes, même moins performants que les humains, permettent l'exploration de quantités de données inatteignables par les humains. Même en reprenant le même diagnostic contenu dans [Moore05a, Moore05b], on peut arriver à une conclusion différente. En effet, plutôt que de rejeter l'ensemble des approches qui ont montré une réelle efficacité opératoire ces dernières années, on peut penser que l'avenir est à une *meilleure* utilisation des données, même dans le cadre d'une augmentation de celles-ci. L'utilisation des méthodes à partir des données n'est pas incompatible avec le développement d'autres méthodes. Soulignons néanmoins l'approche pionnière de R.K. Moore, qui a très tôt fait le parallèle entre les systèmes de reconnaissance (Automatic Speech recognition, ASR) et le système humain (Human Speech Recognition, HSR) (par exemple [Moore01]). Ce parallèle est très prometteur (voir section 6.2.2), en particulier avec le développement de l'imagerie médicale qui permet d'avoir un accès à ce qui se passe dans le cerveau. Cependant, le développement d'une théorie unifiée, capable d'expliquer et de prédire le comportement conjoint des systèmes automatique et humain [Moore05a], nécessite encore beaucoup de travail expérimental afin de nous permettre d'avoir plus de connaissances sur ces deux « systèmes ».

6.2.2 Les tentatives d'évolution

Apport mutuel des modèles statistiques aux modèles de perception humaine

Parmi les personnalités qui ont rapproché les deux reconnaissances, par les humains et par la machine, nous citerons en premier lieu R.K Moore et M.A. Huckvale. M.A. Huckvale a très tôt introduit la position alors iconoclaste que les systèmes de reconnaissance pouvait offrir une vue nouvelle pour les sciences cognitives, et en particulier pour la reconnaissance de mots par les humains [Huckvale96, Huckvale97, Huckvale98]. Il suggère que les sciences cognitives dépassent les réticences provenant de la divergence chomskienne, et qu'il est nécessaire d'effectuer une « re-convergence » entre les deux

communautés de la reconnaissance automatique et des sciences cognitives [7]. Il propose en particulier d'étudier les systèmes de reconnaissance comme s'ils étaient des humains, et réciproquement d'étudier les humains avec les mêmes outils objectifs que ceux utilisés pour étudier les systèmes automatiques. Ce programme s'est étendu à la synthèse de la parole [Huckvale02], et la production d'un modèle automatique, fondé sur des connaissances de ce que fait l'humain [Huckvale01].

Dans [Moore01], un parallèle très complet est fait entre les buts des études de la reconnaissance par des humains et des machines, même si les communautés qui s'occupent de ces deux thèmes de recherche sont très largement différentes. Il est suggéré de plus que la séparation en sous-problèmes très distincts dans le cas de la reconnaissance par des humains est un facteur néfaste pour la mise au point d'un modèle intégrant les diverses connaissances. Enfin, nous avons vu (section 6.2.1) que R.K. Moore aboutit dans [Moore05a] à la conclusion qu'il est nécessaire de développer une nouvelle science transdisciplinaire, l'Informatique Cognitive, qui aurait pour but de développer une théorie unifiée capable d'expliquer et de prédire le comportement du traitement du langage par les humains et par les machines.

[Dusan05] aborde le même problème de l'apport possible de ce que l'on sait en perception humaine, aux systèmes de reconnaissance, en prenant l'exemple des indices non-linguistiques (ou para-linguistiques) qui sont important pour la perception humaine et ne sont pour l'instant que peu pris en compte dans les systèmes de reconnaissance. Dans la même conférence O. Scharenborg [Scharenborg05b] adopte un point de vue plus intéressant, et qui rejoint le point de vue de M. Huckvale : la reconnaissance humaine peut apporter à la reconnaissance automatique, mais réciproquement, la reconnaissance automatique en tant que modèle computationnel de la reconnaissance de mot, peut être utilisé comme base pour un modèle computationnel de la reconnaissance de mot par les humains. Dans [Scharenborg05a, Scharenborg05b] est fait le parallèle entre reconnaissance humaine (Human Speech recognition, HSR) et reconnaissance automatique (Automatic Speech Recognition, ASR) et le développement d'un modèle computationnel de la reconnaissance de mot par un humain, SpeM ; inversement, en utilisant SpeM, [tenBosch05] explore les relations entre les scores acoustiques et les distances symboliques, dans l'exploration d'un graphe dans un système de reconnais-

7. On peut remarquer que ce plaidoyer est pionnier en particulier de celui de R.K. Moore sur le développement d'une « Cognitive Informatics » [Moore05a]

sance.

Des travaux de ces différents chercheurs, on peut conclure que le parallèle entre reconnaissance par les humains et par les machines, et cela même si ce domaine est encore peu exploré (voir par exemple [Vasilescu09]) est porteur d'un grand potentiel.

Linguistique à l'instrument

Il s'agit ici de l'une des évolutions récentes parmi les plus importantes en ce qui concerne les applications des systèmes automatiques : considérer les outils de traitement automatique comme des *instruments* permettant d'accéder aux très grandes quantités de données (de toutes sortes, et en particulier de parole), afin de pouvoir en extraire une connaissance phonétique, sociolinguistique, lexicale, syntaxique, ...

Cette évolution fait écho au « Portrait de linguiste(s) à l'instrument », dans lequel B. Habert [Habert06] fait état d'une évolution de la linguistique, qui jusqu'alors était dominée par l'approche générativiste, et déniait l'intérêt d'instrument. B. Habert introduit les différences en épistémologie entre *instrument* (annotation, transcription), *outil*, qui sera réservé aux logiciels multi-usage, non spécifiquement orientés vers le traitement des données langagières, et *dispositif expérimental*, qui est la forme instable, en phase de développement d'un instrument (« On pourrait dire qu'un instrument, c'est un dispositif expérimental qui a réussi »). Les instruments sont nécessairement imparfaits, et « les scories rencontrées relèvent de deux ordres d'explication : les limitations des techniques mises en place, d'une part, les incertitudes des jugements humains correspondant à ces classifications d'autre part. ». Mais les observations produites par les instruments sont stables : ils produisent une sortie reproductible et prévisible. B. Habert introduit enfin deux idées clés : les instruments permettent de voir de nouveaux phénomènes, ce sont des outils de perception ; il est nécessaire d'adapter les données aux instruments : un instrument est un capteur imparfait, qui sert à prélever une information, et il est donc nécessaire de « voir » avec cet instrument des données qui peuvent être visibles, et pour lesquelles la précision de l'instrument permet d'extraire une information pertinente.

Le workshop récent sur « New Tools and Methods for Very-Large-Scale Phonetics Research » à l'université de Pennsylvanie en janvier 2011, avait comme ambition de mettre en avant les études portant sur l'utilisation des méthodes automatiques pour les études phonétiques, et les études portant sur

de grands corpus. Ce workshop, organisé par Mark Liberman, faisait écho à sa présentation « The Future of Computational Linguistics : or, What would Antonio Zampolli do ? » au cours de la remise du prix Antonio Zampolli durant LREC 2010 [Liberman10a]. Dans cette présentation, M. Liberman fait le parallèle entre la situation actuelle, et l'état de la science en 1610 « Hypothesis : 2010 is like 1610 (. . .) We've invented the linguistic telescope and microscope (. . .) We can observe linguistic patterns ». Il montre que les sciences du langage peuvent être un paradigme des « e-Sciences », fondée sur l'utilisation intensive des ordinateurs, utilisant des bases de données de très grande dimension, dans un environnement hautement distribué.

Les systèmes automatiques de traitement de la parole (reconnaissance du locuteur, segmentation audio, transcription orthographique et phonémique) peuvent être considérés comme des instruments capables d'explorer des quantités de données jusqu'alors inatteignables, et de faire émerger des problématiques ou de valider des hypothèses linguistiques ; cette utilisation nécessite cependant un rapprochement des différentes communautés, car les systèmes automatiques, comme l'a souligné B. Habert, sont des *instruments* et non des *outils* : ils nécessitent une mise au point, des développement spécifiques, en un mot un « savoir-faire », qui n'est pas réductible à la mise à disposition des communautés linguistiques de « boîtes à outils ».

Chapitre 7

Propositions

7.1 Introduction

Dans les sections précédentes, j'ai présenté certains travaux, sur l'évaluation des systèmes (section 6.1.5), sur les critiques du mode actuel de développement des systèmes de traitement du langage (section 6.2.1). J'ai ensuite présenté ce que ces avancées en traitement automatique et ces critiques ont amené comme propositions pour les sciences du langage : soit une inflection nette de l'approche en traitement automatique, par exemple par la convergence avec les sciences cognitives (section 6.2.2), soit une extension de cette approche, en utilisant les systèmes automatiques comme des instruments pour explorer les corpus et découvrir de nouvelles propriétés (section 6.2.2).

Je vais aborder à présent dans ce chapitre une vue plus personnelle de ce que ces travaux m'ont inspiré comme voies de recherche potentielle, ou de structuration de la recherche, en particulier pour aller vers une science empirique du traitement de la parole. Je commencerai par définir ce que peut être l'observable de cette science, en tant qu'extension de la notion de corpus, et ce que cela implique pour les systèmes statistiques. Je présenterai ensuite en quoi l'analyse des erreurs est intéressante par rapport à cette définition de l'observable, et introduirai une typologie des erreurs en fonction de leur capacités à être réduites en fonction de méthodes existantes ou à découvrir. Je donnerai quelques idées sur ce que la mise en place d'une telle démarche empirique impliquerait, dans un monde idéal, comme structuration des thèmes au sein des conférences du domaine. J'aborderai enfin une proposition permettant de mettre en œuvre une démarche empirique fructueuse à la fois

pour le domaine du traitement automatiques et pour les sciences du langage en général, c'est-à-dire la mise en place de structures pérennes autour d'outils de l'état de l'art et des équipes pouvant les manipuler, sur le modèle du CERN dans le domaine de la physique des particules.

7.2 Le statut du corpus

7.2.1 Qu'est-ce qu'un corpus

La définition que l'on trouve dans le dictionnaire est ambiguë : une acception du terme, utilisée en science humaine et en philosophie, donne à *corpus* celui d'un recueil réunissant, en vue de leur étude scientifique, la *totalité* des documents disponibles d'un genre donné ; une deuxième acception, plus vague, est celle utilisée en linguistique : « Ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique ». Pendant longtemps, en traitement du langage naturel, on a introduit une distinction forte entre « collection », qui contient un ensemble de textes recueillis, sans sélection a priori, d'une ou plusieurs sources (par exemple la collection du journal « Le Monde » de 1987 à nos jours, qui est largement utilisé aussi bien en traitement du langage naturel qu'en traitement de la parole) et « corpus », qui doit contenir des échantillons de différents registres, de différentes sources, de différentes époques (typiquement le Brown Corpus, construit dans les années 60 par H. Kucera et W. N. Francis). Cette sélection a priori pose bien des problèmes, car elle suppose déjà un modèle (implicite ou explicite) de ce que l'on veut trouver ; la sélection d'échantillons suppose que ceux-ci soient « représentatifs », mais de quelle réalité ? Se pose également la question de la taille, qui dépend du type de phénomène que l'on recherche. De nos jours, le terme « corpus » est utilisé indifféremment pour les deux types d'objets (corpus et collection), et l'on qualifie de corpus aussi bien le British National Corpus (BNC) qui contient 100 millions de mots recueillis selon un schéma similaire à celui du Brown Corpus, que le English GigaWord, qui contient une collection longitudinale de textes en provenance de 7 agences de presse, totalisant 4 milliards de mots.

7.2.2 Pourquoi un observable ?

L'utilisation des corpus s'est imposée historiquement dans le domaine du traitement de la parole, (voir section 6.1.3) à la suite des évaluations ARPA et du développement et de la distribution de corpus (de textes et de parole) tout d'abord par le LDC [1], puis par d'autres acteurs, dont ELRA [2] en Europe. L'utilisation conjointe des évaluations et des corpus a montré une efficacité réelle, tant par les retombées scientifiques en linguistique de corpus (voir section 6.2.2), que dans le domaine des technologies et des applications (voir section 6.1.5). Le développement d'une science empirique du langage nécessite de définir précisément ce que l'on observe, ce que l'on mesure, que l'on peut résumer par « quel est son observable ».

Mon explication des progrès passés est l'utilisation non-explicite d'un paradigme empirique, qui a donc à voir avec l'utilisation de corpus (d'apprentissage, de développement, d'évaluation), et d'évaluation utilisant ces corpus. Dans ce paradigme, ce qu'on observe n'est pas un fait isolé de la langue (par exemple une phrase particulière énoncée par un locuteur particulier dans un contexte défini), mais des mesures effectuées sur un ensemble de faits (en l'occurrence un corpus d'évaluation). Ce que l'on mesure objectivement est bien sûr dans le cas de la transcription de la parole des taux d'erreur, mais c'est également l'écart entre la modélisation obtenue sur le corpus d'apprentissage, et les performances sur le corpus de développement et d'évaluation. On mesure cet écart en terme de performance absolue, mais également de comportement qualitatif et quantitatif sur différents sous-ensembles de ces corpus d'évaluation. Ces évaluations sont pertinentes par rapport à un certain nombre de paramètres implicites ou explicites que l'on prête aux données langagières que sont supposés traiter les systèmes automatiques (parole lue ou spontanée, spécifique à un tâche donnée, bande large ou étroite, locuteurs professionnels ou amateurs, etc...).

Le corpus est donc ici un échantillon représentatif d'un ensemble de données langagières délimitées par des plages de valeurs sur certains paramètres. J'appellerai cet ensemble *espace langagier isoparamétrique* [3] (ELI), qui est

1. Linguistic Data Consortium, `http://www.ldc.upenn.edu/`
2. European Language Resources Association, `http://www.elra.info/`
3. cette tentative succincte de formalisation « épistémologique » emprunte un certain nombre de termes à la théorie des observables en mécanique quantique, mais cet emprunt est purement lexical : je ne prétends pas que l'espace langagier soit un espace de Hilbert, ni que l'observable est un opérateur linéaire hermitien...

donc notre observable, où les mesures sur l'observable sont les sorties de systèmes automatiques sur un échantillon d'un ELI.

Cet espace est souvent défini implicitement par un corpus d'apprentissage, supposé contenir un grand nombre d'exemples de certains faits langagiers dans la plage de variation des paramètres le définissant. Les corpus (d'apprentissage, de développement et d'évaluation) sont donc dans ce formalisme des projections dans le monde réel d'un ensemble de faits de langage, délimité par un jeu de paramètres, certains explicites (choisis explicitement lors de la constitution du corpus, par exemple parole lue, par des locuteurs professionnels de telle tranche d'âge, etc) ou implicites (présents implicitement dans les corpus, de par le choix particulier d'une instance de l'espace langagier, par exemple les émissions de France Inter de 7h à 9h pendant 5 jours).

Je considère que l'utilisation du paradigme empirique est non-explicite, car le travail d'explicitation des paramètres définissant l'espace langagier que l'on explore n'est la plupart du temps pas fait ou de manière très partielle. Un travail fondamental (voir section 7.2.3) doit être d'expliciter ces paramètres.

J'ai dit que la mesure sur l'observable est la sortie de systèmes automatiques sur un corpus (le corpus dit de d'évaluation dans la dénomination usuelle). Ce que constitue exactement la mesure dépend de l'usage que l'on en fait. Dans l'approche classique dit du « paradigme de l'évaluation », on regarde principalement le meilleur taux d'erreur sur ce corpus : la mesure est alors la performance attendue sur l'espace langagier isoparamétrique en fonction de l'implémentation de nouvelles techniques ou l'amélioration d'anciennes (voir section 6.1.1). Mais pour le développement d'une science empirique, l'étude sur l'observable peut se faire à l'aide d'autres mesures :

1. L'écart entre les performances sur le corpus de développement et le corpus d'évaluation. Cela peut être un problème de taille de corpus d'apprentissage : si le corpus d'apprentissage est trop petit, il risque de ne pas être représentatif des événements du corpus de développement et d'évaluation, il convient donc de l'agrandir. Cela peut être aussi le diagnostic d'un problème au niveau de la définition des paramètres de l'ELI : le corpus est mal défini, et contient des paramètres importants qui ne sont pas explicités . Un tel cas peut être visualisé dans la figure 7.1 : en fonction des paramètres que l'on connaît on choisit les parties apprentissage/développement/évaluation afin qu'ils soient proches les uns des autres (partie gauche de la figure), mais si en

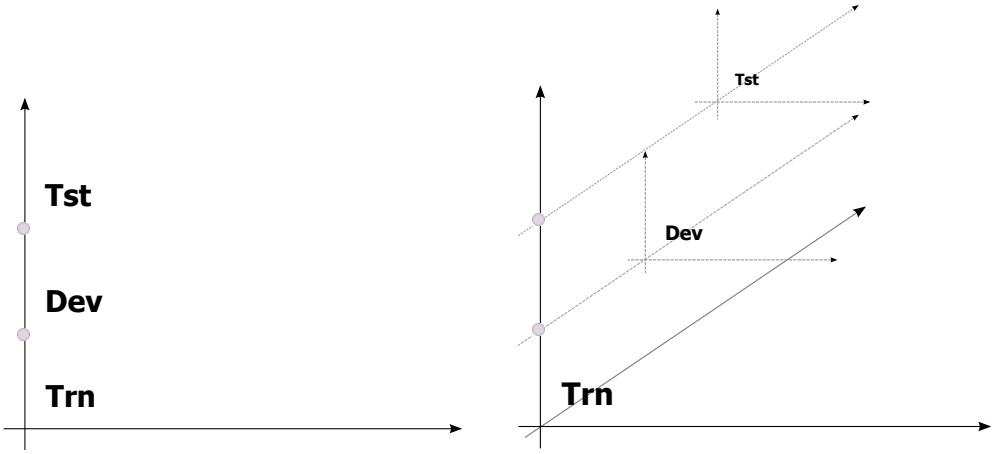


FIGURE 7.1 – Visualisation de la proximité supposée des corpus de *train, dev et test* dans un ELI implicite (partie gauche), qui sont de fait éloignés par non prise en compte d'un paramètre important (partie droite)

fait un paramètre important n'a pas été pris en compte, ces différents corpus sont en fait éloignés, et les mesures ne sont pas stable (partie droite de la figure). Par exemple, si on choisit comme ELI l'espace des conversations en français, entre personnes du même sexe se connaissant (même famille ou amis), mais que l'on ne spécifie pas le lieu de recueil, on se retrouve avec des données qui contiennent des caractéristiques et donc des résultats de reconnaissance complètement différents si les données sont recueillies en banlieue (qui contiendront de l'argot des mots de langue étrangère, une prosodie très spéciale, des sujets très particuliers...) ou en centre ville. Dans ce cas, le lieu de recueil est un paramètre qui surpasse d'autres paramètres (par exemple le sexe).

2. L'évolution des performances lors de l'accroissement du corpus d'apprentissage au sein d'un ELI. Une stagnation doit nous orienter vers la recherche de nouveaux paramètres à expliciter, puisque la quantité de données n'agit plus sur la précision ; une régression nous oriente encore plus dans le fait qu'un paramètre important, implicite, existe, qui biaise les résultats ; une amélioration nous oriente sur le fait que nous restons dans les limites de l'ELI.
3. Les erreurs faites sur un ELI (voir section 7.3) ; celles-ci permettent

d'orienter vers un autre ELI (dans le cas d'erreurs en trop grand nombre, ou mal identifiées), ou d'orienter les recherches afin d'améliorer les systèmes.

7.2.3 Paramètres des modèles statistiques

Dans cette section, je vais préciser l'implication de cette notion d'observable, pour ce qui concerne en particulier les systèmes utilisant des modèles statistiques, c'est-à-dire pratiquement la totalité des systèmes publiés à l'heure actuelle. Je reprendrai donc les mêmes concepts que précédemment, en les appliquant au problème de la modélisation statistique.

En fonction de la notion d'observable définie dans la section précédente, et d'un point de vue épistémologique, la nature du traitement opéré par les systèmes automatiques, est de mettre en œuvre un modèle du monde, contenant l'ensemble des connaissances modélisables ; ces connaissances sont par exemple la nature des paramétrisations acoustiques utiles (MFCC, PLP), le type de jeu de phonèmes utilisé dans le dictionnaire de prononciation pour les modèles acoustiques, la taille du lexique utilisé dans le modèle de langage, etc. Mais le cœur de la connaissance du monde contenue dans les systèmes de traitement de la parole réside dans les modèles probabilistes ou plus généralement statistiques, fondés sur des corpus d'apprentissage et réglés sur des corpus de développement. Ces modèles expliquent au mieux les données observées [4], et permettent une certaine généralisation, au moins au sein de l'ELI, généralisation attestée par les performances sur des données d'évaluation.

Lors de l'apprentissage de modèles probabilistes (et plus généralement de modèles statistiques), certains paramètres sont modélisés de manière explicite, et d'autres de manière implicite. Par exemple, on a l'utilisation d'une modélisation explicite pour les modèles acoustiques, si l'on essaye de faire des modèles différents selon le type de phonème, son contexte plus ou moins étendu, le sexe du locuteur, la largeur de bande (téléphonique ou non), des accents, etc ; dans le cas de modèles de langage, on va par exemple faire des modèles séparés selon les sources, utiliser des contextes de 3, 4, 5 mots, en prenant en compte des parties du discours, ou en prenant en compte

4. Les techniques d'apprentissage statistique ont souvent pour but d'assurer une optimalité (d'après certaines mesures) de la capacité des modèles à expliquer les données d'apprentissage.

le contexte du document, du thème, des mots environnants. D'autres paramètres sont appris implicitement ; par exemple, si l'on ne distingue pas par sexe, mais que l'on utilise un grand nombre de gaussiennes, on peut penser que certaines représenteront cette distinction.

Lorsque l'on explicite certains paramètres, on apporte un a priori fort, qui peut revenir à partitionner le corpus d'apprentissage, ou à en limiter la variabilité. Par exemple, si l'on estime des modèles acoustiquesdépendants du sexe, cela revient à dire que utiliser comme a priori que couper le corpus d'apprentissage en 2 parties distinctes, la parole des hommes et la parole des femmes est avantageux par rapport à n'avoir qu'une seule classe, plus grosse, mais moins précise, contenant tous les locuteurs ; de même, si l'on apprend des modèles de langage différents selon les sources différentes. L'utilisation de parties du discours pour les modèles de langage ne segmente pas le corpus, mais suppose que les dites parties du discours sont pertinentes. Si la partition (dans le cas de modèles séparés) ou la structuration (dans le cas d'un simple a priori) ne sont pas pertinentes étant données la taille et la nature du corpus, on peut donc s'attendre à une perte ; cependant, même si l'explicitation n'est pas pertinente, mais que la taille du corpus implique une large redondance, on peut ne pas observer cette perte.

Parmi les problèmes récurrents lors de la mise au point d'un système, on retrouve donc deux questions qui sont corrélées :

- quels sont les paramètres que l'on doit expliciter ?
- la taille et la nature du corpus permettent-elles d'apprendre des modèles en fonction de ces paramètres explicites ?

Par rapport à la vision du corpus comme projection de l'observable que constitue l'ELI, il est crucial de connaître les paramètres importants, de savoir leur importance relative, et de connaître la taille de l'échantillon nécessaire pour les apprendre.

Afin de faire un parallèle caricatural, imaginons que l'on considère la langue comme un jet de dé à 6 faces. On pourra être tenté étant donné un corpus d'un grand nombre de jets, de voir si l'on retrouve des corrélations entre le résultat du jet, et la forme des coins, la structure de la table, la matière du dé, certains tirages précédents, etc... Mais, en première approximation, on ne trouvera que seul le nombre de faces et la forme du dé sont importants [5].

5. S'il existe d'autres corrélations, et que l'on est capable de les trouver, cela peut être un bon moyen de devenir riche ; ainsi ce joueur qui au XIXe siècle, en étudiant systéma-

Il est donc important, pour explorer les limites de l'espace isoparamétrique qui le définisse, d'émettre un certain nombre d'hypothèses sur les paramètres importants, puis de les tester. Dans le cadre d'un cycle d'augmentation de la taille d'un corpus-observable (c'est-à-dire en augmentant la taille du corpus en conservant les paramètres explicites fixes), on peut ainsi détecter certaines incohérences ou la présence de paramètres cachés, parce que les performances des systèmes stagnent ou même régressent ; en effet, un tel comportement vient soit (1) que l'on n'a pas spécifié de manière explicite des paramètres intrinsèques à la base (passage implicite → explicite), soit (2) que le corpus tel que défini par ces limites ne contient pas assez de paramètres implicites différents, et qu'il est donc important de changer la provenance du corpus (dans le même ELI) afin d'explorer une nouvelle dimension de l'ELI. Ce comportement des systèmes me conduit à préconiser une étude précise des erreurs (voir section 7.3) afin de mettre à jour de nouvelles dimensions intrinsèques du corpus (voire son éclatement en plusieurs sous-corpus) dans le cas (1). Dans le cas (2) où donc l'on peut considérer que les limites du corpus sont atteintes et où l'instrument d'observation que constitue le ou les systèmes est le plus précis, cela revient à spécifier quels sont les champs à explorer afin de concevoir de nouveaux modèles à incorporer aux systèmes afin d'améliorer les systèmes.

On voit que dans un tel schéma d'exploration des données, on ne produit pas directement un modèle explicite de la parole, mais que l'on permet son élaboration par une étude précise des paramètres qui régissent l'espace langagier, de leur importance relative et de leur lois de probabilité. Il faut bien souligner ici qu'un modèle explicite n'est en rien incompatible avec un modèle statistique ou probabiliste ; en particulier on peut penser à un modèle explicite qui permet d'expliquer les propriétés moyennes des paramètres sur un corpus-observable, mais qui doit être « convolué » avec les lois probabilistes apprises sur les données, pour « expliquer » les faits individuels contenus dans le corpus.

---

tiquement les fréquences de sortie des numéros d'une roulette particulière, en a déduit certains numéros qui avaient plus de probabilité de sortir, et gagna une fortune à Monte-Carlo ; depuis, les casinos vérifient très soigneusement la qualité de leur matériel...

7.3 Analyse d'erreurs

L'étude des erreurs est un domaine crucial, en particulier dans le paradigme d'une science empirique du traitement de la parole. Cette étude est multiforme ; elle couvre par exemple l'étude comparée des erreurs des systèmes et des humains [Vasilescu09], l'étude diagnostique des erreurs [Goldwater10]. Il faut souligner ici que l'étude des erreurs est intéressante quand il y a des erreurs en nombre assez faible pour que l'on puisse les classifier aisément : pour des tâches où le taux d'erreur est trop important, les différents types d'erreurs se chevauchent, interfèrent profondément, rendant toute tâche d'analyse très difficile. Aussi, l'étude des erreurs est intéressante en particulier sur des tâches et des corpus qui ne sont plus forcément en « 1ère ligne », c'est-à-dire qui ne font plus l'objet de développement intense afin de faire baisser les taux de manière drastique ; au contraire, dans la figure 6.1, les tâches intéressantes à explorer sont les tâches stables, que l'on a jugées « résolues », ce qui est le plus souvent faux, mais qui signifie que l'on a atteint les limites du paradigme de l'évaluation comparative, et que l'on s'attend à ce que les efforts en terme de développement simple (en particulier par augmentation de la taille du corpus d'apprentissage), n'apporteront qu'un gain faible alors que les résultats sont assez « satisfaisants ». C'est en particulier pour ces tâches que les erreurs résiduelles, qui semblent rétives aux modélisations classiques, sont intéressantes à étudier. D'un autre côté, nous avons souligné que les systèmes de traitement de la parole pouvaient (devaient) être utilisés comme instruments pour explorer les corpus (section 6.2.2), afin de tester certaines hypothèses, de découvrir certains faits linguistiques, phonétiques etc ; dans ce paradigme, l'étude des erreurs nous permet de mesurer la précision des systèmes comme instruments, selon les corpus.

Examiner les erreurs est utile pour découvrir les principales directions à explorer afin de déterminer de nouvelles techniques et à terme améliorer les performances des systèmes. L'analyse des erreurs peut s'effectuer à plusieurs niveaux : en classifiant les erreurs quantitativement et qualitativement, en utilisant une expertise linguistique (phonétique, lexicale, syntaxique, grammaticale, sémantique, pragmatique), en comparant les erreurs des systèmes avec les erreurs produites par des humains [Lippmann97, Vasilescu09] . . . A ce titre il est intéressant de ne pas regarder les erreurs uniquement d'un système par rapport à une référence absolue, ou de regarder les erreurs en tant que diagnostic (s'agit-il d'erreurs dues au modèle de langage, au modèle acoustique, à la segmentation, des erreurs de recherche, . . .) mais également de

définir une typologie en se plaçant dans l'optique de la *résolution* des erreurs.

Définir une typologie n'amène pas une résolution directe des problèmes, que ce soit pour les erreurs ou pour les langues ou les champignons, mais définir un certain nombre de types pour faciliter l'analyse est pertinent dans le cas où les phénomènes sont nombreux, complexes et intriqués, ce qui est bien sûr le cas des erreurs en traitement de la parole. Si nous voulons résoudre des problèmes, il s'agit de les identifier en séparant le problème global en sous-problèmes atteignables ; à ce titre un typologie des erreurs en fonction de leur solution est utile, car nous ne devons pas tenter de résoudre avec les outils méthodologiques existants des problèmes qui ne peuvent pas être résolus, ou qui ne sont pas des problèmes, ou encore qui sont d'une importance mineure.

Un système de traitement de la parole est un instrument complexe. De nombreuses stratégies différentes peuvent être mises en jeu, dans des agencements différents, avec des réglages tout aussi complexes et multiformes ; tout ceci conduit de fait à une grande variété de systèmes différents selon les laboratoires qui les ont conçus [6], ayant un comportement différent face à une même tâche, même à taux d'erreurs égal. Cette diversité est paradoxale, la totalité des systèmes de l'état de l'art étant fondés sur les mêmes principes, utilisant des modélisations très proches, apprises sur des bases de données partagées (au moins lors des évaluations officielles). De plus, Il existe des méthodes stables, utilisées par tous les systèmes, comme par exemple l'utilisation de modèles de phones en contexte, ou les méthodes d'adaptation du type MLLR [Leggetter95]. Une matérialisation de cette diversité est l'efficacité de la technique du Rover[Fiscus97], qui permet, par une combinaison des sorties de plusieurs systèmes différents, d'obtenir un système ayant significativement un taux d'erreurs plus faible [7]. Dans les systèmes actuels ces techniques et modèles sont autant de briques de Lego dont l'agencement différent conduit effectivement à des édifices différents. Afin de pouvoir explorer les limites des systèmes et en particulier des modélisations actuelles sur une tâche (ou plutôt sur un ELI), je préconise l'usage du *Rover Oracle* (RO), c'est-à-dire le choix oracle de la bonne solution si elle existe dans le graphe des différentes possibilités fournies par les différents systèmes ; le Rover Oracle nous offre

6. En fait, même pour un même système, par exemple CMU Sphinx, selon le laboratoire qui l'aura implémenté, on aura également une variabilité de comportement très importante.

7. D'autres techniques, comme l'utilisation d'adaptation croisée de systèmes permet d'obtenir des taux d'erreurs identiques à l'utilisation du ROVER, voire plus efficaces, lorsque uniquement 2 systèmes sont disponibles ; cependant, nous nous placerons ici dans le cas où de nombreuses (> 3) sorties de systèmes sont disponibles

FIGURE 7.2 – Visualisation de classe des erreurs irréductibles, vide dans la figure de gauche, et égale à l'intersection stricte des erreurs des 3 systèmes S1, S2 et S3.

ainsi une approximation de la limite supérieure de ce que peut atteindre un système de l'état de l'art, qui utiliserait de manière optimale les différentes modélisations présentes dans les différents systèmes, et ayant également le réglage optimal de ses différents paramètres.

En plus d'une indication sur la limite inférieure du taux d'erreurs sur un corpus donné, le Rover Oracle nous permet un accès à différentes classes d'erreurs intéressantes. La classe d'erreurs la plus intéressante pour le chercheur, et qui permet d'approximer la précision des systèmes de traitement en tant qu'instrument, est la classe des **erreurs irréductibles** (EI) (voir figure 7.2), définie par (en notation ensembliste) :

$$EI = ERR(RO)$$

où $ERR(RO)$ est l'ensemble des erreurs produites par le Rover Oracle. On obtient ainsi la classe des solutions correctes qu'aucun système n'avait envisagées. Une part de ces erreurs sont cependant du bruit, par exemple les erreurs dans la référence [8], mais cette notion de bruit doit être étendue à tous les phénomènes qui ne sont pas pertinents pour une tâche donnée, et qui donc ne doivent pas être considérés comme erreur ; par exemple, pour une tâche classique de transcription de la parole, il est d'usage de ne pas considérer que le fait de ne pas reconnaître une pause remplie (*euh* en français) doit être

8. Une partie de ces « erreurs » dans la référence ne sont pas uniquement des fautes, mais est aussi le reflet des incertitudes du jugement humain [Habert06] qui a conduit à la production de cette référence ; il est nécessaire alors d'avoir un mécanisme qui prenne en compte cette incertitude, sous la forme de références multiples.

pénalisé. On définit une version adaptée de la classe des erreurs irréductibles n'incluant pas ces erreurs non pertinentes :

$$EI' = ERR(RO) - Br(RO)$$

où $Br(RO)$ représente les erreurs de bruit produites par le système Rover Oracle. Une fois ce bruit retiré, on se trouve en face de toutes les erreurs réellement difficiles, et pour lesquelles une nouvelle modélisation serait nécessaire. Si pour une tâche donnée nous nous trouvons devant un taux EI' élevé, cela permet juste un diagnostic sur la tâche ; par exemple, en l'état actuel des systèmes, reconnaître simultanément ce que disent 2 personnes parlant en même temps avec un même niveau d'énergie reste inaccessible, et l'examen des erreurs ne nous apportera rien. Un taux EI' élevé doit donc conduire selon le cas à changer de tâche, ou à recueillir une grande quantité de données.... Si pour une tâche donnée, le taux EI' est faible, alors nous pouvons commencer à examiner les différentes classes d'erreurs en diagnostic (en utilisant par exemple une connaissance linguistique), afin de permettre à terme une amélioration. Il est important d'appuyer sur ce dernier point : examiner précisément les erreurs n'est possible que si le système est déjà performant, car sinon les différentes causes d'erreurs se superposeront et rendront tout diagnostic impossible.

Autre classe d'erreurs intéressantes, la classe des **erreurs atteignables** d'un système S ($EA(S)$) (voir figure 7.3). Sa définition est :

$$EA(S) = ERR(S) - ERR(RO)$$

De manière explicite, $EA(S)$ représente la classe des erreurs du système S qui ont pu être corrigées en utilisant le Rover Oracle. Cette classe est celle qui doit être examinée en commun avec les développeurs d'autres systèmes car, pour un système S, elle représente l'ensemble des erreurs pour lesquelles au moins un autre système a pu trouver la bonne solution. Typiquement, il s'agit des erreurs sur lesquelles se concentrent les différents sites pendant le workshop suivant une évaluation, essayant désespérément de comprendre exactement ce qui a été fait dans le meilleur système. Il serait bon également de répertorier les sous-classes de $EA(S)$ selon les systèmes qui ont été capable de trouver la bonne solution (voir figure 7.4) ; une telle information permet, couplée avec le description des systèmes, d'avoir une première indication sur quelle technique ou quelle donnée supplémentaire peut apporter

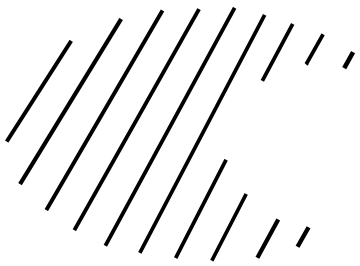


FIGURE 7.3 – Visualisation de la classe des erreurs atteignables du système S1, comme la classe des erreurs de S1 qui ne sont pas des erreurs pour les systèmes S2 et S3.

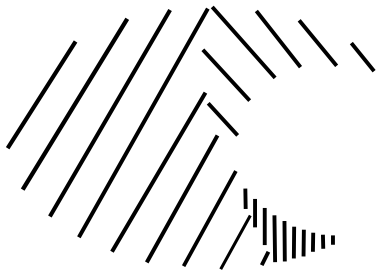


FIGURE 7.4 – Visualisation de la classe des erreurs atteignables du système S1, selon le système qui permet de les atteindre ; ainsi la partie hachurée jaune correspond aux erreurs de S1 qui ne sont des erreurs ni pour S2, ni pour S3.

FIGURE 7.5 – Visualisation de la zone de flou introduit par différents réglages fin des systèmes

un bénéfice sur telle catégorie d'erreurs ; cela permet également de répertorier, pour chaque système, quel autre site doit être interrogé pour obtenir des informations pertinentes pour réduire cette classe d'erreurs.

Les systèmes, à l'heure actuelle, sont des instruments très complexes du fait, entre autres, qu'ils contiennent un grand nombre de techniques de modélisation et d'adaptation. Ces techniques, bien que cohabitant dans un même schéma théorique (la modélisation statistique de la parole) interagissent de manière difficile à prédire ou à modéliser. En particulier, les systèmes nécessitent une phase de réglage d'un certain nombre de paramètres, propres au système en question et aux techniques mises en œuvre ; le plus souvent cette phase de réglage des différents paramètres est sous-optimale (on ne teste pas tous les paramètres dans toutes les valeurs). Cependant, en ce qui concerne l'étude des erreurs en toute rigueur, cette phase de réglage introduit une zone d'incertitude (limitée), de flou sur les erreurs d'un système (voir figure 7.5). Pour avoir accès à cette zone de flou, il faudrait pouvoir tester l'ensemble des paramètres pour la chaîne complète (pour le meilleur système, intégrant toutes les techniques) ce qui serait d'un coût rédhibitoire ; une expérience intéressante serait cependant de pouvoir estimer la surface de cette zone

d'incertitude due au réglage pour au moins un système sur quelques tâches, pour pouvoir estimer de manière fiable la précision des systèmes de traitement automatique de la parole en tant qu'instrument pour explorer les ELI (voir section 7.2.2), ou les corpus (voir section 6.2.2)

Les quelques idées que je viens d'exposer sur une typologie des erreurs en fonction des performances des systèmes est très succincte et préliminaire, et doit se coupler avec une étude précise et diagnostique, de chaque classe d'erreurs. En particulier, cette typologie est très dépendante des performances du meilleur système qui va évoluer au cours du temps ; si les performances évoluent, les frontières entre les différentes classes d'erreurs vont également évoluer. Autre facteur déterminant, le nombre d'erreurs : dans la mesure où j'ai fait l'hypothèse que l'étude des erreurs était intéressante là où il y en avait peu, cela oblige, si l'on veut pouvoir trouver des corrélations utiles à effectuer des analyses sur de grands corpus d'évaluation. 10 heures de parole, avec un taux d'erreurs de 5% produisent environ 5000 erreurs, ce qui est bien peu pour extraire des classes ; on peut donc extrapoler que ces analyses seront réellement fructueuses sur des corpus de l'ordre de 100 heures. Enfin la mise au point d'outils permettant de faire ce travail exploratoire sur une tâche (et une langue) donnée où suffisamment de systèmes sont disponibles devrait permettre de mieux planifier le travail de recherche sur ce problème.

7.4 Structuration de la production scientifique

Le développement des sciences du langage dans une démarche relevant authentiquement d'une science empirique est déjà visible, nous l'avons souligné, grâce à l'évaluation comparative et à l'utilisation des systèmes automatiques comme instrument. Cependant, un certain nombre de points freinent ce développement, parmi lesquels la manière dont les sujets de recherche sont abordés dans les conférences existantes du domaine [9].

Les articles scientifiques sont le moyen principal de mettre en exergue les connaissances que nous produisons ; de plus la production scientifique est un point essentiel dans le développement actuel d'une science, puisqu'elle constitue le moyen principal d'évaluation des chercheurs et des laboratoires de recherche. Les articles ont pour but de qualifier et de quantifier la connaissance que nous avons extraite des expériences, afin de la communiquer à l'ensemble

9. je rappelle que je me restreins volontairement au domaine du traitement de la parole, bien que certains parallèles existent avec l'ensemble des sciences du langage.

de la communauté, voire lui permettre de reproduire les expériences ayant produit cette connaissance.

Dans la plupart des sciences expérimentales, il faut bien se rendre compte du statut de *reproductibilité* des résultats : s'il est bien sûr nécessaire de pouvoir reproduire les résultats les plus marquants ou les plus originaux, afin de les valider définitivement, cette reproductibilité reste et doit rester théorique pour la plupart des résultats mineurs, qui sont le lot de 98% des articles. En effet, c'est parce que l'on se doit d'accorder un crédibilité aux résultats publiés *sans avoir à refaire les expériences*, qu'une science expérimentale atteint un degré de maturité qui lui permet d'avancer efficacement. Cette confiance, cette crédibilité, ne sont pas (c'est un euphémisme) toujours présentes pour les articles publiés dans les conférences majeures, voire même pour les articles de journaux. Paradoxalement, on peut même dire qu'un domaine scientifique où la plupart des expériences doivent être reproduites à plusieurs endroits pour être validées, n'est *pas* un domaine où les expériences sont reproductibles. Nous nous retrouvons ici dans la situation de la chimie du moyen-âge, où le manque flagrant de produits chimiques stables, par manque de connaissances sur les éléments chimiques, aboutissait à une très faible reproductibilité : une grande partie du temps des alchimistes était consacrée à essayer de reproduire des expériences décrites dans des grimoires, qui utilisaient des produits mal définis, et donc aboutissaient à des résultats peu stables. La chimie moderne a émergé de la prise de conscience que seuls des éléments chimiques stables et connus pouvaient permettre la reproductibilité des résultats, après la publication en 1664 de « Sceptical Chemist » de Boyle pour avoir une définition satisfaisante de l'élément chimique, en tant que corps indécomposable, mettant fin au principe alchimique mystique de « sympathie universelle » qui a bloqué la compréhension des phénomènes chimiques. Comme pour la chimie des premiers âges, nous devons rechercher ce qui permettra la confiance dans les résultats publiés par rapport aux propriétés annoncées des corpus utilisés ; actuellement la reproductibilité est à peu près assurée pour un même système de traitement automatique, si l'on reprend le même corpus exactement [10], mais pas un autre qui aurait les mêmes caractéristiques annoncées, car celles-ci sont souvent approximatives. Par exemple, pour le traitement de la parole spontanée, plusieurs corpus

10. Ce qui est déjà un progrès immense par rapport à la situation des années 60 à 80 où chaque papier faisait état de résultat sur des données qui n'étaient pas utilisées ailleurs, voire pas publiques.

s'en réclament comme par exemple le corpus CSJ, « Corpus of Spontaneous Japanese » [Maekawa00], qui contient 658 heures de parole, en provenance à 95% de présentations académiques et de prises de parole en public (en particulier intervention des journalistes de plateau dans les émissions télévisuelles) ; comment comparer le contenu de cette parole, peu spontanée selon nos critères occidentaux, puisque justement les présentations publiques sont préparées (en particulier au Japon), et d'autres corpus comme celui de Nijmegen [Torreira10], qui contient des conversations entre amis, enregistrées à leur insu ? Ici, comme pour l'alchimie, nous devons pouvoir décrire la matière qui nous permet de produire nos résultats avec une précision plus grande. Pour revenir à ma définition de l'observable, il est nécessaire de pouvoir rattacher avec précision un corpus avec son ELI, afin de pouvoir en déduire des propriétés généralisables ; le CSJ et le corpus de Nijmegen ne sont pas des instances du même ELI, car ils sont représentatifs de faits de langage ayant des paramètres intrinsèques et explicites bien différents, alors qu'ils sont tous les deux dénommés comme « corpus de parole spontanée ».

Le système actuel de présentations des résultats scientifiques, avec un nombre important de workshops, mono-thématiques pour la plupart, et quelques très rares congrès très généralistes, ne favorise pas le développement d'une science expérimentale, qui demande en particulier que certaines thématiques, portant sur les instruments (comme en astronomie ou en physique des particules) ou sur la mise au point de techniques permettant d'obtenir des éléments de qualité qui seront la base d'expériences fructueuses et reproductibles (comme en chimie ou en biologie), soient fortement valorisées. L'apparition de la conférence LREC (*International Conference on Language Resources and Evaluation*) a été un premier pas fructueux pour la mise en avant de papiers se reportant à la mise au point d'évaluation ou de corpus. Autre symptôme de cette inadéquation du mode de production actuelle, on peut observer pour un nombre croissant de papiers, la multiplication des thèmes abordés, influencés en cela par le mode d'évaluation des articles en parole, qui tend actuellement *pour un même papier*, à exiger une approche novatrice et/ou une méthode originale, une implémentation réussie, et une expérimentation significative. Bien sûr, certains papiers répondent à tous ces critères, mais nous voyons, poussés par la pression bibliométrique qui obligent les chercheurs à produire un nombre toujours croissant d'articles, un grand nombre de papiers qui contiennent une partie « méthode » dans l'aspect novateur est souvent gonflé par une dénomination avantageuse et une mathématisation souvent confuse étant donné le format réduit dans les conférences,

suivie d'une implémentation dans un système n'étant pas de l'état de l'art (et donc, étant donnée la complexité des systèmes actuels, pour laquelle il ne sera pas possible d'inférer des résultats pour un système de l'état de l'art), sur une base de données réduite en taille et/ou en complexité, qui aura l'avantage de produire facilement des résultats (par exemple TIMIT ou TIdigit en parole, ou le Wall Street Journal dans le domaine de l'écrit), mais où également il sera difficile de connaître son exacte portée sur une base de données plus réaliste. En particulier, un certain nombre d'articles, sous la dénomination « modélisation » ne sont que la formalisation mathématique d'hypothèses obtenues a posteriori, issues d'observations sur une base de données unique ; pour se situer dans un cadre réellement expérimental, ces hypothèses devraient être validées sur d'autres données avant d'en tirer un modèle, mais surtout il est nécessaire que ce modèle ait une qualité de généralisation et de *prévision* d'un certain nombre de résultats, ce qui en retour permet de mettre au point des expériences pour valider ce modèle. De ces papiers, il est difficile de tirer une connaissance directe, sauf à ré-implémenter la méthode, et la tester soi-même dans un système de l'état de l'art et sur des données suffisantes en taille et en difficulté.

Un moyen de permettre de plus apprendre d'expériences ou de modèles peut passer vers une façon différente de présenter les résultats, par une structuration différente de la production scientifique. Cette structuration doit venir de la réponse à l'interrogation de J.R. Pierce [Pierce69] (voir section 5) : quel est notre but ? Or on peut dégager (au moins) trois sous-buts [11] qui, en ce qui concerne le traitement de la parole, permettraient de réduire et de focaliser la recherche.

- Comment l'homme produit et comprend la parole ? Ici se placent les démarches issues de l'intelligence artificielle, le rapprochement entre la reconnaissance par les humains (HSR) et par les machines (ASR), l'intégration des résultats issus de la neuro-imagerie, l'intégration de la sémantique, « aller vers le sens » [Levinson95].
- Comment caractériser les faits de langue ? ici se situe l'exploration des *espaces langagiers isoparamétriques* (ELI), ou plus simplement des corpus, et en particulier l'exploration des contours par l'étude des erreurs, des paramètres pertinents ou redondants, mais également l'utilisation des systèmes automatiques pour explorer les corpus afin d'y valider des hypothèses linguistiques.

11. Ces trois questions abordent bien sûr des domaines qui se recouvrent.

– Comment produire des applications et des instruments efficaces ?
  - méthodes : une méthode ou algorithme nouveau doit être implémenté dans un système proche de l'état de l'art, et testé sur un certain nombre de corpus issus de différents ELI, pour lesquels cet(te) méthode/algorithme peut être pertinent(e).
  - systèmes : perfectionner les systèmes comme instrument pour explorer les corpus, et exploration d'autres ELI.
  - applications : développer des applications utilisant les systèmes mis au point, et par là même avoir accès à de nouveaux corpus.

Ces trois sous-buts nécessitent, pour ce qui concerne l'organisation de la recherche, des approches, des expériences différentes. La séparation de la production selon ces trois sous-buts permettraient de mieux structurer la production, et donc la lisibilité et la réutilisabilité des résultats.

La structuration obtenue est orthogonale à la structuration actuelle en « thèmes », et est une prolongation de la tendance que l'on voit en pratique dans certains workshops et certaines sessions spéciales. Elle peut donc parfaitement être initiée dans le paysage actuel des congrès et workshops, à condition que la règle soit clairement explicitée, en particulier pour ce qui concerne l'évaluation des articles.

7.5 Des centres instrumentaux pour le traitement des langues

Nous l'avons vu dans la section 6.1.5, certaines équipes/laboratoires obtiennent par leur participation aux évaluations majeures une *certification* qui leur accorde une crédibilité. Cette crédibilité leur apportera tout à la fois une plus grande facilité à faire publier leurs résultats et une plus grande confiance de la communauté dans la validité de ces résultats. Or, comme l'ont mis en avant certains auteurs (voir section 6.2.1, en particulier [Bourlard96]) un défaut de l'évaluation comparative, est une propension naturelle à *uniformiser* les systèmes : si une technique tend à prouver son efficacité, elle sera adoptée par les autres participants à l'évaluation, et à terme, les systèmes seront quasi-identiques, modulo les innovations de l'évaluation en cours. En caricaturant la situation actuelle, nous avons le choix entre une uniformité fiable ou une innovation peu fiable, puisque si une innovation est introduite par un équipe qui n'a pas été certifiée ou par un système qui n'a pas été calibré par

une évaluation, il n'y aura pas de confiance dans le résultat, au moins tant que celui-ci n'aura pas été reproduit par un couple système/équipe certifié.

De facto, il y a une énorme perte de temps et de travail dans ce modus vivendi, car il n'accorde pas assez de place pour les équipes qui n'ont pas les moyens humains ou matériels de participer à des évaluations importantes, donc de développer de front une recherche originale et le développement de systèmes performants. Cet état de fait peut causer (outre un risque de frustration pour ces équipes), la multiplication d'évaluations à faible coût d'investissement, mais à faible intérêt, afin de permettre à tous de participer au processus d'évaluation comparative ; ceci est également un facteur de perte de temps et de moyens.

Si les systèmes de traitement du langage peuvent être considérés comme des instruments, il doivent être considérés comme des instruments complexes, difficile à maîtriser et coûteux à développer et à maintenir, et donc pour lesquels il est intéressant de développer une structure de mutualisation, comme le sont les anneaux d'accélération utilisés en physique des particules, au CERN [12]. Le coût de mise au point des instruments est bien sûr de quelques ordres de grandeur plus faibles pour les systèmes de traitement du langage, mais il faut souligner l'intérêt de mutualiser, comme dans le cas de la physique des particules, le coût de développement et de maintenance des instruments. Cependant, il y a un intérêt certain à avoir plusieurs instruments différents, car nous avons vu que les méthodes pouvaient avoir un comportement différent suivant les systèmes ; par ailleurs, l'utilisation de plusieurs systèmes peut amener des informations complémentaires (voir section 7.3). On peut donc penser à la mise en place de cette mutualisation au sein de plusieurs laboratoires qui seraient alors des *centres instrumentaux* mutualisés, dont le nombre dépend de la capacité de la zone géographique à en faciliter le développement. Un laboratoire peut par ailleurs devenir centre instrumental en fonction de ses résultats récurrents à diverses évaluations. Le statut de centre instrumental amènerait des moyens (financiers et humains) supplémentaires ; les centres instrumentaux seraient impliqués dans les projets expérimentaux d'autres laboratoires qui n'auraient pas le statut de centre instrumental ; en échange, le centre accueillerait des équipes (choisis sur des projets nécessitant l'emploi d'un système état de l'art et/ou de corpus) et des chercheurs (choisis également sur des projets mais aussi uniquement d'après leur compétence),

12. Organisation européenne pour la recherche nucléaire, `public.web.cern.ch/public/`

et recevraient l'aide active du personnel qualifié du centre instrumental. Un tel schéma ne peut efficacement fonctionner sans un système de bourses permettant aux équipes et aux chercheurs d'avoir les moyens de mener leurs expériences au sein du centre instrumental, avec une rétribution à celui-ci. Au terme de leur passage, on peut imaginer différents modes de fonctionnement ; par exemple, les équipes ou les individus retourneraient dans leur laboratoire d'origine, en partageant leurs résultats (et leurs publications) avec le centre instrumental d'accueil, et continueraient à avoir un lien avec l'équipe d'accueil et un accès à l'instrument pendant un certain temps, afin d'assurer le suivi des expériences. L'ensemble des laboratoires utilisateurs et le centre instrumental qui partageraient un même instrument formeraient un réseau, ce qui permettrait pour un laboratoire utilisateur, d'obtenir des partenariats industriels, en collaboration avec le centre instrumental, que l'instrument soit sous licence libre ou commerciale.

En ce qui concerne l'instrument ou les instruments à utiliser, il y a deux solutions, les solutions sous licences commerciales (comme les logiciels mis au point par le LIMSI ou BBN) et les boîtes à outils mises à disposition gratuitement ; parmi ces dernières, on peut citer Sphinx [13], le toolkit distribué par le RWTH [14], ou encore le célèbre HTK [15]. Le principe de base de la mise à disposition gratuite est que le logiciel est mis à jour par l'ensemble des utilisateurs, avec des modèles économiques particuliers pour les applications industrielles utilisant le logiciel, les laboratoires contribuant au logiciel ayant également un modèle économique par le savoir-faire dans l'utilisation de ce logiciel. Les avantages sont bien connus et le modèle du logiciel libre est fonctionnel dans un grand nombre de domaines de l'informatique : une mise à jour rapide, une mise à disposition de la communauté des innovations au sein d'une plate-forme. L'avantage de la solution commerciale est souvent une facilité d'intégration dans des applications industrielles (achat de licence) Les boîtes à outil libres sont utilisées depuis de nombreuses années, mais il faut constater pragmatiquement que peu de laboratoires arrivent à obtenir des résultats compétitifs à l'aide de ces boîtes à outils, à part bien sûr les laboratoires qui ont mis au point ces logiciels, et ceux qui ont fortement investi dans la maîtrise d'une boîte à outil. Ce problème de choix entre logiciel libre et la licence commerciale est secondaire : il faut constater

13. `cmusphinx.sourceforge.net/`
14. `www-i6.informatik.rwth-aachen.de/rwth-asr`
15. `htk.eng.cam.ac.uk/`

au vu des résultats obtenus par les différentes équipes pendant les évaluations, que le logiciel n'est qu'une composante du résultat final, et que la taille, la compétence et l'expérience de l'équipe utilisant ce logiciel sont des paramètres au moins aussi important que le logiciel lui-même. L'utilisation d'une suite logicielle (quel que soit sa nature) qui serait distribué à tous les laboratoires, à charge pour eux de maîtriser le logiciel, et de développer les modèles ne semble pas de nature à permettre d'augmenter significativement la qualité expérimentale moyenne des expériences, car elle ne résout pas les problèmes cruciaux de la gestion des bases de données multilingues, de la maîtrise des outils de traitement (normalisation des textes, fabrication des modèles de langage, paramétrisations, modélisation acoustique), la maîtrise du logiciel et surtout de l'expérience et de la compétence de l'équipe en ce qui concerne le menée d'expériences de grande ampleur. Le développement de centres instrumentaux pour le traitement de la parole pourrait être un modèle de recherche original et européen. Par rapport au modèle américain qui peut se résumer à développer quelques centres d'excellence privés ou public, financés par les projets gouvernementaux, autour de centres d'évaluation et de mises à disposition de corpus (NIST [16] et LDC [17]), eux-mêmes financés directement ou indirectement par des projets gouvernementaux, on ne peut que constater que, en Europe, malgré l'existence de quelques excellents laboratoires et d'une agence de distribution de données, ELDA [18], il n'existe pas de modèle efficace européen.

Dans le schéma de recherche que je propose, l'ensemble des expériences pourraient être mené avec des systèmes de l'état de l'art et sur des bases de données pertinentes. Les moyens de mise au point des instruments seraient concentrés sur quelques centres, et les autres moyens étant dévolus à des expériences précises. Le savoir acquis autour de l'instrument serait partagé entre les différentes équipes du réseau, et pérennisé au sein de l'équipe gérant le centre instrumental. Ce schéma conserve la fiabilité directement issue de l'évaluation comparative, mais permet une plus grande innovation pour un même coût.

16. `www.itl.nist.gov/`
17. `www.ldc.upenn.edu/`
18. `http://www.elda.org/`

Chapitre 8

Conclusion

J'ai abordé dans cette partie les avantages de l'évaluation comparative, en reprenant des analyses et des critiques d'auteurs qui ont essayé de décrire et de formaliser la manière dont le domaine du traitement de la parole a évolué au cours des quarante dernières années. Durant cette période, le paysage scientifique de notre domaine a fortement évolué, en mettant en place à la fois une méthode d'objectivisation des résultats par l'utilisation de l'évaluation comparative sur des corpus communs, et les moyens de cette méthode, c'est-à-dire la distribution de corpus de grande taille, et de complexité croissante, la possibilité d'une évaluation objective par un tiers, et des moyens mis à disposition des équipes de recherche par les acteurs gouvernementaux et industriels, mis en confiance par une vision à court et moyen terme des potentialités du domaine.

Cependant, la mise en place de l'évaluation comparative et ses bénéfices visibles actuels n'est qu'une étape dans la mise en place d'une structure scientifique à même de permettre un développement à long terme de notre domaine, qui passe en particulier selon moi par la pérennisation d'une vraie science empirique. Cette pérennisation passe par :

– Une plus grande formalisation des buts, et des moyens à mettre en œuvre pour atteindre ces buts. Cela signifie définir clairement l'observable de cette science empirique, le moyen de l'atteindre, de l'observer, puis comment apprendre à partir de celui-ci. Je suggère d'utiliser comme observable une extension de la notion de corpus, l'espace langagier isoparamétrique, qui est défini par les paramètres intrinsèques et extrinsèques qui sont choisis pour extraire des corpus représentatifs de cet espace. On peut étudier cet espace en observant le comportement

des systèmes sur des corpus issus de cet espace, en interrogeant ces paramètres (sont-ils pertinents, y a-t-il assez de paramètres différents ?) et en examinant les erreurs des systèmes.
- Une structuration de la production scientifique. Elle passe en particulier vers la remise au goût du jour de cette question simple : quel est notre but ? s'agit-il de modéliser le fonctionnement de l'humain dans sa globalité, d'étudier les faits de langue en tant qu'objets d'analyse, ou de construire des systèmes efficaces ? tous ces buts sont pertinents, mais les moyens à mettre en œuvre sont différents.
- La prise en compte des critiques et des frustrations. Par des critiques récurrentes de certains acteurs majeurs, mais également par d'autres indices, comme par exemple la connaissance objective que l'on est capable d'extraire des articles publiés lors d'un congrès comme Interspeech ou ICASSP, on peut en déduire que le domaine du traitement de la parole a des faiblesses structurelles, et qu'il n'utilise pas aussi efficacement qu'il le pourrait le potentiel humain qui s'y implique. Une amélioration peut passer par une meilleure structuration de la production scientifique (voir ci-dessus) mais surtout par une mutualisation des systèmes de traitement automatique comme instruments, systèmes complexes et coûteux à mettre en œuvre et à maintenir au meilleur niveau.

Si j'ai soulevé certaines critiques, par exemple sur le mode de production des articles et de leur contenu, il s'agit de critiques sur les structures qui les génèrent, et jamais sur les chercheurs. Si nous sommes individuellement responsables du travail que nous faisons, il n'est pas raisonnable de penser qu'il en est de même sur le choix des thématiques, les sujets des projets que nous soumettons aux appels d'offres, ou la façon dont nous pouvons publier : sur tous ces sujets nous sommes orientés, bridés par les structures académiques ou étatiques de recherche, dont le but est d'organiser et d'orienter la recherche. Seules ces structures qui organisent, coordonnent, financent la recherche ont la possibilité de faire réellement changer les choses, mais je suis également persuadé que ces structures ne sont pas imperméables à des changements, et c'est parce que j'ai cette conviction que j'ai écrit ce document.

A l'heure actuelle, parce que les applications technologiques actuelles ou à venir, qu'elles soient à visées sociétales (aide au personnes âgées, aide au handicap, apprentissage des langues...), gouvernementales (renseignement, identification, ...), ou industrielles (recherche d'information, sous-titrage, serveurs vocaux, ...) sont nombreuses, le domaine du traitement de la parole bénéficie, avec de fortes disparités entre les sous-domaines et entre les

laboratoires, de moyens relativement importants. Cependant, la connaissance que nous générons pourrait être plus importante, à la fois en formalisant plus notre discipline de manière à faire émerger une science empirique stable, et en utilisant mieux le potentiel actuel en terme de moyens humains et matériel. Nous devons être capable de structurer notre discipline, de manière à générer une connaissance régulière, et d'être capable d'absorber les risques inhérents aux grands cycles technologiques et économiques dont nous sommes tous tributaires.

Conclusion Générale

Dans ce document, que je n'ai pas voulu trop technique[1], j'ai pu décrire dans la première partie un parcours de recherche en parole, finalement pas si sinueux. J'ai eu l'avantage de travailler au sein d'un laboratoire et d'une équipe, où j'ai pu participer à la mise au point de systèmes de traitement de la parole parmi les meilleurs mondiaux, ainsi qu'à la mise en place du paradigme de l'évaluation comparative. Grâce à cette situation privilégiée, j'ai pu collaborer à de nombreuses études, où j'ai principalement apporté ma pierre en travaillant sur la modélisation linguistique. J'ai calculé que parmi les 145 articles référencés dans ma bibliographie, j'ai eu 110 co-signataires différents (permanents, doctorants ou post-doctorants) ce qui laisserait supposer un certaine dispersion ; il n'en est rien, et ces travaux sont en premier lieu le résultat du travail d'un équipe restreinte avec qui j'ai étroitement collaboré depuis plus de 25 ans : parmi ces 145 papiers, plus de 80 étaient cosignés avec Lori Lamel ou Jean-Luc Gauvain, plus de 50 avec Martine Adda-Decker. Cette collaboration fructueuse m'a permis d'apprendre énormément au niveau scientifique, mais m'a également permis d'être parmi les spectateurs privilégiés de l'évolution de la communauté internationale en traitement de la parole depuis 25 ans.

C'est donc en tant qu'acteur de base et spectateur, que j'ai réfléchi ces dernières années, à partir de lecture et de rencontres[2], au statut du traitement de la parole, en tant que science à part entière. Dans la deuxième partie de ce mémoire d'habilitation, j'ai essayé de restituer brièvement certains faits et certaines lectures qui ont nourri ma réflexion, en particulier sur le statut de l'évaluation comparative. Ce concept a fait avancer à grands pas la recherche en parole et a structuré notre domaine depuis plus de 20 ans, mais a toujours un statut assez spécial : incontournable, voire omniprésent, mais mal défini. Si quelques personnalités marquantes se sont penchées sur son berceau, et ont largement communiqué l'intérêt, la manière de faire, les points-clés, etc, peu de travaux ont cherché à trouver un support plus théorique à ce paradigme.

1. Il est possible que certains lecteurs aient été rebutés par la lecture de ce document peu technique. J'en suis désolé pour eux, mais qu'ils sachent, si cela peut les consoler, que le fait qu'il soit peu technique n'a pour moi ajouté aucune facilité à l'écriture ; cela m'a été dicté par la thématique de ce document, c'est-à-dire le développement d'une science empirique du traitement de la parole, qui vise un public plus large que les seuls professionnels de la modélisation du langage.

2. Ma rencontre avec Benoît Habert a été à cet égard déterminante, par sa connaissance de l'épistémolologie, ainsi que par ses écrits sur l'utilisation d'instruments en science du langage.

Or, une formalisation est souvent une étape indispensable pour définir les contours d'une nouvelle science et chercher de nouvelles voies de recherche. C'est dans ce but que j'ai proposé quelques pistes, qui nécessiteraient d'être approfondis pour être validées. C'est ce à quoi je désire consacrer une grande part de ma recherche à venir, que ce soit sur l'étude des erreurs, ou la définition de l'observable. Enfin, et toujours pour favoriser le développement d'une science empirique de la parole, j'ai fait quelques propositions concrètes, en ce qui concerne la structuration de la production scientifique et le développement de centres instrumentaux : loin d'être des vœux pieux, elles me paraissent réalisables à court terme, si en tant que communauté de recherche, nous œuvrons pour les rendre réelles.

Bibliographie

Bibliographie personnelle

Bibliographie

[Adda11A] G. Adda, G. Chollet, S. Essid, T. Fillon, M. Garnier-Rizet, C. Hory, L. Beltaifa Zouari, Traitement des modalités “audio” et “parole” In *Sémantique et multimodalité en analyse de l'information*, sous la direction de Marine Campedel et Pierre Hoogstoël, pp. 143-188, collection "Recherche d'Information et Web", B. Bouchon-Meunier Ed, Hermes/Lavoisier, ISBN 978-2-7462-3139-9, Avril 2011.

[Adda11B] G. Adda, F. Cailliau, A.-L. Daquo, M. Garnier-Rizet, S. Guillemin-Lanne, P. Suignard, C. Waast-Richard La transcription automatique et la fouille de données conversationnelles pour l ?analyse de la relation client In *Sémantique et multimodalité en analyse de l'information*, sous la direction de Marine Campedel et Pierre Hoogstoël, pp. 143-188, collection "Recherche d'Information et Web", B. Bouchon-Meunier Ed, Hermes/Lavoisier, ISBN 978-2-7462-3139-9, Avril 2011.

[AddaDecker11] M. Adda-Decker, G. Adda, L. Lamel, Les systèmes de transcription automatique de la parole comme instruments de mesure de grands corpus oraux In *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*, Nguyen, N. (eds), Hermes/Lavoisier *à paraître 2011.* .

[Fort11] K. Fort, G. Adda, and K. Bretonnel Cohen, Amazon Mechanical Turk : Gold Mine or Coal Mine ? Computational Linguistics 37(2), 413-420, 2011

[Sagot11] B. Sagot, K. Fort, G. Adda, J. Mariani, B. Lang, Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. TALN 2011, Montpellier, pp 199-210, 27 juin - 1er juillet 2011.

2010

[Adda10A] G. Adda, Language resources and Amazon Mechanical Turk : legal, ethical and other issues LISLR2010, "Legal Issues for Sharing Language Resources workshop", LREC2010, Malta, 17 mai 2010

[Adda10B] G. Adda, Using the Amazon Mechanical Turk for the production of Language Resources FLaReNet Forum 2010, "Language Resources of the future - the future of Language Resources", Barcelona, 11-12 février 2010

[Quintard10] L. Quintard, O. Galibert, D. Laurent, S. Rosset, G. Adda, V. Moriceau, X. Tannier, B. Grau, A. Vilnat, Question answering on web data : the QA evaluation in Quaero LREC 2010. Seventh International Conference on Language Resources and Evaluation, Valetta, Malta , 2010.

[Snoeren10] N.D. Snoeren, M. Adda-Decker, G. Adda The Study of Writing Variants in an Under-resourced Language : Some Evidence from Mobile N-Deletion in Luxembourgish LREC 2010. Seventh International Conference on Language Resources and Evaluation, Valetta, Malta : 2010.

2009

[Bernard09] G. Bernard, S. Rosset, O. Galibert, E. Bilinski, G. Adda, The LIMSI participation to the QAst 2009 track Working Notes of CLEF Workshop. Corfu, Greece. September 2009.

[Rosset09] S. Rosset, O. Galibert, G. Bernard, E. Bilinski, G. Adda, The LIMSI multilingual, multitask QAst system Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers. C. Peters et al. (Eds)

2008

[AddaDecker08A] M. Adda-Decker, T. Pellegrini, E. Bilinski, and G. Adda. Developments of lëtzebuergesch resources for automatic speech processing and linguistic studies. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.

[AddaDecker08B] M. Adda-Decker, C. Barras, G. Adda, P. Paroubek, P. Boula de Mareuil, and B. Habert. Annotation and analysis of overlapping speech in political interviews. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.

[GarnierRizet08] M. Garnier-Rizet, G. Adda, F. Cailliau, J.-L. Gauvain, S. Guillemin-Lanne, L. Lamel, S. Vanni, and C. Waast-Richard. Callsurf : Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.

[Rosset08] S. Rosset, O. Galibert, G. Bernard, E. Bilinski, G. Adda, The LIMSI participation to the QAst track Working Notes of CLEF 2008 Workshop. Cross-Language Evaluation Forum. In conjunction with ECDL 2008

[Dechelotte08] D. Déchelotte, G. Adda, A. Allauzen, H. Bonneau-Maynard, O. Galibert, J.-L. Gauvain, P. Langlais, and F. Yvon. LIMSI's Statistical Translation Systems for WMT'08. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 107-110, Columbus, Ohio, June 2008.

2007

[Dechelotte07A] D. Déchelotte, H. Schwenk, G. Adda, and J.-L. Gauvain. Improved Machine Translation of Speech-to-Text outputs. In InterSpeech'07, Antwerp, Belgium, August 2007.

[Dechelotte07B] D. Déchelotte, H. Schwenk, H. Bonneau-Maynard, A. Allauzen, and G. Adda. A state-of-the-art statistical machine translation system based on Moses. In MT Summit, pages 127-133, Copenhagen, September 10-14 2007.

[Adda07] G. Adda, M. Adda-Decker, C. Barras, P. Boula de Mareüil, B. Habert, and P. Paroubek. Speech Overlap and Interplay with Disfluencies in Political Interviews. In International workshop on Paralinguistic Speech - between models and data, ParaLing 2007, pages 41-46, Saarbruecken, August 2007.

[Lamel07B] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu. The LIMSI 2006 TC-STAR EPPS Transcription Systems. In Proceedings of IEEE-ICASSP, pages 997-1000, Honolulu, Hawaii, April 2007.

[Rosset07A] S. Rosset, O. Galibert, G. Adda, and E. Bilinski. The LIMSI participation to the QAst track. In Working Notes for the CLEF 2007 Workshop, Budapest and Hungary, September 2007.

[Rosset07B] S. Rosset, O. Galibert, G. Adda, and E. Bilinski. The LIMSI Qast systems : comparison between human and automatic rules generation for question-answering on speech transcriptions. In IEEE ASRU, December 2007.

[Lamel07A] L. Lamel, E. Bilinski, J.-L. Gauvain, G. Adda, C. Barras, and X. Zhu. The LIMSI RT07 Lecture Transcription System. In S. Renals, S. Bengio, and J. Fiscus, editors, Lecture Notes in Computer Science, Bethesda, MD, May 2007. Springer Verlag.

2006

[AddaDecker06] M. Adda-Decker, P. Boula De Mareüil, G. Adda, N. Nguyen. Analyses phonétiques et phonologiques du corpus PFC par alignement automatique dans le projet VARCOM Colloque international PFC 2006. Approches phonologiques et prosodiques de la variation sociolinguistique : le cas du français, Louvain-la-Neuve, Belgique : 2006

[Baude06] Corpus oraux : guide des bonnes pratiques 2006. Collaboration au groupe de réflexion coordonné par Olivier Baude - CNRS Editions, 2006

[Lamel06A] L. Lamel, E. Bilinski, G. Adda, J.-L. Gauvain, and H. Schwenk. The LIMSI RT06s Lecture Transcription System. In S. Renals, S. Bengio, and J. Fiscus, editors, Lecture Notes in Computer Science, Vol. 4299 - Proc. 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006), Washington, May 2006. Springer Verlag.

[Lamel06B] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu. The LIMSI 2006 TC-STAR Transcription Systems . In TC-STAR Workshop on Speech-to-Speech Translation, pages 123-128, Barcelona, June 2006.

[Matsoukas06] S. Matsoukas, J.-L. Gauvain, G. Adda, T. Colthurst, C.-L. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang. Advances in Transcription of Broadcast News and Conversational Telephone Speech within the Combined EARS BBN/LIMSI System. IEEE Transactions on Audio, Speech and Language Processing, 14(5) :1541-1556, 2006.

2005

[AddaDecker05] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*,46 :119-139, 2005.

[BoulaDeMareuil05] P. Boula de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Baras, G. Adda and P. Paroubek A quantitative study of disfluencies in French broadcast interviews In *Proceedings of Disfluency In Spontaneous Speech (DISS) Workshop* , Aix-en-Provence, september 2005.

[Gauvain05A] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk Where Are We in Transcribing French Broadcast News ? In *Proceedings of Interspeech*, Lisbon, september 2005.

[Gauvain05B] Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Veronique Gendner, Lori Lamel, Holger Schwenk Le système TRS du LIMSI Atelier Ester 30-31 mars 2005. Avignon.

[Lamel05A] L. Lamel, H. Schwenk, J.-L. Gauvain, G. Adda, and E. Bilinski. Improvements in Transcribing Lectures and Seminars. Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), Edinburgh, 2005.

[Lamel05B] L. Lamel, G. Adda, E. Bilinski, and J.-L. Gauvain Transcribing Lectures and Seminars In *Proceedings of Interspeech*, Lisbon, september 2005.

[Prasad05] R. Prasad, S. Matsoukas, C.-L. Kao, J.Z. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System In *Proceedings of Interspeech*, Lisbon, september 2005.

2004

[AddaDecker04] Martine Adda-Decker, Benoit Habert, Claude Barras, Gilles Adda, Philippe Boula de Mareüil et Patrick Paroubek Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage In *Actes des JEP*, Fez, avril 2004.

[Barras04] C. Barras, G. Adda, M. Adda-Decker, B. Habert, P. Boula de Mareüil and P. Paroubek Automatic Audio and Manual Transcripts Alignment, Time-code Transfer and Selection of Exact Transcripts In *Proceedings of LREC*, Lisbon, May 2004.

[Chen04] L. Chen, J.-L. Gauvain, and L. Lamel ad G. Adda. Dynamic Language Modeling for Broadcast News In *Proceedings of ICSLP*, Jeju Island, October 2004.

[Gauvain04A] Jean-Luc Gauvain, Gilles Adda, Lori Lamel, Fabrice Lefevre and Holger Schwenk Transcription de la parole conversationnelle In *Actes des JEP*, Fez, avril 2004.

[Gauvain04B] J.-L. Gauvain, G. Adda, L. Lamel, F. Lefevre, and H. Schwenk. Transcription de la parole conversationnelle. *Traitement Automatique des Langues*, 2004, Volume 45 numéro 3.

[Lamel04] L. Lamel, J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk Speech Transcription in Multiple Languages In *Proceedings of IEEE-ICASSP*, Montreal, May 2004.

[Nguyen04] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.L. Gauvain, G. Adda, H. Schwenk and F. Lefevre The 2004 BBN/LIMSI 10xRT english broadcast news transcription system In *Proceedings of DARPA RT04 Workshop*, Palisades NY, November 2004.

[Prasad04] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, G. Thattai, O. Kimball, R. Schwartz, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda and F. Lefevre The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech System In *Proceedings of DARPA RT04 Workshop*, Palisades NY, November 2004.

[Schwartz04] R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C.-L. Kao, D. Liu, O. Kimball, J. Ma, J. Makhoul, S. Matsoukas, L. Nguyen, M. Noamany, R. Prasad, B. Xiang, D. Xu, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and L. Chen Speech recognition in multiple languages and domains : The 2003 BBN/LIMSI EARS system In *Proceedings of IEEE-ICASSP*, Montreal, May 2004.

2003

[AddaDecker03] M. Adda-Decker, B. Habert, C. Barras, G. Adda, Ph. Boula de Mareuil, and P. Paroubek A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models In *Proceedings of ISCA DiSS '03 - Disfluency in Spontaneous Speech Workshop*, Gothenburg, September 2003.

[Chen03] L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda Unsupervised language model adaptation for broadcast news In *Proceedings of IEEE-ICASSP*, Hong Kong, April 2003.

[Gauvain03A] J.L. Gauvain, L.F. Lamel, G. Adda, L. Chen, H. Schwenk, The LIMSI RT03 BN systems DARPA/NIST RT 03 Workshop, Boston : 2003

[Gauvain03B] J.L. Gauvain, L.F. Lamel, H. Schwenk, G. Adda, L. Chen, Speech-to-text research at LIMSI EARS Mid Year Workshop, Berkeley : 2003

[Gauvain03C] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen and F. Lefèvre Conversational telephone speech recognition In *Proceedings of IEEE-ICASSP*, Hong Kong, April 2003.

[Gauvain03D] J.L. Gauvain, L.F. Lamel, H. Schwenk, G. Adda, L. Chen, CTS progress at LIMSI DARPA/NIST RT 03 Workshop, Boston : 2003

[Gauvain03E] J.L. Gauvain, L.F. Lamel, H. Schwenk, G. Adda, Experiments with Fisher data DARPA EARS STT Dec'03 Workshop, Saint Thomas : 2003

[Lamel03] L.F. Lamel, J.L. Gauvain, L. Chen, G. Adda, Training acoustic models with TDT DARPA EARS STT Sep'03 Workshop, Martigny, Switzerland : 2003

2002

[AddaDecker02] M. Adda-Decker, P. Boula de Mareüil, G. Adda, and L. Lamel. Investigating syllabic structure and its variation in speech from French radio interviews In *Proceedings of ISCA ITRW Pronunciation modeling and Lexicon Adaptation for Spoken Language*, Estes Park, September 2002.

[Gauvain02A] J.L. Gauvain and L. Lamel and G. Adda The LIMSI Broadcast News Transcription System *Speech Communication*, 37(1-2) :89-108, 2002.

[Gauvain02B] J.L. Gauvain, L.F. Lamel, H. Schwenk, G. Adda, F. Lefevre, The LIMSI April 2002 SWB and BN systems DARPA "Rich Transcription" Workshop. Vienna, Austria, May 7-8, 2002, 2002

[Lamel02A] L. Lamel, J.L. Gauvain, and G. Adda Lightly supervised and unsupervised acoustic model training *Computer Speech and Language*, 16(1) :115-229, 2002.

[Lamel02B] L. Lamel, J.-L. Gauvain, and G. Adda Unsupervised Acoustic Model Training In *Proceedings of IEEE-ICASSP*, Orlando, May 2002.

2001

[Chen01A] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker Language Model Adaptation for Broadcast News Transcription In *Proceedings of ISCA ITRW 2001 Adaptation Methods for Speech Recognition*, Sophia-Antipolis, August 2001.

[Chen01B] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker Using Information Retrieval Methods for Language Model Adaptation In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, Aalborg, September 2001.

[Gauvain01A] J.L. Gauvain, L. Lamel, and G. Adda. Audio partitioning and transcription for broadcast data indexation. *Multimedia Tools and Applications - MTAP Journal*, 14(2) :187–200, 2001.

[Gauvain01B] J.-L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, C. Barras, L. Chen, Y. de Kercadio Processing Broadcast Audio for Information Access In *Proc. of the Joint EACL - ACL Meeting*, Toulouse, juillet 2001.

[Gauvain01C] J.-L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, C. Barras, L. Chen, and Y. de Kercadio Processing Broadcast Audio for Information Access In *Proceedings of the ACL 39th annual meeting*, Toulouse, July 2001.

[Lamel01A] L. Lamel, J.-L. Gauvain, and G. Adda. Investigating lightly supervised acoustic model training. In *Proceedings of IEEE-ICASSP*, Salt Lake City, May 2001.

[Lamel01B] L. Lamel, F. Lefèvre, J.-L. Gauvain, and G. Adda. Portability issues for speech recognition technologies. In *Proceedings of HLT 2001*, pages 9–16, San Diego, March 2001.

2000

[Adda00A] G. Adda, M. Adda-Decker, J.L. Gauvain, L.F. Lamel, Le système de dictée du LIMSI pour l'évaluation AUPELF'97 In : K. Chibout, J. Mariani, N. Masson, F. Néel (Eds.) "Ressources et Evaluation en Ingénierie des Langues". Collection Champs Linguistiques. De Boeck, Paris, 2000

[Adda00B] G. Adda, M. De Calmès, L.F. Lamel, G. Perennou, M. Rajman, S. Rosset, J. Zeiliger, Ressources pour l'apprentissage, le développement

et l'évaluation des systèmes de dictée vocale en français : corpus de texte, de parole et lexical In : K. Chibout, J. Mariani, N. Masson, F. Néel (Eds.) "Ressources et Evaluation en Ingénierie des Langues". Collection Champs Linguistiques. De Boeck, Paris, 2000

[Adda00C] G. Adda, M. Adda-Decker, Normalisation de textes en français : une étude quantitative pour la reconnaissance de la parole. In : K. Chibout, J. Mariani, N. Masson, F. Néel (Eds.) "Ressources et Evaluation en Ingénierie des Langues". Collection Champs Linguistiques. De Boeck, Paris, 2000

[Adda00D] G. Adda, J. Lecomte, J.J. Mariani, P. Paroubek, M. Rajman Les procédures de mesure automatique de l'action Grace pour l'évaluation des assignateurs de parties du discours pour le français In : K. Chibout, J. Mariani, N. Masson, F. Néel (Eds.) "Ressources et Evaluation en Ingénierie des Langues". Collection Champs Linguistiques. De Boeck, Paris, 2000

[AddaDecker00A] M. Adda-Decker, G. Adda, Morphological decomposition for ASR in German Workshop on Phonetics and Phonology in ASR. Saarbrücken, Germany, March 1-3, 2000.

[AddaDecker00B] M. Adda-Decker, G. Adda, and L. Lamel. Investigating text normalization and pronunciation variants for German broadcast transcription. In *Proc. ICSLP'2000*, pages 266–269, Beijing, Oct 2000.

[Barras00] C. Barras, G. Adda, M. Adda-Decker, L. Chen, J.L. Gauvain, Y. De Kercadio, L.F. Lamel, An audio transcriber for broadcast document indexation Démonstration au Congrès RIAO. Paris, 12-14 avril 2000, 2000

[Chen00] L. Chen, L. Lamel, G. Adda, and J.L. Gauvain. Broadcast news transcription in Mandarin. In *Proc. ICSLP'2000*, pages II–1015–1018, Beijing, Oct 2000.

[Gauvain00A] J.L. Gauvain, L. Lamel, and G. Adda. Transcribing broadcast news for audio and video indexing. *Communications of the ACM*, 43(2) :64–70, Feb 2000.

[Gauvain00B] J.L. Gauvain, L. Lamel, Y. de Kercadio, and G. Adda. Transcription and indexation of broadcast data. In *Proceedings of IEEE-ICASSP*, pages 1663–1666, Istanbul, Jun 2000.

[Gauvain00C] J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Chen C. Barras, M. Jardino, L. Lamel, and H. Schwenk. An overview of speech recognition

activities at LIMSI. In *Sino-French Symposium on Speech and Language Processing*, Beijing, Oct 2000.

[Gauvain00D] J.L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. Kercadio. The LIMSI SDR system for TREC-9. In *Proc. of the Text Retrieval Conference, TREC-9*, Gaithersburg, Nov 2000.

[Gauvain00E] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI 1999 hub-4e transcription system. In *Proc. DARPA Speech Transcription Workshop*, Gaithersburg, May 2000.

[Gauvain00F] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System DARPA Speech Transcription Workshop. College Park, USA, May 16-19, 2000.

[Lamel00] L. Lamel, J.L. Gauvain, and G. Adda. Lightly supervised acoustic model training. In *ISCA ITRW Workshop on Automatic Speech Recognition : Challenges for the new Millenium*, pages 150–154, Paris, Sep 2000.

1999

[Adda99A] G. Adda, M. Jardino, and J.-L. Gauvain. Language modeling for broadcast news transcription. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, pages 1759–1762, Budapest, Sep 1999.

[Adda99B] G. Adda, J. Lecomte, J. Mariani, P. Paroubek, and M. Rajman. L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 2(2) :119–129, 1999.

[Adda99C] G. Adda, J. Lecomte, J. Mariani, P. Paroubek, and M. Rajman. Métrique et premiers résultats de l'évaluation grace des étiqueteurs morpho-syntaxiques pour le francais. In *TALN'99*, Conférence sur le traitement automatique du langage Naturel, Cargèse, 12-17 juillet 1999.

[Gauvain99A] J.-L. Gauvain, L. Lamel, G. Adda, and M. Jardino. The LIMSI 1998 hub-4e transcription system. In *Proc. of the DARPA Broadcast News Workshop*, pages 99–104, Herndon, VA, Feb 1999.

[Gauvain99B] J.-L. Gauvain, L. Lamel, G. Adda, and M. Jardino. Recent advances in transcribing television and radio broadcasts. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, pages 655–658, Budapest, Sep 1999.

[Gauvain99C] J.-L. Gauvain, Y. Kercadio, L. Lamel, and G. Adda. The LIMSI SDR system for TREC-8. In *Proc. of the Text Retrieval Conference, TREC-8, notebook*, Gaithersburg, Nov 1999.

[Gauvain99D] J.-L. Gauvain, L. Lamel, and G. Adda. Audio partitioning and transcription for broadcast data indexation. In *Proc. CBMI'99*, Toulouse, Oct 1999.

[AddaDecker99] M. Adda-Decker, G. Adda, J.-L. Gauvain, and L. Lamel. Large vocabulary speech recognition in French. In *Proceedings of IEEE-ICASSP*, Phoenix, Mar 1999.

1998

[Adda98] G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajman. The GRACE French part-of-speech tagging evaluation task. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, volume I, pages 433–441, Granada, May 1998.

[AddaDecker98A] M. Adda-Decker, G. Adda, , J.L. Gauvain, and L. Lamel. On the use of speech & text corpora for automatic speech recognition in French. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, volume II, pages 783–788, Granada, May 1998.

[AddaDecker98B] M. Adda-Decker, G. Adda, J.L. Gauvain, and L. Lamel. Elements pour la mise au point de système de reconnaissance grand vocabulaire de français. In *Proc. XXIIièemes Journées d'Etudes sur la Parole*, pages 367–370, Martigny, June 1998.

[Gauvain98A] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI 1997 Hub-4E transcription system. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 75–79, Landsdowne,VA, February 1998.

[Gauvain98B] J.-L. Gauvain, L. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *International Conference on Speech and Language Processing*, volume 4, pages 1335–1338, Sydney, Australia, Dec 1998.

[Habert-et-al98] B. Habert, G. Adda, M. Adda-Decker, P. Boula de Marëuil, S. Ferrari, O. Ferret, G. Illouz, and P. Paroubek. Towards tokenization evaluation. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, volume I, pages 427–431, Granada, May 1998.

[Lamel98] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, and J.L. Gauvain. A multilingual corpus for language identi-

fication. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, volume II, pages 1115–1122, Granada, May 1998.

1997

[Adda97B] G. Adda and M. Adda-Decker. Normalisation de textes en français : une étude quantitative pour la reconnaissance de la parole. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 289–296, Avignon, France, April 1997.

[Adda97C] G. Adda, M. de Calmès, L. Lamel, G. Pérennou, M. Rajman, S. Rosset, and J. Zeiliger. Ressources pour l'apprentissage, le développement et l'évaluation des systèmes de dictée vocale en français : corpus de texte, de parole et lexical. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 305–309, Avignon, France, April 1997.

[Adda97D] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel. Text normalization and speech recognition in French. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, volume 5, pages 2711–2714, Rhodes, September 1997.

[AddaDecker97] J.M. Dolmazon, F. Bimbot, G. Adda, M. El Bèze, J. C. Caërou, J. Zeiliger, and M. Adda-Decker. Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 13–18, Avignon, France, April 1997.

[Gauvain97A] J.L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing broadcast news shows. In *Proceedings of the IEEE-ICASSP*, volume II, pages 715–719, Munich, April 1997.

[Gauvain97C] J.L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker. Transcription of broadcast news. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, volume 2, pages 907–910, Rhodes, September 1997.

[Gauvain97B] J.L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing broadcast news : The LIMSI Nov96 Hub4 system. In *Proceedings of ARPA Spoken Language Technology Workshop*, pages 56–63, Chantilly, Virginia, February 1997.

[Adda97E] G. Adda, M. Adda-Decker, J. L. Gauvain, and L. Lamel. Développement du système de dictée vocale du limsi. Atelier de l'ARC-B1 de l'Aupelf-Uref, 14 avril 1997. Avignon.

[Adda97A] G. Adda, M. Adda-Decker, J. L. Gauvain, and L. Lamel. Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 35–40, Avignon, France, April 1997.

[Lecomte97] J. Lecomte, G. Adda, J. Mariani, P. Paroubek, and M. Rajman. Progress report on the grace evaluation program for french part-of-speech taggers. In *Salt workshop on evaluation in speech and language technology*, Sheffield, June 17-18 1997.

[Paroubek97] P. Paroubek, G. Adda, J. Mariani, and M.Rajman. Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 245–252, Avignon, France, April 1997.

1996

[Adda96] G. Adda, Séminaire ARPA 1996 sur la Reconnaissance de la Parole Francil. Réseau FRANcophone de l'Ingénierie de la Langue, Numéro 3, avril 1996.

[AddaDecker96A] M. Adda-Decker, G. Adda, L.F. Lamel, and J.-L. Gauvain. Developments in large vocabulary, continuous speech recognition of German. In *Proceedings of the IEEE-ICASSP*, Atlanta, May 1996.

[AddaDecker96B] M. Adda-Decker, L.F. Lamel, J.-L. Gauvain, and G. Adda. Activities in multilingual speech recognition at LIMSI. In *Proc. of the CRIM/FORWISS Workshop on Progress and Propects of Speech Research and Technology*, October 1996. invited.

[Lamel96A] L. Lamel and G. Adda. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *International Conference on Speech and Language Processing*, pages 6–9, Philadelphia, October 1996.

[Lamel96B] L. Lamel, M. Adda-Decker, J.L. Gauvain, and G. Adda. Spoken language processing in a multilingual context. In *International Conference on Speech and Language Processing*, pages 2203–2206, Philadelphia, October 1996. invited paper.

[Lamel96C] L. Lamel, G. Adda, and M. Adda-Decker. Les lexiques de prononciation dans les systèmes de reconnaissance de la parole. In *Proc. Séminaire GDR-PRC CHM Lexique et communication parlée*, pages 1–10, Toulouse, October 1996.

[Lamel96D] L. Lamel, M. Adda-Decker, G. Adda, and J.-L. Gauvain. Reconnaissance multilingue de grands vocabulaires. In H. Meloni, editor, *Fondements et Perspectives en Traitement Automatique de la Parole*. AUPELF-UREF, 1996.

[Gauvain96A] J.-L. Gauvain, L. F. Lamel, G. Adda, and D. Matrouf. Developments in continuous speech dictation using the 1995 ARPA NAB news task. In *Proceedings of the IEEE-ICASSP*, volume I, pages 73–76, Atlanta, May 1996.

[Gauvain96B] J.L. Gauvain, L. F. Lamel, G. Adda, and D. Matrouf. The LIMSI 1995 Hub3 System. In *Proceedings of ARPA Spoken Language Technology Workshop*, February 1996.

1995

[Adda95] G. Adda, P. Blache, J. Mariani, P. Paroubek, and M. Rajman. Action GRACE. mise en place du paradigme d'évaluation. application au domaine de l'analyse morpho-syntaxique.b In Philippe Blache, editor, *Le Traitement Automatique du Langage Naturel, 14, 15 et 16, Marseille*, pages 72–77. GDR-PRC Communication Homme-Machine, Pôle Langage Naturel, June 1995.

[Lamel95] L.F. Lamel, M. Adda-Decker, G. Adda, J.L. Gauvain, Reconnaissance multilingue de grands vocabulaires Ecole thématique "Fondements et perspectives en Traitement Automatique de la Parole". CNRS, Marseille-Luminy, 17-25 juillet 1995

1994

[Jardino94A] M. Jardino and G. Adda. Automatic determination of a stochastic bi-gram class language model. In *Proc. of ICGI-94, Alicante, Espagne*, page 57, 1994.

[Jardino94B] M. Jardino, G. Adda. Automatic determination of a stochastic bi-gram class language model. Lecture Notes in Computer Science, Rafael C. Carrasco, Jose Oncina eds., Volume 862 / 1994 Springer Berlin / Heidelberg.

[Gauvain94A] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Continuous speech dictation system in French. In *International Confe-*

rence on Speech and Language Processing, Yokohama, Japan, September 1994.

[Gauvain94B] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Continuous speech dictation system at LIMSI. In *Proc.of the CRIM/FORWISS Workshop on Progress and Propects of Speech Research and Technology, Munich*, September 1994.

[Gauvain94C] J.-L. Gauvain, L. F. Lamel, G. Adda, and J. Mariani. Recent progress in speech-to-text conversation at LIMSI. In *Esprit Speech Project Workshop book*. Springer Verlag, 1994.

[Gauvain94D] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Communication*, 15 :21–37, September 1994.

[Gauvain94E] J.-L. Gauvain, L. F. Lamel, G. Adda, and J. Mariani. Speech-to-text conversion in french. *Int. J. Pat. Rec and A. I.*, 8(1) :99–131, 1994.

[Gauvain94F] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. The LIMSI continuous speech dictation system. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 319–324, March 1994.

[Gauvain94G] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. The LIMSI continuous speech dictation system : Evaluation on the arpa wall street journal task. In *Proceedings of the IEEE-ICASSP*, volume 1, pages 557–560, Adelaide, April 1994.

[Gauvain94H] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. The LIMSI nov93 wsj system. In *Proceedings of ARPA Workshop on Spoken Language Technology*, Plainsboro, N.J., March 6-8 1994

1993

[Jardino93A] M. Jardino and G. Adda. Language modeling for CSR of large corpus using automatic classification of words. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, page 1191, Berlin, September 1993.

[Jardino93B] M. Jardino and G. Adda. Automatic word classification simulated annealing. In *Proceedings of the IEEE-ICASSP*, page II 41, Minneapolis, April 1993.

[Gauvain93B] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Large vocabulary speech recognition in English and French. In *Proceeding of IEEE Workshop on Automatic Speech Recognition*, December 1993.

[Gauvain93A] J-L Gauvain, L. Lamel, and G. Adda. LIMSI nov92 evaluation. In *Darpa spoken language systems technology workshop*, MIT, Cambridge, MA, January 1993.

[Gauvain93C] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker-independent continuous speech dictation. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, Berlin, September 1993.

1992

[AddaDecker92] M. Adda-Decker and G. Adda. Experiments on stress-dependent phone modelling for continuous speech recognition. In *Proceedings of IEEE-ICASSP*, San Francisco, 23-26 March 1992.

[Nait92] M. Nait-Lahcen, G. Adda, and S. Bornerand. Une méthode centiseconde pour la reconnaissance d'un grand vocabulaire de mots isolés. In *19e JEP*, Bruxelles, 19-22 mai 1992.

1990

[Vittorelli90] V. Vittorelli, G. Adda, J. Mariani, and al. Polyglot : multilingual speech recognition and synthesis. In *ICSLP90*, Kobe,Japan, 18-22 November 1990.

1988

[Adda88] G. Adda. algorithme de conversion sténotype-graphème. Journées de présentation du système de conversion sténotype-graphème pour le sous-titrage d'émissions télévisées CCETT, Rennes, octobre 1988.

1987

[Adda87] G. Adda, M. Eskenazi, and P-E. Stern. The use of rough spectral features for large vocabulary recognition. In *European Conference on Speech Technology*, volume 1, pages 171–175, Edinburgh, September 1987.

[Adda87B] G. Adda. Reconnaissance de grands vocabulaires : une étude syntaxique et lexicale. Thèse de docteur-ingénieur, Université de Paris-Sud, Orsay, décembre 1987.

1986

[Adda86] G. Adda, M. Eskenazi, and P-E. Stern. Reconnaissance de grands vocabulaires : utilisation et évaluation de traits grossiers. In *15ème JEP*, pages 219–223, Aix-en-Provence, 1986.

1985

[Neel85B] F. Néel, G. Adda, and al. Problèmes liés aux sous-titrages d'émissions télévisées avec léger différé. *Handitec*, 1985.

[Neel85A] F. Néel, G. Adda, and al. Problèmes liés aux sous-titrages d'émissions télévisées avec léger différé. In *Colloque francophone sur la technologie au service des personnes handicapees*, Paris, décembre 1985.

[Bellilty85] D. Bellilty, G. Adda, and al. Dictée vocale en mots isoles. In *14ème JEP*, pages 231–233, Paris, 1985.

1984

[Adda84A] G. Adda and al. Transcription de sténotypes en français écrit. In *13ème JEP*, pages 15–16, Bruxelles, 1984.

[Adda84B] G. Adda, A. Andreewsky, J. Avrain, E. Bsalis, M. Desi, C.E. Fluhr, R. Harani, J.J. Mariani, F. Neel, F. Poirier, Les problèmes lexico-syntaxiques en reconnaissance de la parole continue Symposium soviético-français sur " Le dialogue acoustique de l'homme avec la machine ". Moscou, septembre 1984, 1984

1983

[Adda83B] G. Adda, C. Fluhr, C. Morel, F. Neel, Utilisation d'un système à saisie rapide dans une source de sous-titrage ANTIOPE : un exemple des améliorations que peut apporter l'introduction de l'intelligence artificielle dans les équipements de production vidéographique Rev. Radiodiffusion-Télévision, Nř80, nov-déc 1983, 1983.

[Adda83A] G. Add and al. Utilisation d'un système automatique de transcription du code sténotypique en français écrit pour le sous-titrage des émissions de télévision. In *Journées Telemat*, Marseille, juin 1983.

Bibliographie autres auteurs

Bibliographie

[Andreewsky78] A. Andreewsky et al., Une expérience d'aide linguistique à la reconnaissance automatique de la parole. Note CEA N. 2055, octobre 1979.

[Baker75] J.K. Baker, Stochastic modeling for automatic speech understanding. Speech Recognition, Academic Press, p. 521-542, 1975.

[Barras01] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, Transcriber : development and use of a tool for assisting speech corpora production Speech Communication, 33(1), pp. 5-22, 2001.

[Bellegarda04] J.R. Bellegarda Statistical language model adaptation : review and perspectives Speech Communication, **42**(1) p. 93-108, 2004.

[Bellman57] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.

[Bonneau-Maynard06] H. Bonneau-Maynard et al., Results of the French Evalda-Media evaluation campaign for literal understanding Proc. LREC, Gènes, 2006.

[Bonnet86] A. Bonnet, J.-P. Haton, J.-M. Truong-Ngoc, Systèmes-experts : vers la maîtrise technique InterEditions, Paris, 1986.

[Bourlard96] H. Bourlard, H. Hermansky, N. Morgan, Towards increasing speech recognition error rates. Speech Communication 18(3), mai 1996, pp. 205-231.

[Brown92] P.F. Brown, V.J. Della Pietra, J.C. Lai, P.V. de Souza, R.L. Mercer Class-based n-gram models of natural language. Computational Linguistics 18(4) :467--479, 1992.

[Chase98] L. Chase, A review of the American switchboard and callhome speech recognition evaluation programs. First International Conference on Language Resources and Evaluation, vol. II, p. 789-793, Grenade, mai 1998.

[Chen98] S. Chen and J. Goodman, An empirical study of smoothing techniques for language modeling, Technical report, Harvard University, 1998.

[Chollet82] G. Chollet and C. Gagnoulet, Evaluating speech recognizers and data bases using a reference system. Proceedings of the IEEE-ICASSP, Paris, 1982.

[Crystal77] T.H. Crystal, M.K. Hoffmann, A.S. House, Statistics of phonetic category representation of speech for application ro word recognition Institute for Defense Analyses/CRD, Princeton, New Jersey, septembre 1977, working paper num. 528.

[Debili77] F. Debili, Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage. Thèse de docteur-ingénieur, Paris VII, septembre 1977.

[Dempster77] A.P. Dempster, M.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm Journal of the Royal Statistical Society Series B (methodological). **39** :1-38, 1977.

[Derouault85] A.-M. Derouault, Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques. Thèse d'état, Paris VII, avril 1985.

[Derouault86] A.-M. Derouault et B. Merialdo, Natural Language Modeling for Phoneme-to-Text Transcription IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6), pp. 742-749, novembre 1986

[Doddington81] G.R. Doddington, and T.B. Schalk, Speech Recognition : Tuming Theory to Practice IEEE Spectrum, Sept. 1981, p. 26-32.

[Dolmazon97] J.M. Dolmazon et al., ARC B1 - Organisation de la 1e campagne AUPELF pour l'évaluation des systèmes de dictée vocale. 1ères JST FRANCIL, Avignon, April 1997.

[Durand03] J. Durand, B. Laks, C. Lyche Le projet Phonologie du français contemporain (PFC). La Tribune Internationale des Langues Vivantes 33 3-9, 2003.

[Dusan05] S. Dusan and L.R. Rabiner, On Integrating Insights from Human speech Perception into Automatic speech recognition. Proceeding of Interspeech, Lisbonne , Septembre 2005.

[Federico96] M. Federico, Bayesian Estimation methods for n-gram language model estimation. Proceeding of ICSLP'96, pp. 240-243, 1996.

[Federico99] M. Federico, Efficient language model adaptation through MDI estimation Proceeding of Eurospeech'99, pp. 1583-1586, 1999.

[Fiscus97] J. G. Fiscus, A post-processing system to yield word error rates : Recognizer Output Voting Error Reduction (ROVER). Proceedings ASRU Workshop, 1997.

[Galliano09] Sylvain Galliano, Guillaume Gravier, Laura Chaubard, The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. Proceedings Interspeech, 6-10 Septembre, Brighton, pp. 2583-2586, 2009.

[Garofolo00] J.S. Garofolo, C.G.P. Auzanne, E.M. Voorhees, The TREC Spoken Document Retrieval Track : A Success Story, in Text Retrieval Conference (TREC) 8, pp. 16-19, 2000.

[Geoffrois08] E. Geoffrois, An Economic View on Human Language Technology Evaluation. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, 28-30 mai 2008.

[Geoffrois08b] E. Geoffrois, Evaluation in Quaero. Quaero/imageCLEF workshop, Aarhus, Danemark, septembre 2008.

[Geoffrois09] E. Geoffrois, L ?évaluation des systèmes de traitement des données non structurées Séminaire DGA 1er juillet 2009, Paris.

[Goldwater10] S. Goldwater, D. Jurafsky, and C. D. Manning. Which words are hard to recognize ? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52, 181-200. 2010.

[Gonzalo10] J. Gonzalo, Benchmarking and Evaluation Campaigns : the good, the bad and the metrics. Presentation au Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods, LREC2010, Valletta, Malta, May 23, 2010

[Gravier04] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. The ESTER evaluation campaign of Rich Transcription of French Broadcast News . Proc. Language Evaluation and Resources Conference, Lisbon, 2004.

[Habert97] Benoît Habert, Adeline Nazarenko et André Salem, Les linguistiques de corpus. Armand Colin, Paris, 1997.

[Habert06] B. Habert, Portrait de linguiste(s) à l'instrument. À la quête du sens : Études littéraires, historiques et linguistiques en hommage à

Christiane Marchello-Nizia, Céline Guillot, Serge Heiden et Sophie Prévost éditeurs,Foreign Language Study , 2006.

[Hemphil90] C.T. Hemphil, J.J. Godfrey and G.R Doddington, ATIS spoken language systems pilot corpus, Proceedings of DARPA Speech and Natural Language Workshop, Pittsburgh, PA, June 1990.

[Hermansky04] H. Hermansky and N. Morgan, Show What You Know : Musings on the Reporting of Negative Results in Speech Recognition Research, Invited Editorial Note in Journal of Negative Results in Speech and Audio Sciences, 2004.

[Huckvale96] M. Huckvale, Learning from the experience of building automatic speech recognition systems. UCL working papers, Speech Hearing and Language, 1996.

[Huckvale97] M. Huckvale, 10 things engineers have discovered about speech recognition. NATO ASI Workshop on Speech Pattern Processing, Jersey, 1997

[Huckvale98] M. Huckvale, Opportunities for Re-convergence of Engineering and Cognitive Science Accounts of Spoken Word Recognition. Proc. IOA Conference Speech and Hearing, Windermere, November 1998.

[Huckvale01] M. Huckvale, Learning on the job : the application of machine learning within the speech decoder. Institute of Acoustics, Workshop on Innovation in Speech processing. stratford-on-Avon, mai 2001.

[Huckvale02] M. Huckvale, Speech Synthesis, Speech Simulation and Speech Science Proc. International Conference on Speech and Language Processing, pp 1261-1264, Denver, 2002.

[Iquierdo07] E. Iquierdo, J. Bennois-Pineau, R. André-Obrecht, Special Issue on Content Based Multimedia Indexing and Retrieval Signal Processing : Image Communication, vol. 22, Issues 7-8, Elsevier, 2007.

[Jelinek76] F. Jelinek Continuous Speech Recognition by Statistical Methods. Proc. of the IEEE, **64**(4), pp. 532-536, 1976.

[Jelinek90] F.J. Jelinek, Self-organized language modeling for speech recognition Alex Waibel and Kay-Fu Lee, (éds), Readings in Speech Recognition p. 450–505. Morgan Kaufmann, Los Altos, CA, 1990.

[Jelinek97] F.J. Jelinek, Statistical Methods for Speech Recognition. The MIT Press, 1997.

[Jones05] Karen Spärck Jones, Some points in a Time. ACL Lifetime achievement Award, Computational Linguistics, 31(1), MIT Press Cambridge, MA, USA, March 2005.

[Julliand70] A. Julliand, D. Brodin, C. Davidovitch, Frequency dictionary of French words. The Hague, Mouton, 1970

[Katz87] S.M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Trans. Acoustics, Speech & Signal Processing, ASSP-35(3), p. 400-401, mars 1987.

[Laasri88] H. Lâasri, B. Maître, T. Mondot, F. Charpillet, J.-P. Haton, ATOME : A blackboard architecture with temporal and hypothetical reasoning. Proceedings of the 8th European conference on Artificial intelligence ECAI '88, 1988.

[Lamel91] L.F. Lamel, J.-L. Gauvain, M. Eskenazi BREF, a Large Vocabulary Spoken Corpus for French. Proceedings of Eurospeech, Genève, septembre 2001

[Lamel92] L.F. Lamel et J.-L. Gauvain, Continuous speech recognition at LIMSI. DARPA ANNT Speech Program, Sep. 1992.

[Leggetter95] C.J. Leggetter and P.C. Woodland, Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models Computer Speech and Language, vol. 9, p. 171-185, 1995.

[Levinson95] S E Levinson, Speech recognition technology : a critique. Voice Communication between Humans and Machines, David B. Roe and Jay G. Wilpon, Editors,National Academy of Sciences, 1994, pp. 159-164.

[Liberman10a] M. Liberman, The Future of computational linguistics : or, what would Antonio Zampolli do ? Antonio Zampolli Prize speech, presented at LREC2010, Valletta, Malta, May 21, 2010

[Liberman10b] M. Liberman, Language resources and evaluation in the US : from fraud prevention to community building. Présentation aux Journées DGA traitement de la parole, du langage et des documents multimedias, 6-7juillet 2010

[Liberman10c] M. Liberman, Obituary Fred Jelinek Computational Linguistics December 2010, Vol. 36, No. 4 : 595 ?599.

[Lippmann97] R. Lippmann, Speech recognition by machines and humans. Speech Communication, vol. 22, pp. 1-16, 1997.

[Maekawa00] K. Maekawa, H. Koiso, S. Furui, H. Isahara, Spontaneous speech corpus of Japanese. Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000), Athènes, pp. 947-952, 2000.

[Makhoul95] J. Makhoul, R. Schwartz, State of the art in continuous speech recognition. Proceedings of The National Academy of Sciences 1995 Oct 24 ;92(22) :9956-9963.

[Mariani09] J.J. Mariani (éditeur), Language and Speech Processing, ISBN :978-1-84821-031-8, Wyley-ISTE, Janvier 2009.

[Merialdo85] A.-M. Derouault, B. Mérialdo, Un système de transcription automatique de la sténotypie. In *Actes des 14e JEP*, Paris, 1985.

[Moore77] R.K. Moore, Evaluating speech recognizers. IEEE-ASSP, pp. 178-183, 1977

[Moore01] R.K. Moore and A. Cutler, Constraints on theories of human vs. machine recognition of speech. Proceedings SPRAAC workshop on Human Speech Recognition as pattern Classification, Max-Planck-nstitute for Psycholinguistics, Nijmegen 11-13 juillet 2001.

[Moore03] R.K. Moore, A comparison of the data requirements of automatic speech recognition systems and human listeners, Proceedings of Eurospeech, pp. 2582-2584, Berlin septembre 2003

[Moore05a] R.K. Moore Cognitive Informatics : The Future of Spoken Language Processing ? Keynote talk, SPECOM - 10th Int. Conf. on Speech and Computer, Patras, Greece, 17-19 October 2005.

[Moore05b] R.K. Moore, Research Challenges in the Automation of Spoken Language Interaction. Keynote talk, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005), Aalborg University, Denmark, 10-11 November 2005.

[Norvig11] P. Norvig, On Chomsky and the Two Cultures of Statistical Learning. http ://norvig.com/chomsky.html

[Pallett82] David S. Pallett, Workshop on Standardization for Speech I-O technology Gaithersburg, Maryland, 1982

[Pallett85] David S. Pallett, Performance Assessment of Automatic Speech Recognizers, Journal of Research of the National Institute of Standards and Technology, Vol. 90, Num. 5, septembre-octobre 1985

[Pallett92] D.S. Pallett, J.G Fiscus and J.S. Garofolo, Resource management corpus : September 1992 test set benchmark results. Proceedings of the ARPA Workshop on Continuous Speech Recognition, Stanford, CA, septembre 1992.

[Pallett93] D.S. Pallett, J.G Fiscus and J.S. Garofolo, Benchmark tests for the darpa spoken language program. Proceedings of ARPA Workshop on Human Language Technology, Princeton, NJ, 1993.

[Pallett95] D.S. Pallett et al., 1994 benchmark tests for the DARPA spoken language program. Proceedings of ARPA Spoken Language Technology Workshop, Austin, TX, 1995.

[Pallett97] D.S. Pallett, J.G Fiscus, W.M. Fisher and J.S. Garofolo, Use of broadcast news materials for speech recognition benchmark tests. Proceedings of the European Conference on Speech Technology, EuroSpeech, vol. IV, p. 1903-1906, Rhodes, 1997.

[Papineni02] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU : a method for automatic evaluation of machine translation in ACL-2002 : 40th Annual meeting of the Association for Computational Linguistics pp. 311 ?318, 2002.

[Paroubek00] P. Paroubek, Language Resources as by-Product of Evaluation : the MULTITAG example. Second International Conference on Language Resources and Evaluation (LREC2000), pp. 151-154, Athènes, juin 2000.

[Pianta10] E. Pianta, The role of evaluation for research and development of information and communication technologies. Présentation aux Journées DGA traitement de la parole, du langage et des documents multimedias, 6-7 juillet 2010

[Pierce69] J.R. Pierce, Whither Speech Recognition ? Letter to the Editor of J.Acoutic Soc.Amer, 46, pp. 1049-1051, 1969.

[Pisoni85] D.B. Pisoni, H.C. Nusbaum, P.A. Luce, L.M. Slowiaczek, Speech Perception, word recognition, and the structure of the lexicon. Speech Communication, Vol.4, num 1-3, pp. 75-95, 1985.

[Price88] P.Price, W.M Fisher, J. Bernstein, D.S. Pallett, The DARPA 1000-word Resource Management Database for Continuous Speech Recognition, Proceedings of International Conference on Acoustics, Speech, and Signal Processing, New York, avril 1988.

[Prouts80] B. Prouts, Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur Thèse de Docteur-Ingénieur, Université de Paris XI, Orsay, 1980.

[Rabiner93] L.R Rabiner, B.H. Juang, Fundamentals of Speech Recognition. Prentice-Hall, 1993.

[Rosenfeld94] R. Rosenfeld. Adaptive Statistical Language Modeling : A Maximum Entropy Approach. PhD thesis, School of Computer Science - Carnegie Mellon University, Pittsburgh, PA 15213, 1994.

[Rosenfeld00] R. Rosenfeld, Two decades of Statistical Language Modeling : Where Do We Go From Here ? Proceedings of the IEEE ASSP **88**(8) :1270-1278, août 2000.

[Scharenborg05a] O. Scharenborg, D. Norris, L. ten Bosch, J.M. McQueen, How Should a Speech Recognizer Work ? Cognitive Science 29, pp. 867 ?918, 2005

[Scharenborg05b] O. Scharenborg, Parallels between HSR and ASR : How ASR can Contribute to HSR. Proceeding of Interspeech, Lisbonne , Septembre 2005.

[Schwenk07] H. Schwenk, Continuous space language models Computer Speech & Language, Volume 21, Issue 3, pp. 492-518 , July 2007.

[Snover06] M. Snover, B. Dorr, R. Schwartz, L. Micciulla et J. Makhoul, A Study of Translation Edit Rate with Targeted Human Annotation, Proc AMTA 2006.

[Taylor80] M.M. Taylor, Issues in the evaluation of speech recognition systems. Technical report, Defense & Civil Intitute of Environmental Medicine, 1980.

[Torreira10] F. Torreira, M. Adda-Decker, M. Ernestus , The Nijmegen corpus of casual French. Speech Communication, **52** p. 201-212, 2010.

[tenBosch05] L. ten Bosch, O. Scharenborg, ASR Decoding in a Computational Model of Human Word Recognition, Proceeding of Interspeech, Lisbonne , Septembre 2005.

[Vasilescu09] I. Vasilescu, M. Adda-Decker, L. Lamel, P. Hallé, A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English. Proceedings Interspeech, 6-10 Septembre 2009, Brighton, pp. 144-147.

[Walker00] M. Walker, L.Y. Hirschman and J.Y Aberdeen, Evaluation for DARPA Communicator spoken dialog systems Proc. LREC, Athènes, 2000.

[Young97] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken A.J. Robinson, and P.C. Woodland. Multilingual large vocabulary speech recognition : the european SQALE project. Computer Speech and Language, 11(1) :73 ?89, January 1997.