Reconnaissance de la parole dans un contexte de cours magistraux : évaluation, avancées et enrichissement

Salima Mdhaffar

THÈSE DE DOCTORAT DE

LE MANS UNIVERSITÉ
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Salima Mdhaffar

Reconnaissance de la parole dans un contexte de cours magistraux : évaluation, avancées et enrichissement

Thèse présentée et soutenue à « Avignon » en Visioconférence, le « 01/07/2020 »
Unité de recherche : Laboratoire d'Informatique de l'Université du Mans (LIUM)
Thèse N° : 2020LEMA1008

Composition du Jury :

Rapporteurs :	Georges Linarès	Professeur	Université d'Avignon
	Irina Illina	Maître de conférences, HDR	Université de Lorraine
Examinateurs :	Olivier Galibert	Ingénieur de recherche	LNE
	Camille Guinaudeau	Maître de conférences	Université de Paris Saclay
	Sylvain Meignier	Professeur	Le Mans université
Dir. de thèse :	Yannick Estève	Professeur	Université d'Avignon
Co-encdrants :	Antoine Laurent	Maître de conférences	Le Mans université
	Nicolas Hernandez	Maître de conférences	Université de Nantes
	Solen Quiniou	Maître de conférences	Université de Nantes

Remerciements

C'est un plaisir de pouvoir adresser ici ma profonde reconnaissance à tous ceux qui m'ont aidé à aller au bout de cette aventure.

Je tiens tout d'abord à exprimer ma gratitude et mes vifs remerciements à mon directeur de thèse M. Yannick Estève, Professeur à l'université d'Avignon, pour son leur accueil bienveillant, sa disponibilité, ses encouragements, sa gentillesse, ses conseils précieux et son apport scientifique. Il a su me soutenir et m'encourager tant sur le plan professionnel que sur le plan personnel.

Mes remerciements vont également à mes encadrants, M. Antoine Laurent, Maître de conférences à Le Mans Université, M. Nicolas Hernandez, Maître de conférences à l'université de Nantes et Mme Solen Quiniou, Maître de conférences à l'université de Nantes pour leurs conseils, aides et encouragements, ainsi que l'intérêt continu qu'ils ont porté à mon travail.

Je remercie également M. Richard Dufour, Maître de conférences à l'université d'Avihnon, pour sa disponibilité, ses orientations et sa collaboration. Nos discussions, toujours très fructueuses, ont beaucoup compté dans l'orientation de mes recherches et l'aboutissement de cette thèse.

Je tiens aussi à exprimer mes remerciements à M. Georges Linarès, Professeur à l'université d'Avignon, ainsi que Mme Irina Illina, Maître de conférences à l'université de Lorraine, qui ont accepté de juger ce travail et d'en être les rapporteurs. Je remercie également M. Olivier Galibert, Ingénieur de recherche au laboratoire national de métrologie et d'essais, M. Sylvain Meignier, Professeur à Le Mans université, et Mme Camille Guinaudeau, Maître de conférences à l'université de Paris Saclay, pour avoir accepté de juger cette thèse et pour l'intérêt qu'ils ont porté à mon travail.

Mes remerciements s'adressent aussi aux membres du projet ANR PASTEL, avec lesquelles j'ai pu avoir de nombreux échanges et collaborations.

Cette thèse a eu lieu dans trois laboratoires académiques : le laboratoire d'informatique de l'université du Mans, le laboratoire d'informatique de l'université de Nantes et le laboratoire d'informatique d'Avignon. Mes remerciements sincères vont également à toutes les personnes que j'ai pu côtoyer quotidiennement au sein de ses trois équipes. Je pense particulièrement à Antoine Caubrière, Kevin Vythelingum, Mercedes Garcia Martinez, Fethi Bougares, Sahar Ghannay, Natalia Tomashenko, Manon Macary, Amira Barhoumi, Adrien Bardet, Emmanuelle Billard, Étienne Micoulaut, Abdessalem Bouchekif, Ozan Caglayan, Grégor Dupuy, Anne Cécile Erreau, Aicha Bakki, Jihen Karoui, Rajoua Anane, Loïc Barrault, Nathalie Camelin, Nicolas Dugué, Marie Tahon, Sylvain Meignier, Malik Koné, Basma El Amal Boussaha, Amir Hazem, Florian Boudin, Randa Abdelmoneem, Sondes Abderrazek, Afef Arfaoui, Adrien Gresse, Sarkis Moussa, Tesnim Naceur, Mathias Quillot, Dina Tarek, Malek Hajjem, Gaelle Laperriere, Driss Matrouf, Titouan Parcollet, Céline, Carlos González, Anaïs Chanclu, Cyril Sahuc.

Rien n'aurait été possible sans le soutien et l'encouragement de mon père, ma mère et mes soeurs. Mes pensées vont aussi à ma famille d'accueil au Mans, Jannick Lambert et Gérard Lambert, qui étaient à mes côtés durant cette thèse.

Je tiens à remercier très vivement mes amis Leila, Mariem Khelifa, Mariem Fourati, Nadia, Tesnim, Sondes, Bilel, Saif, Saleh, Salah, Jawhar, Hajer de leur présence fidèle et vraie. Les moments passés ensemble furent parmi les meilleurs et le seront encore.

Enfin, merci à tous ceux qui, de prés ou de loin, m'ont encouragé et soutenu tout au long de ce travail.

A vous tous, Merci !

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte d'étude

Les pratiques d'enseignement sont actuellement en profonde mutation, offrant des services pédagogiques innovants et des modes diversifiés d'accès aux contenus pédagogiques. Les universités sont particulièrement concernées par ces problématiques d'accès aux contenus, dont une partie des réponses se trouvent dans les environnements collaboratifs qu'elles mettent en place, tels que les environnements numériques de travail (ENT). Les enseignants ont alors l'occasion de mettre à disposition des étudiants un ensemble de ressources liées aux cours qu'ils dispensent (supports de cours, exercices, vidéos...).

Les avancées récentes dans le domaine de la reconnaissance de la parole et du traitement automatique du langage naturel ont rendu possible le développement de nouvelles applications à vocation pédagogique. Dans le contexte d'un cours magistral, le discours de l'enseignant constitue une ressource très importante à partir de laquelle plusieurs applications peuvent être envisagées. Parmi ces applications, nous pouvons citer (1) la structuration automatique pour faciliter la navigation et la recherche au sein du contenu du cours, (2) l'enrichissement de documents pédagogiques avec des liens vers des ressources pédagogiques externes, (3) la traduction automatique dont le but est d'aider les étudiants étrangers à mieux comprendre le cours, etc.

Différents travaux dans la littérature ont montré l'interêt de développer de telles applications dans un cadre pédagogique comme ceux de (HO et al., 2005 ; HWANG et al., 2012 ; SHADIEV et al., 2014).

La majorité de ces applications pédagogiques exploite la sortie d'un système de reconnaissance de la parole. Dans ce contexte, on attend alors des systèmes de reconnaissance automatique de la parole qu'ils soient capables de transcrire précisément le discours de l'enseignant.

Problématique et objectifs de la thèse

Les systèmes de reconnaissance de la parole sont généralement développés pour une tâche donnée dans un contexte d'utilisation connu, comme par exemple la transcription d'émissions radiophoniques et télévisées, ou la transcription de conversations téléphoniques. Dans la pratique, quand les contextes d'utilisation sont similaires à celles de l'apprentissage, ces systèmes s'avèrent efficaces. Par contre, ces systèmes seront moins efficaces lorsqu'il s'agit de transcrire des documents portant sur d'autres sujets. Les données spécifiques à chaque situation ne sont pas toujours volumineuses, l'adaptation permet d'exploiter des données spécialisées dans le domaine visé, généralement rares, en les combinant avec des données hors-domaine existantes et disponibles en grande quantité. Dans le cadre de ce travail, il s'agit dans un premier temps de fournir la transcription de cours magistraux portant sur divers domaines. Il est important d'avoir recours à des techniques d'adaptation des modèles de langage afin d'obtenir de meilleures transcriptions, puisque chaque cours présente un thème différent.

Pour adapter le modèle de langage et enrichir le vocabulaire, il faut disposer d'une source d'information contenant les mots manquants dans leur contexte linguistique. Les données du web constituent une source très intéressante de données textuelles de taille importante. Dans la première partie de cette thèse, nous proposons un protocole d'adaptation qui collecte des données web en utilisant les diapositives préparées par un enseignant pour son cours.

La métrique la plus utilisée pour évaluer la qualité d'un système de reconnaissance de la parole est le taux d'erreurs sur les mots (PALLETT, 2003). Cette métrique consiste à compter les erreurs selon des types prédéfinis qui sont l'insertion, la suppression et la substitution déterminées par un alignement qui minimise la distance de Levenshtein entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Cette métrique est très utile pour évaluer la performance globale des systèmes de transcription. Mais, dans le cadre d'une transcription de cours magistral, le but est de réduire les erreurs pour les phrases contenant des mots du nouveau domaine, afin par exemple de pouvoir s'appuyer sur ces mots pour apparier des données pédagogiques externes avec le contenu du cours. De plus, les systèmes de reconnaissance de parole sont souvent conçus comme une brique parmi d'autres applications de traitement de langage naturel qui utilisent les transcriptions automatiques pour effectuer d'autres tâches. Chaque brique utilisée a un impact sur la brique suivante, une erreur pouvant en entraîner une autre. Le but n'est pas de développer un SRAP pour qu'il obtienne le meilleur score indépendamment de la tâche visée, mais de développer un système qui va contribuer à la performance globale de l'application finale. Dans ce contexte, nous proposons un protocole d'évaluation qui permet d'évaluer la performance des mots du domaine dans des transcriptions multi-domaines ainsi la performance pour certaines tâches exploitant la transcription : la recherche d'informations et l'indexabilité.

L'utilisation de données issues du Web comme corpus d'adaptation a suscité beaucoup d'intérêt depuis quelques années. Toutefois, en raison du caractère évolutif d'Internet, les résultats d'adaptation peuvent ne pas être reproductibles d'une période temporelle à une autre. Dans la seconde partie de cette thèse, nous étudions la reproductibilité des résultats d'adaptation des modèles de langage sur une période d'un an.

En raison de l'absence d'informations structurelles telles que les limites de phrases et de paragraphes ou l'alignement avec les diapositives, les transcriptions automatiques de cours magistraux peuvent être fastidieuses à parcourir, ce qui rend difficile la navigation et la récupération d'informations pertinentes. Nous nous sommes donc intéressés à la structuration de la transcription. Nous intégrons l'information de changement de diapositives pour la segmentation thématique et nous proposons une méthode d'alignement des segments de transcription et des diapositives.

Le projet PASTEL

Financé par l'ANR [1] (Agence National de la Recherche), le projet PASTEL s'inscrit dans le domaine de la reconnaissance de la parole, le traitement automatique des langues naturelles et les environnements informatiques pour l'apprentissage humain.

L'objectif du projet est d'explorer le potentiel de la transcription automatique pour l'instrumentation de situations pédagogiques mixtes, où les modalités d'interaction sont présentielles ou à distance, synchrones ou asynchrones.

Quatre laboratoires académiques sont impliqués dans PASTEL :

- **LIUM :** le Laboratoire d'Informatique de l'Université du Maine ;
- **LS2N :** le Laboratoire des Sciences du Numérique de Nantes ;
- **ORANGE :** l'équipe CONTENT et l'équipe MAS ;
- **CREN :** le Centre de Recherche en Éducation de Nantes.

La collaboration entre ces équipes de recherche permet l'échange scientifique et technologique autour de quatre lots scientifiques. Chaque partenaire participe au moins à un lot. Les lots sont les suivants :

- **Lot 1 : Analyse des besoins et des usages et validation**
  Les conditions de réussite d'un projet reposent sur l'identification des besoins et des usages des utilisateurs et sur l'évaluation des différents prototypes proposés. L'objectif de ce lot est de préparer des scénarios pour les situations pédagogiques identifiées et d'analyser les conditions de réussite de ces scénarios. Il s'agit d'identifier les différents

1. http ://www.agence-nationale-recherche.fr/

éléments et leurs articulations (identifier les acteurs représentatifs, leur rôle et fonctions, les activités, ressources, outils, etc.) et de proposer et analyser des pistes d'intégration des scénarios dans les prototypes.

- **Lot 2 : Traitement automatique de la parole**
  Dans le cadre applicatif du projet PASTEL, les technologies de reconnaissance de la parole sont sollicitées pour traiter des discours très spécialisés, puisque liés à des cours spécifiques et portant sur divers sujets. Le système de reconnaissance de la parole devra être capable de transcrire avec une meilleure qualité des cours magistraux de disciplines variées dispensés par des enseignants différents. Ce lot vise à optimiser les performances de la reconnaissance automatique de la parole dans les deux principaux cas d'usage visés par le projet : la reconnaissance en différé, où toutes les ressources (données et serveurs de calcul) peuvent être utilisées dans un temps non strictement contraint, et la reconnaissance en temps réel, dont les implications sont très fortes en termes de qualité. L'adaptation des modèles de langage et l'évaluation de la transcription sont les axes de recherche les plus importants du lot 2.
- **Lot 3 : Enrichissement du matériel pédagogique**
  Les transcriptions automatiques peuvent également servir pour la recherche automatique de matériaux pédagogiques et d'informations complémentaires. En travaillant à la structuration automatique du discours de l'enseignant, par exemple en découpant ce discours en segments thématiques, il est possible d'extraire de ces segments des mots-clés, ou de caractériser ces segments par d'autres moyens, de manière telle qu'il soit possible de lier un segment thématique à un ensemble de sources d'informations disponibles par ailleurs et non produites par l'enseignant.
- **Lot 4 : Instrumentation de l'apprentissage et de la conception**
  L'objectif de ce lot est dédié à l'instrumentation des activités pédagogique de type cours magistral ou travaux dirigés. Il s'agit de proposer une plateforme contenant la transcription en temps réel et une chaîne éditoriale d'édition et de structuration de contenus pédagogiques en ligne.

La répartition des lots entre les différents équipes est représentée dans la figure 1. Cette thèse porte sur les lots 2 "Traitement automatique de la parole" et 3 "Enrichissement du matériel pédagogique".

Organisation du manuscrit

Ce document est organisé en deux grandes parties. Nous présentons, dans la première partie, l'état de l'art des systèmes de reconnaissance de la parole.

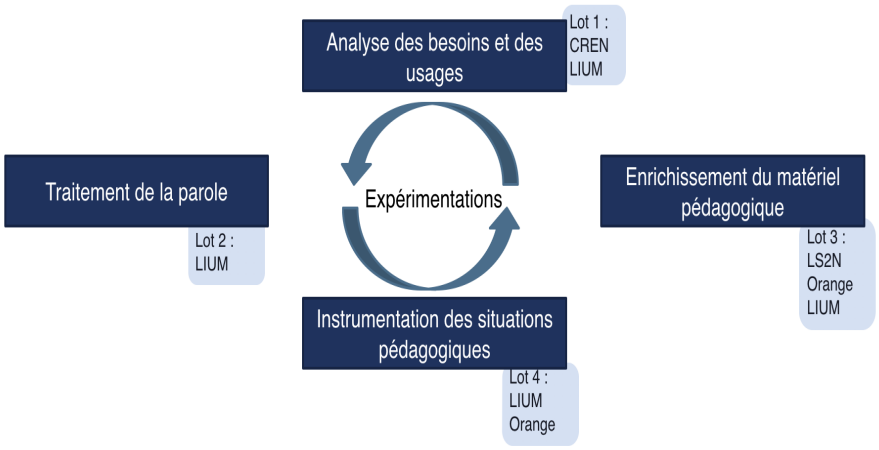


FIGURE 1 – Partition des lots du projet PASTEL

— Chapitre 1 : nous présentons le principe de fonctionnement d'un système de reconnaissance automatique de la parole et les composants fondamentaux pour construire un système de reconnaissance de parole à l'état de l'art.

— Chapitre 2 : ce chapitre présente un état de l'art des travaux de l'adaptation l'adaptation des modèles de langage n-grammes et neuronaux.

— Chapitre 3 : nous reprenons des notions de base liées à la segmentation thématique et nous présentons les méthodes de segmentation les plus citées dans la littérature.

— Chapitre 4 : nous présentons dans ce chapitre un état de l'art sur les travaux de traitement automatique de la parole dans le contexte d'éducation.

La seconde partie concerne le travail réalisé durant cette thèse.

— Chapitre 5 : Nous présentons le corpus PASTEL en détaillant le schéma d'annotation et nous détaillons le système de reconnaissance de la parole, à l'échelle de lequel tous les expériences sont menées dans cette thèse.

— Chapitre 6 : nous abordons la problématique de l'évaluation des systèmes de reconnaissance de parole pour l'adaptation des modèles de langage en proposant trois métriques d'évaluation.

— Chapitre 7 : nous proposons un travail axé sur la reproductibilité des résultats obtenus

par les systèmes de reconnaissance de la parole à l'aide de modèles de langage adaptés à partir de données collectées sur Internet.

— Chapitre 8 : nous proposons une structuration automatique de la transcription des cours magistraux.

Une conclusion clôt ce document en évoquant un résumé de nos travaux ainsi que les perspectives concernant les différents travaux réalisés.

PREMIÈRE PARTIE

État de l'art

Chapitre 1

Reconnaissance de la parole

Contents

1.1 Introduction

De nos jours, l'essor technologique et scientifique dans le domaine de la reconnaissance automatique de la parole permet de fournir des Systèmes de Reconnaissance Automatique

de la Parole (SRAP) performants dans différentes conditions d'utilisation. Cependant, ces systèmes restent sensibles à la variation de thèmes. En effet, la transcription automatique est encore généralement une séquence de mots restreinte seulement aux mots contenus dans le vocabulaire du SRAP.

Ces transcriptions représentent le matériau d'entrée pour de nombreuses applications telles que le résumé automatique, le dialogue homme-machine, la compréhension de la parole, la traduction de la parole, la détection d'entités nommées, le résumé automatique, la recherche d'informations. Les enjeux sont donc multiples pour la reconnaissance de la parole, puisque le SRAP peut être vu comme un composant d'un système plus important.

Dans ce chapitre nous présentons les principes de base d'un système de reconnaissance automatique de la parole, en détaillant le processus permettant de passer d'un signal de parole à la transcription de ce signal. Dans cette optique, nous présentons dans un premier temps l'architecture générale d'un SRAP, puis nous détaillons ses composants. Dans le cadre de cette thèse, où il s'agit dans un premier temps de fournir la transcription des cours magistraux portant sur divers domaines, il est important d'avoir recours à des techniques d'adaptation des modèles de langage afin d'obtenir de meilleures transcriptions. Nous détaillons alors dans ce chapitre la modélisation linguistique.

1.2 Transcription automatique de la parole

1.2.1 Principes généraux

Les systèmes de reconnaissance automatique de la parole ont pour objectif de produire la séquence de mots prononcée dans un signal de parole. Le fonctionnement de la plupart de ces systèmes s'appuie sur des principes probabilistes (JELINEK, 1976). Il s'agit de chercher la séquence de mots $W^* = w_1, w_2, ...w_n$ à partir d'un ensemble d'observations acoustiques $X = x_1, x_2, ...x_n$ maximisant la probabilité suivante :

$$W^* = \underset{W}{argmax}\ P(W|X) \tag{1.1}$$

où $P(W|X)$ est la probabilité d'émission de la séquence de mots W sachant X. En appliquant le théorème de Bayes, cela peut se décrire ainsi :

$$W^* = \underset{W}{argmax}\ \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

$P(X)$ est considérée constante et retirée de l'équation 1.2. Par conséquent, on obtient :

$$W^* = \underset{W}{argmax}\ P(X|W)P(W) \tag{1.3}$$

où $P(W)$ représente la probabilité a priori d'une séquence de mots W et $P(X|W)$ est estimée par le modèle acoustique. $P(X|W)$ correspond à la probabilité de rencontrer la séquence d'observation X lorsque la séquence de mots W est prononcée. La maximisation $argmax$ est réalisée par le processus de décodage. La figure 1.1 présente une schématisation générale du fonctionnement d'un SRAP et les sections suivantes décrivent chacun de ses composants.

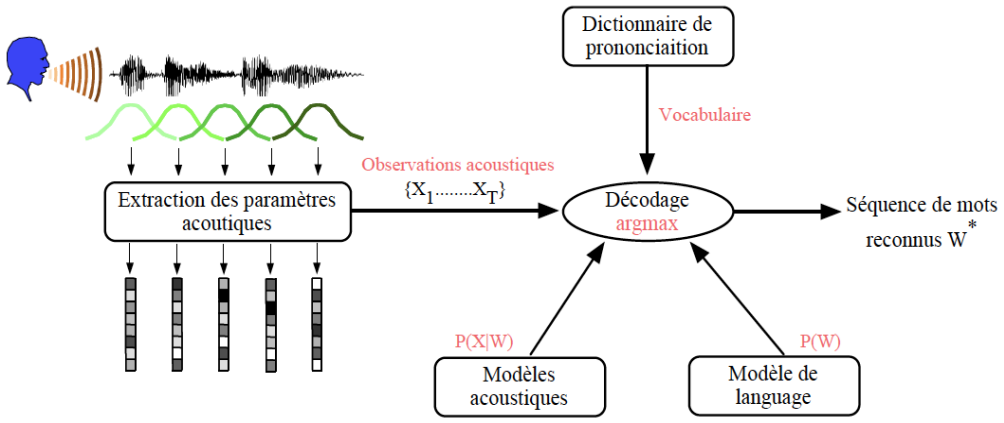


FIGURE 1.1 – Architecture d'un système de reconnaissance de la parole (GHANNAY, 2017)

1.2.2 Extraction de paramètres

Le signal audio contient plusieurs informations autres que le contenu linguistique tels que l'identité du locuteur, l'émotion du locuteur, la langue adoptée, les conditions d'enregistrement, l'environnement sonore.

Un SRAP a pour but d'extraire le contenu linguistique contenu dans le signal audio indépendamment des autres informations. L'extraction de paramètres s'effectue sur une fenêtre glissante à court terme dans laquelle le signal est considéré comme stationnaire, typiquement d'une longueur de 20 à 40 ms, avec un déplacement de 10 ms. En sortie de ce module, le signal est représenté comme une suite de vecteurs de paramètres qui sont appelés vecteurs acoustiques. La figure 1.2 illustre le processus d'extraction de paramètres à partir d'un signal audio.

Les techniques de paramétrisation les plus citées calculent les coefficients :

- MFCC (Mel-Frequency Cepstral Coefficients) (DAVIS et MERMELSTEIN, 1980),
- LPC (Linear Predictive Codes) (ABE, 1992),

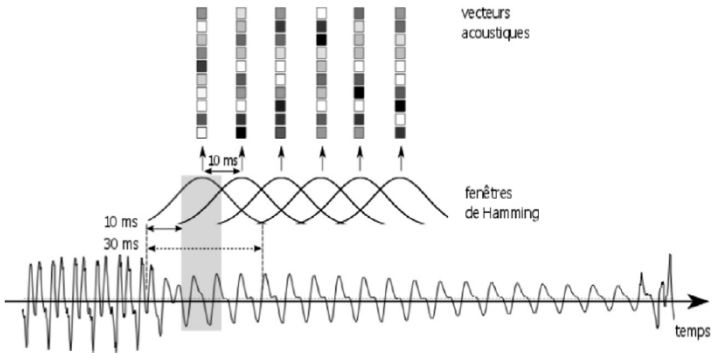


FIGURE 1.2 – Extraction de paramètres à partir d'un signal audio

- PLP (Perceptual Linear Prediction) (HERMANSKY et COX JR, 1991),
- LPCC (Linear Predictive Cepstral Coefficients) (MARKEL et GRAY, 1982),
- RASTA-PLP (Relative Spectral PLP) (HERMANSKY et al., 1992),
- TRAPS (HERMANSKY et SHARMA, 1999)
- Bottleneck (BN) (GRÉZL et al., 2007 ; YU et SELTZER, 2011),
- etc

1.2.3 Modèles acoustiques

Un modèle acoustique a pour objectif d'estimer la probabilité $P(X|W)$ définie dans la formule 1.3. Étant donnée une séquence de vecteurs de paramètres extraits du signal de parole via la phase d'extraction de paramètres (Section 1.2.2), le but des modèles acoustiques est de calculer la probabilité qu'une unité linguistique particulière (phonème, syllabe, mot, phrase, etc...) ait généré cette séquence.

Les modèles de Markov cachés (MMC, ou en anglais Hidden Markov Model, HMM) sont parmi les modèles les plus utilisés pour la modélisation acoustique du signal (JELINEK, 1976). Chaque unité acoustique est représentée par un modèle de Markov caché. Les unités de base modélisées par ces systèmes sont souvent les phonèmes en contexte qui, par leur concaténation, permettent de former des mots. Les modèles de Markov cachés sont des automates stochastiques à états finis (figure 1.3). Un MMC est caractérisé par les paramètres suivants :

- Un ensemble de N états,
- un ensemble A de probabilités de transition discrètes d'un état i à un état j ($A = \{a_{ij}\}$) où $i \leq j$ avec

$$a_{ij} = P(s_t = j | s_{t-1} = i), \tag{1.4}$$

où

$$\sum_{j=1}^{N} a_{ij} = 1 \; \forall i \in [1; N] \tag{1.5}$$

- un ensemble B de probabilités d'observation associées aux états où $b_i(o)$ indique la probabilité d'émettre l'observation o_t à partir d'un état i à l'instant t

$$b_i(o_t) = P(o_t|s_t = i) \tag{1.6}$$

- un ensemble Π de probabilités initiales π_i indiquant la probabilité d'être dans l'état i à l'instant initial avec

$$\sum_{i=1}^{N} \pi_i = 1 \tag{1.7}$$

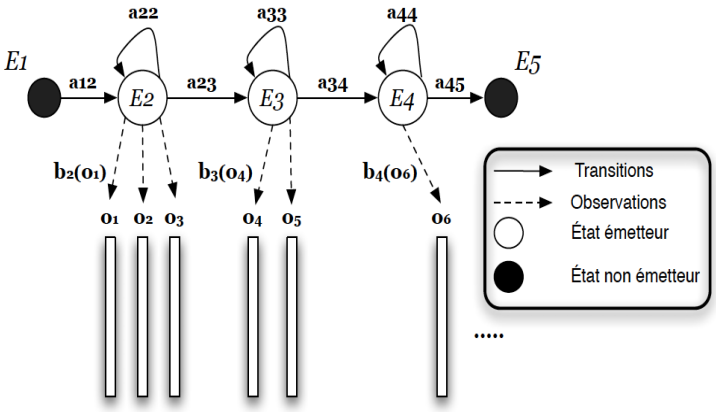


FIGURE 1.3 – Exemple d'un MMC à 5 états, dont 3 émetteurs (BOUGARES, 2012)

Plusieurs méthodes tels que les mélanges de modèles gaussiens et les réseaux de neurones profonds ont été proposés pour calculer les probabilités d'observations $b_i(o)$.

1.2.3.1 Les mélanges de modèles gaussiens (GMM/HMM)

Les modèles de mélange de gaussiennes (GMM) sont basés sur l'estimation d'une densité de probabilité suivant une loi Gaussienne. Dans un modèle GMM/HMM, chaque probabilité émise sur un état j est calculée comme une somme de poids de M_j Gaussiennes :

$$b_j(o_t) = P(o_t|j) = \sum_{m=1}^{M_j} \omega_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}), \sum_{m=1}^{M_j} \omega_{jm} = 1 \tag{1.8}$$

avec ω_{jm} le poids de la m ème composante du mélange pour l'état j à l'instant t et chaque Gaussienne ayant une densité de probabilité continue égale à :

$$N(o_t; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} exp(-\frac{1}{2}(o_t - \mu)^T \Sigma^{-1}(o_t - \mu)) \tag{1.9}$$

avec d est la dimension du vecteur acoustique $o_t \in R^d$.

1.2.3.2 Les réseaux de neurones profonds (DNN/HMM)

Les réseaux de neurones profonds (DNN - Deep Neural Network) ont été utilisés pour la modélisation acoustique depuis quelques années dans les travaux de (BOURLARD et WELLEKENS, 1989 ; MORGAN et BOURLARD, 1990) en utilisant les perceptrons multicouches (multilayer perceptron MLP). Mais, à cause de la puissance de calcul limitée à cette époque, ces modèles n'étaient pas très performants.

Les récents progrès techniques du matériel informatique ont permis de dépasser certaines limitations des premières approches neuronales. Les réseaux neuronaux peu profonds (MLP) ont été remplacés par les architectures neuronales profondes, avec de nombreuses couches cachées, voire des architectures neuronales plus complexes tels que les CNN (LECUN et al., 1990), LSTM, TDNN (PEDDINTI, POVEY et KHUDANPUR, 2015). Beaucoup de travaux ont montré que les modèles acoustiques DNN/HMM obtiennent de meilleures performances en comparaison avec les modèles acoustiques HMM/GMM dans de nombreuses tâches de reconnaissance de la parole (DAHL et al., 2011b ; DAHL et al., 2011a ; HINTON et al., 2012 ; YU, DENG et DAHL, 2010 ; LING, 2019). La figure 1.4 illustre l'architecture d'un modèle acoustique de type DNN/HMM. La couche de sortie du DNN utilise la fonction Softmax pour calculer la probabilité de chaque état j du HMM sachant l'observation o_t à l'instant t.

Pour une étude détaillée à propos des modèles acoustiques DNN/HMM, il est intéressant de se reporter à (LI et al., 2015).

1.2.4 Modèles de langage

Un modèle de langage (ML) représente un composant fondamental dans un système de reconnaissance de la parole. Il a pour objectif d'estimer la probabilité $P(W)$ utilisée dans la formule 1.3. Cette probabilité désigne la probabilité que la séquence de mots W appartienne à une langue donnée. Elle s'exprime par :

$$P(W) = P(w_1) \prod_{i=1}^{h} P(w_i|h) \tag{1.10}$$

où k est le nombre de mots dans la séquence W et h est l'historique du mot w_i.

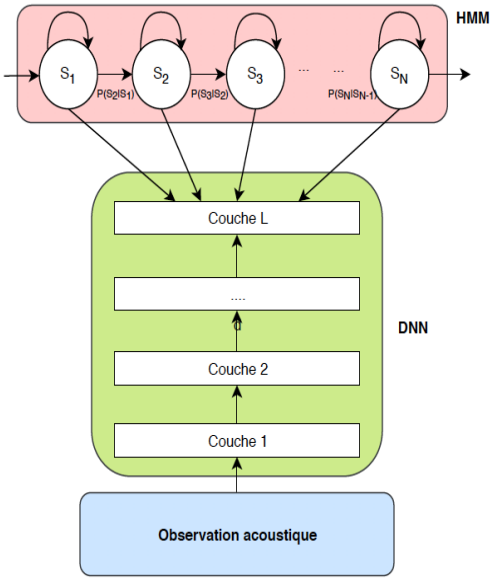


FIGURE 1.4 – Une architecture DNN/HMM pour la modélisation acoustique (ELLOUMI, 2019)

Les modèles de langages peuvent être des modèles n-grammes ou des modèles neuronaux. Ces modèles sont décrits en détails dans la section 1.3.

1.2.5 Dictionnaire de prononciation

Le dictionnaire de prononciation joue un rôle important dans le processus de reconnaissance automatique de la parole en faisant le lien entre la modélisation acoustique et la modélisation linguistique. Il détermine la concaténation des unités de modélisation acoustique (les phonèmes) pour construire les unités lexicales.

Un dictionnaire de prononciation fournit des représentations phonémiques pour chaque mot. Un mot peut avoir une ou plusieurs prononciations.

Les représentations phonétiques sont renseignées manuellement par des experts ou générées par un système de conversion graphèmes-phonèmes. Citons à titre d'exemple l'outil de conversion graphèmes-phonèmes LIA-PHON proposé par (BÉCHET, 2001) pour le français.

1.2.6 Décodage

Le principe du décodage consiste à trouver la séquence de mots qui maximise conjointement le produit des probabilités acoustiques et des probabilités linguistiques (formule 1.3). Les stratégies de décodage existantes peuvent être caractérisées par les aspects suivants :

1. La construction d'un espace de recherche à la volée :

L'espace de recherche est construit au même moment du décodage. Un exemple de système de reconnaissance de parole qui utilise ce type de décodage est CMU Sphinx.

2. L'espace de recherche est pré-construit :

   Un exemple de système de reconnaissance de parole qui utilise ce type de décodage est Kaldi.

1.2.7 Sorties des systèmes de reconnaissance de la parole

Grâce à la phase de décodage décrite précédemment (Section 1.2.6), il est possible de produire en sortie du système plusieurs représentations. La plus courante de ces représentations est la meilleure hypothèse, ou 1-best, qui est l'hypothèse la plus probable trouvée par le système.

Il est possible de retenir plusieurs hypothèses de reconnaissance. Ces hypothèses peuvent être fournies sous la forme de :

- **Liste des N-meilleures hypothèses ou N-best** : cette liste contient les N meilleures hypothèses trouvées pour chaque groupe de souffle [1] ordonnées selon le score calculé par le système de reconnaissance de la parole.
- **Graphes de mots** : une représentation alternative à la liste de N-best est le graphe de mots. Les graphes de mots sont des graphes orientés et acycliques dont chaque arête est étiquetée avec un mot et un score. Chaque noeud est étiqueté avec un point temporel (l'instant supposé où un mot se termine et un autre débute). L'avantage d'une telle représentation est qu'elle est compacte dont la recherche d'un chemin au sein du de laquelle est facile. Cette représentation peut être le point de départ pour d'autres analyses.
- **Réseaux de confusion** : un autre type de représentation de la sortie du système de reconnaissance de parole sont les réseaux de confusion (MANGU, BRILL et STOLCKE, 2000). Ils sont représentés aussi sous la forme d'un graphe sans contrainte temporelle forte (ordre topologique). Les mots localement identiques sont fusionnés. Les mots en concurrence sont regroupés sur des ensembles de confusions communs. Chaque mot obtient un score qui est sa probabilité a posteriori (obtenue à partir du treillis de mots) divisée par la somme des probabilités a posteriori des mots en concurrence avec lui (y compris l'absence de mot modélisé par ε). Dans un réseau de confusion, la recherche de la meilleure la meilleure solution est obtenue en minimisant les erreurs de transcription, et non en déterminant la suite de mots W possédant la plus grande probabilité $P(A|W)P(W)$. La figure 1.5 illustre un exemple de réseau de confusion.

1. Un groupe de souffle correspond à la parole prononcée par un locuteur entre deux respirations (pauses silencieuses).

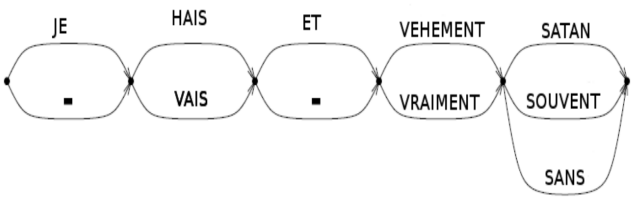


FIGURE 1.5 – Exemple de réseau de confusion (LECOUTEUX, 2008)

1.3 Modélisation linguistique

Comme nous l'avons vu précédemment, le modèle de langage a pour objectif d'estimer la probabilité $P(W)$ utilisée dans la formule 1.3. Ces modèles peuvent être des modèles de langage n-grammes ou des modèles de langage neuronaux.

1.3.1 Modèles de langage n-grammes

Un modèle de langage n-gramme estime la probabilité d'apparition d'un mot sachant les $n-1$ mots qui le précèdent. Généralement, les valeurs les plus utilisées pour n sont $n=3$ (modèles de langage trigrammes) et $n=4$ (modèles de langage quadrigrammes). Dans ce cas, l'équation 1.10 peut s'écrire :

$$P(W) = P(w_1) \prod_{i=2}^{n} P(w_i | w_{i-n+1} ... w_{i-1}) \tag{1.11}$$

avec $P(w_1)$ la probabilité d'observer le mot w_1 et $P(w_i|w_{i-n+1}...w_{i-1})$ la probabilité du mot w_i étant donné son historique $w_{i-n+1}...w_{i-1}$. La probabilité d'apparition d'un mot est généralement estimée par le critère de maximum de vraisemblance.

$$P(w_i|h) = \frac{C(h, w_i)}{C(h)} \tag{1.12}$$

avec $h = w_{i-n+1}...w_{i-1}$ et C représente le nombre d'occurrences d'une séquence de mots dans les données d'apprentissage.

1.3.2 Techniques de lissage

Un SRAP nécessite un corpus d'apprentissage volumineux afin que l'estimation des probabilités linguistiques soit précise. Mais, quelle que soit la taille du corpus d'apprentissage, les modèles n-grammes doivent modéliser des n-grammes non vus lors de l'apprentissage, pour lesquels le modèle doit attribuer une probabilité non nulle. L'utilisation de techniques de lissage

(CHEN et GOODMAN, 1999) a pour but de pallier le problème des événements non vus dans le corpus d'apprentissage. Les techniques les plus connues sont les techniques de décomptes de Good-Turing (GOOD, 1953), de Witten-Bell (WITTEN et BELL, 1991) et de Kneser-Ney (KNESER et NEY, 1995) qui utilisent toutes une stratégie de repli (back-off). Un état de l'art sur les différentes techniques de lissage pour la modélisation du langage est présenté dans (CHEN et GOODMAN, 1999 ; ZHAI et LAFFERTY, 2004 ; ZHAI et LAFFERTY, 2017)

1.3.3 Modèles de langage n-grammes à base de classes

Souvent, les mots ayant un sens ou une morphosyntaxe proche peuvent apparaître dans des contextes similaires. Ceci n'est pas pris en compte par les modèles de langage n-grammes classiques où les mots sont représentés dans un espace discret (le vocabulaire) dans lequel il n'existe aucun partage d'information syntaxique ou sémantique entre les mots.

Les modèles n-grammes à base de classes (BROWN et al., 1992) ont été introduits afin d'aborder ce problème en regroupant les mots et les contextes dans des classes en fonction de leurs utilisations. L'exploitation des informations relatives aux classes permet d'améliorer la généralisation des modèles de langage.

Ce modèle prédit non seulement un mot en fonction des $n-1$ classes le précédant, mais aussi une classe de mots en fonction des $n-1$ classes qui la précèdent. La probabilité d'une classe dans un modèle n-classes se calcule de la façon suivante :

$$P(c_i|h) = \frac{C(h, c_i)}{C(h)} \tag{1.13}$$

avec $h = c_{i-n+1}...c_{i-1}$ et C représente le nombre d'occurrences d'une séquence de classes dans les données d'apprentissage.

Par la suite, pour chaque mot et chacune de ses classes, la probabilité est calculée comme suit :

$$P(w|Classe(w)) = \frac{C(w)}{C(Classe(w))} \tag{1.14}$$

avec $Classe(w)$ est la fonction qui renvoie la classe du mot w.

La probabilité d'un mot au sein d'une séquence est alors estimée par la formule :

$$P(w|h) = P(w|Classe(w)) * P(c_w|h(c_w)) \tag{1.15}$$

avec $h = w_{i-n+1}...w_{i-1}$ et $h(c_w) = c_{i-n+1}...c_{i-1}$.

Cependant, ce type de modèle exige d'avoir un corpus d'apprentissage pré-étiqueté. L'étiquetage manuel est une tâche très coûteuse s'il est effectué manuellement et les résultats obtenus sont moins exacts s'il est effectué d'une façon automatique.

1.3.4 Modèles de langages neuronaux

Malgré les bonnes performances obtenues en utilisant les modèles n-grammes, ces modèles présentent l'inconvénient de la taille de l'historique qui est généralement limitée de 2 à 4 mots. Ces dernières décennies, l'apparition des réseaux de neurones dans la modélisation du langage a connu beaucoup de succès et a permis d'atteindre de performances très intéressantes grâce à leurs capacités à mieux gérer le problème des n-grammes non existants dans le corpus d'apprentissage. Ces modèles de langages neuronaux seront décrits dans les sections suivantes.

1.3.4.1 Modèles de langage feedforward

Les modèles de langage "*feedforward*" (BENGIO et al., 2003 ; SCHWENK et GAUVAIN, 2004 ; SCHWENK, 2007) ont été très utilisés dans la modélisation linguistique.

Les entrées du modèle de langage "*feedforward*" sont les indices des mots précédents qui définissent le contexte n-gramme. La sortie correspond aux probabilités de tous les mots dans le vocabulaire étant donné le contexte n-gramme. L'architecture du réseau est illustrée dans la figure 1.6. Un modèle de langage "*feedforward*" est composé d'au minimum trois couches, à savoir une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie.

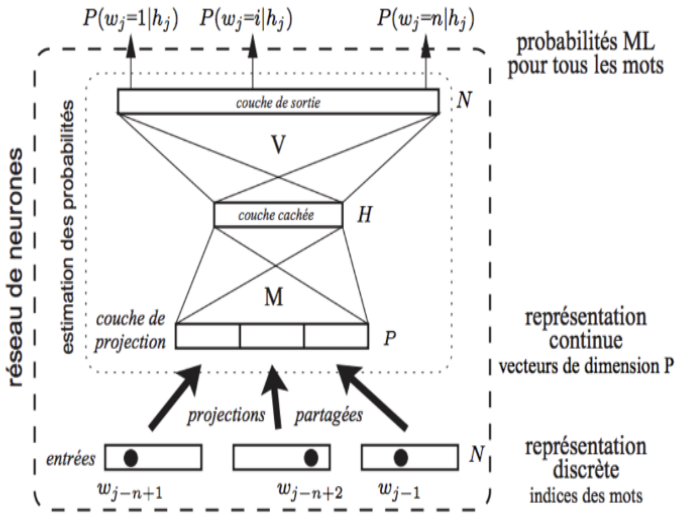


FIGURE 1.6 – Architecture d'un modèle de réseau neuronal "*feedforward*" quadrigramme (SCHWENK et GAUVAIN, 2004)

- **La couche de projection :** chaque mot de l'historique est projeté dans un vecteur de représentation continue de dimension m en utilisant une matrice M de dimension $|V|*m$

avec V est la taille du vocabulaire. La concaténation de ces vecteurs forme la couche de projection.

- **La couche cachée :** la concaténation des vecteurs de représentation constitue l'entrée pour la première couche cachée du réseau. Si le réseau contient plus d'une couche cachée, il peut être vu comme un empilement de couches neuronales dont chaque couche est définie en fonction d'un vecteur d'entrée, d'une matrice de poids et d'un biais. La sortie de ces couches se calcule selon la formule suivante :

$$h_l = \sigma(W_l i_l + b_l) \tag{1.16}$$

avec l est le numéro de la couche, σ est une fonction d'activation non linéaire, W est une matrice de poids et b est le biais.

- **La couche de sortie :** la couche de sortie prend en entrée la sortie de la couche cachée. Elle est constituée par un nombre de neurones qui est égal à la taille du vocabulaire $|V|$. Le but de la couche de sortie est de calculer les probabilités pour chaque mot w dans le vocabulaire en fonction du contexte n-gramme donné par le réseau. La couche de sortie utilise la fonction d'activation softmax afin de garantir que la somme des probabilités soit égale à 1. Elle se calcule selon la formule suivante :

$$y_t = softmax(W h_t + b) \tag{1.17}$$

où W est une matrice de poids et b est le biais.

Avec ce type d'architectures, la taille du vocabulaire de sortie pose un problème en terme de temps de calcul. Des solutions ont été proposées concernant spécifiquement la couche de sortie afin de réduire le coût d'inférence. Une de ces solutions réside à réduire la taille de la couche de sortie en ne considérant que les mots les plus fréquents dans le corpus d'apprentissage (*shortlist*) (SCHWENK, 2007).

1.3.4.2 Modèles de langage récurrents

Les modèles de langage *feedforward* permettent de résoudre le problème de sparsité des données en projetant les mots dans un espace continu de faible dimension grâce à une matrice de projection partagée. Cependant, ces modèles adoptent toujours une prédiction basée sur le contexte n-gramme où seulement les $n-1$ mots précédents sont pris en compte. Les dépendances contextuelles plus longues sont ignorées. Afin de pallier ce problème, les réseaux de neurones récurrents ont été proposés (MIKOLOV et al., 2010 ; MIKOLOV et al., 2011). L'architecture du modèle de langage récurrent est illustrée dans la figure 1.7. Cette architecture présente plusieurs différences par rapport aux modèles *feedforward*.

L'entrée du réseau $x(t)$ à l'instant t ne se compose que du mot précédent $w(t)$ au lieu des $n-1$ mots précédents. Cependant, un vecteur $s(t-1)$, qui représente les valeurs de sortie dans la couche cachée, à partir de l'étape précédente, s'ajoute. Ce vecteur permet de capturer des informations contextuelles d'une séquence via la connexion récurrente.

$$x(t) = w(t) + s(t-1) \tag{1.18}$$

avec b est le biais et f est une fonction d'activation.

Les vecteurs $w(t)$ et $s(t-1)$ sont concaténés dans un seul vecteur afin de former l'entrée de la couche cachée $s(t)$.

$$s(t) = f(Uw(t) + Ws(t-1) + b) \tag{1.19}$$

La couche de sortie $y(t)$ (équation 1.20) est constituée d'un nombre de neurones qui est égal à la taille du vocabulaire V ou à la taille de la *shortlist*.

$$y(t) = softmax(Vs_t + b) \tag{1.20}$$

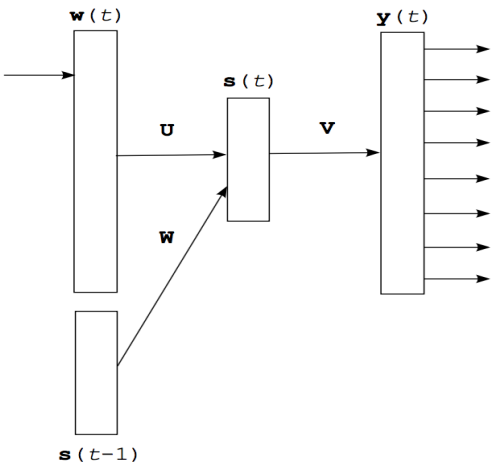


FIGURE 1.7 – Architecture d'un modèle du réseau neuronal récurrent (MIKOLOV, 2012)

1.3.4.3 Les modèles de langage "Long Short-Term Memory (LSTM)"

Quoique les RNN permettent, en théorie, de modéliser des dépendances infiniment longues, ils ne sont pas capables de mémoriser des historiques de grande taille.

Les réseaux de neurones *Long Short-Term Memory* (LSTM) sont des variantes des réseaux de neurones récurrents dont le but est d'apprendre les dépendances à long terme. Ces

réseaux intègrent différentes portes (HOCHREITER et SCHMIDHUBER, 1997) (en anglais *gate*), permettant d'écrire, de mettre à jour ou de lire une mémoire contextuelle, à partir d'informations vues précédemment. Ces portes permettent aux LSTM de modéliser plus efficacement les dépendances.

Les modèles de langage LSTM on été introduits par (SUNDERMEYER, SCHLÜTER et NEY, 2012). Un LSTM est composé d'une mémoire et de trois portes (figure 1.8). La porte d'oubli f (forget) contrôle quelle est la partie de la cellule précédente qui sera oubliée. La porte d'entrée i (input) doit choisir les informations pertinentes qui seront transmises à la mémoire. La sortie o (output) contrôle quelle partie de l'état de la cellule sera exposée en tant qu'état caché. En particulier, ces portes sont calculées comme suit :

$$i_t = \sigma(U_i x_t + W_i s_{t-1} + V_i c_{t-1} + b_i), \tag{1.21}$$

$$f_t = \sigma(U_f x_t + W_f s_{t-1} + V_f c_{t-1} + b_f), \tag{1.22}$$

$$g_t = f(U x_t + W s_{t-1} + V c_{t-1} + b), \tag{1.23}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{1.24}$$

$$o_t = \sigma(U_o x_t + W_o s_{t-1} + V_o c_{t-1} + b_o), \tag{1.25}$$

$$s_t = o_t . f(c_t), \tag{1.26}$$

$$y_t = g(V s_t + M x_t + d) \tag{1.27}$$

avec i_t, f_t, o_t sont respectivement la porte input, la porte forget et la porte output. c_t est la mémoire contextuelle. s_t est la sortie de la couche cachée. $U_i, U_f, U, U_o, W_i, W_f, W, W_o, V_i, V_f, V$ et V_o sont les matrices de poids. b_i, b_f, b, b_o et d sont les biais. f et σ sont les fonctions d'activation.

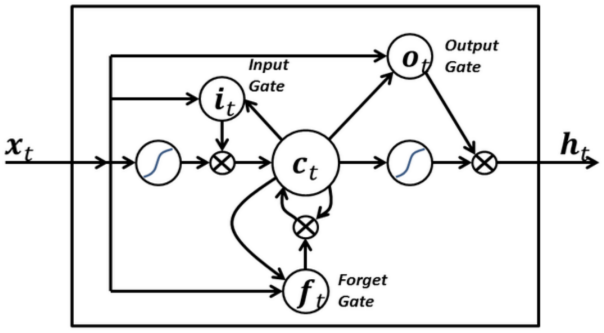


FIGURE 1.8 – Illustration des unités LSTM(s) (CHUNG et al., 2014)

1.3.4.4 Les modèles de langage "Gated Recurrent Unit (GRU)"

Les réseaux de neurones de type GRU *Gated Recurrent Unit* (GRU) (CHO et al., 2014) sont une extension des LSTM. Ils ont l'avantage d'être plus simples et moins coûteux en calcul car ils possèdent moins de paramètres. Ils sont composés de seulement deux types de portes (figure 1.9) au lieu de trois : une porte de ré-initialisation r (reset) qui détermine comment combiner la nouvelle entrée avec la mémoire précédente et une porte de modification u (update) qui permet de décider si l'état caché h doit être mis à jour avec le nouvel état caché h ou non.

$$r_t = \sigma(U_r x_t + W_r s_{t-1} + b_r), \tag{1.28}$$

$$z_t = \sigma(U_z x_t + W_z s_{t-1} + b_z), \tag{1.29}$$

$$\tilde{h}_t = f(U x_t + W(r_t \odot s_{t-1}) + b), \tag{1.30}$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{1.31}$$

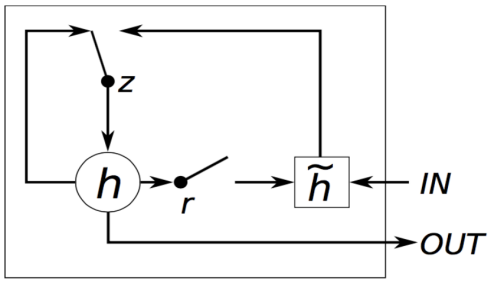


FIGURE 1.9 – Illustration des unités GRU (CHUNG et al., 2014)

Une autre variation du GRU est le GRU-Highway. Les réseaux highway (SRIVASTAVA, GREFF et SCHMIDHUBER, 2015) ont été proposé dont le but est d'optimiser les réseaux neuronaux et augmenter leur profondeur. Les réseaux highway (figure 1.10) servent à calculer une sortie qui est une combinaison entre l'entrée (x dans la figure 1.10) et la sortie d'un réseau neuronal ($F(x)$ dans la figure 1.10). Dans le cas du GRU-Highway, la sortie d'un réseau neuronal correspond à la sortie du GRU.

On a $h_t^{(gru)}$ qui représente la sortie du GRU classique, x_t est l'entrée à l'instant t et f est une fonction sigmoïde. Les équations pour le GRU-Highway sont les équations 1.28 - 1.30 auxquelles s'ajoutent les équations suivantes :

$$h_t^{(gru)} = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{1.32}$$

$$g_t = f(U_g x_t + W_g s_{t-1} + V c_{t-1} + b_g) \tag{1.33}$$

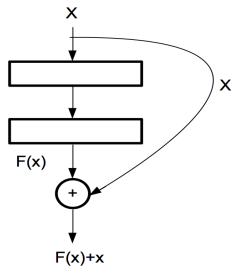


FIGURE 1.10 – Illustration de l'architecture des réseaux highway

$$h_t = g_t \odot h_t^{(gru)} + (1 - g_t)) \odot x_t \tag{1.34}$$

Sur le même principe que le GRU-Highway, on trouve aussi dans la littérature les réseaux LSTM-Highway (KURATA et al., 2017).

1.3.4.5 Apprentissage des modèles neuronaux

L'apprentissage des modèles consiste à ajuster l'ensemble des paramètres (matrices de poids, biais) en minimisant la fonction d'erreur ou de coût : l'entropie-croisée est utilisée avec l'algorithme de rétro-propagation (RUMELHART, HINTON et WILLIAMS, 1988).

L'algorithme commence par une propagation *avant* au cours de laquelle le réseau produit pour l'entrée la probabilité $P(w_i|w_{i-3}, w_{i-2}, w_{i-1})$. Puis, l'entropie croisée est calculée (erreur entre les valeurs de sortie estimées et les valeurs désirées) selon la formule 1.35.

$$E = -\sum_{k=1}^{V} t_k y_k \tag{1.35}$$

avec t_k représente la valeur de sortie estimée et y_k représente la valeur désirée.

Une fois qu'on dispose de la fonction d'erreur, nous devons maintenant apprendre nos paramètres afin de minimiser cette fonction d'erreur sur l'ensemble des exemples d'apprentissage. Pour ce faire, on effectue la propagation *arrière* qui consiste à rétro-propager la dérivée partielle de l'erreur $\frac{\partial E}{\partial W}$ par rapport aux poids du réseau. Et finalement, les poids sont mis à jour en fonction de cette dérivée partielle.

On retrouve généralement plus d'une technique de rétro-propagation. Les modèles de langage "feedforward" sont appris en utilisant l'algorithme standard de rétro-propagation (Back-Propagation-BP). Les réseaux neuronaux récurrents (RNN) sont habituellement entraînés par l'algorithme de rétro-propagation à travers le temps (Back-Propagation Through Time - BPTT). Une introduction détaillée de ces deux techniques se trouve dans l'étude de (MIKOLOV, 2012).

1.3.5 Évaluation du modèle de langage

Généralement, un modèle de langage est évalué en fonction de la performance du système dans lequel il est intégré. Par exemple, pour la reconnaissance de la parole, un modèle de langage est jugé meilleur qu'un autre si son utilisation permet un plus petit nombre d'erreurs sur les mots.

Cependant, un modèle de langage peut être également évalué indépendamment des systèmes dans lesquels il sera intégré. La mesure la plus couramment utilisée pour évaluer les modèles de langage est la perplexité (JELINEK, 1976). La perplexité (PPL) est définie par :

$$PPL = 2^{-\frac{1}{n}\sum_{t=1}^{n} log_2 P(w_t|h)} \tag{1.36}$$

où $P(w_t|h)$ est la probabilité donnée par le modèle n-gramme au mot w_t, h est l'historique du mot w_t et n est le nombre de mots dans le corpus.

Cette mesure d'évaluation trouve une justification avec la théorie de l'information et, intuitivement, elle peut être vue comme étant le nombre moyen de mots (parmi le lexique) équiprobables pour déterminer le prochain mot émis. Plus la perplexité est faible, plus le modèle peut être considéré comme modélisant correctement les données à traiter.

1.4 Évaluation d'un système de reconnaissance de la parole

L'évaluation des systèmes de reconnaissance de la parole a été l'objet de plusieurs campagnes d'évaluation pour la reconnaissance de la parole. La figure 1.11 montre l'historique de l'évaluation des systèmes de reconnaissance automatique de la parole sur des tâches de NIST jusqu'à l'année 2009 [2].

La mesure la plus répandue est le taux d'erreur sur les mots ou *Word Error Rate* en anglais (WER) (PALLETT, 2003). Cette mesure s'appuie sur une comparaison entre la phrase produite par le système de reconnaissance et la phrase correspondante transcrite manuellement. Un alignement mot à mot utilisant la distance de Levenshtein est réalisé entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Ensuite, une comparaison est effectuée selon les différents types d'erreurs sur les mots que peut commettre le système. Les types d'erreurs sont les suivants :

- Insertion (I) : nombre de mots insérés par le système ;
- Substitution (S) : nombre de mots substitués par le système ;
- Supression (D) : nombre de mots supprimés par le système.

Ces trois types d'erreurs sont distingués dans la figure 1.12. Le calcul du WER s'effectue

2. https ://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

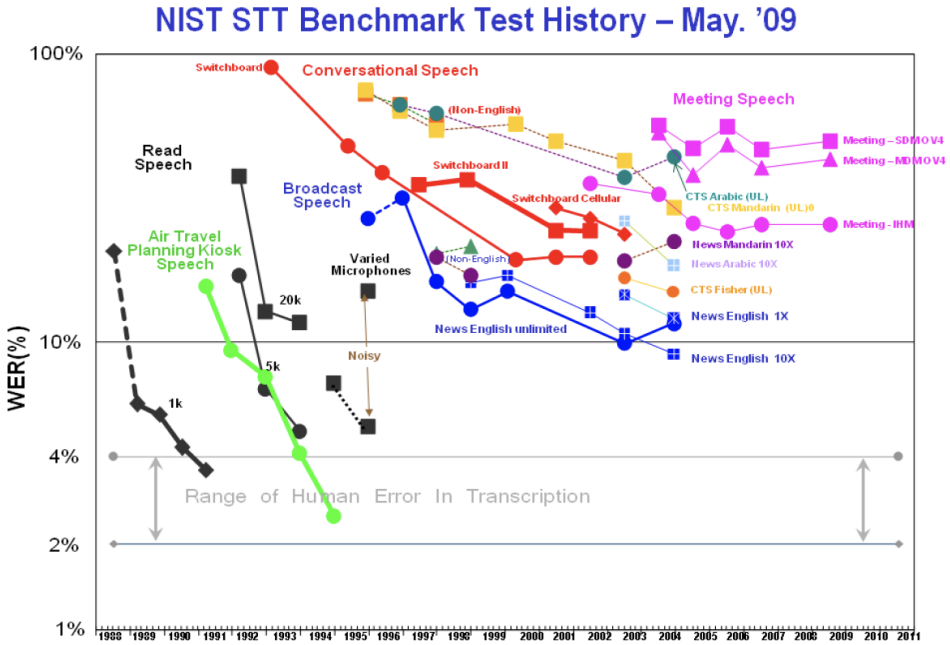


FIGURE 1.11 – Historique d'évaluation des systèmes de reconnaissance de la parole NIST 2009

selon la formule suivante :

$$WER = \frac{I + S + D}{N} \tag{1.37}$$

où N représente le nombre de mots dans la référence.

REF: les chaînes de caractères ** sont notées comme UNE suite de CARACTÈRES entourés de deux QUOTES.
HYP: les chaînes de caractères SE sont notées comme *** suite de CARACTÈRE entourés de deux CÔTÉS.

I D S S

FIGURE 1.12 – Alignement d'une transcription automatique (HYP) et d'une transcription de référence (REF)

1.5 Conclusion

Dans ce chapitre, nous avons décrit le principe de fonctionnement et les différents composants d'un système de reconnaissance automatique de la parole. Nous avons ensuite présenté la modélisation linguistique pour les systèmes de reconnaissance de parole.

L'apparition des architectures neuronales a amélioré les performances des SRAPs. Cependant, les systèmes de reconnaissance de la parole restent très sensibles à la variation de thèmes. Cette sensibilité est principalement due à la nature des données utilisées pour l'apprentissage qui sont inadéquates avec les données du test. L'adaptation des modèles de langage a pour but de réduire l'impact de cette inadéquation. Nous présentons, dans le chapitre suivant, un état de l'art sur les méthodes d'adaptation des modèles de langage pour un système de reconnaissance de la parole.

Chapitre 2

Adaptation des modèles de langage

Contents

2.1 Introduction

Même si ces dernières années la technologie de reconnaissance automatique de la parole a considérablement progressé, principalement grâce aux architectures neuronales pour la modélisation acoustique, un système de reconnaissance automatique de la parole reste sensible à la variation de sujets et à la précision de ses modèles de langage (ML). Un tel système doit être bien préparé pour traiter des documents spécialisés. L'adaptation de modèles de langage a été proposée pour résoudre ce problème d'inadéquation entre les données d'apprentissage et de test ou de production (dans un contexte de déploiement). Les enjeux de l'adaptation linguistique sont nombreux à savoir (1) le type de données d'adaptation à utiliser, (2) la manière de récupérer les données d'adaptation pour le sujet considéré, (3) les informations nécessaires

pour l'adaptation du modèle de langage, (4) la technique d'adaptation à adopter et, (5) les mots qui doivent être ajoutés au vocabulaire. Nous présentons dans ce chapitre les travaux de la littérature liés à ces différents enjeux.

Ce chapitre est organisé comme suit. Nous commençons par le principe et les techniques d'adaptation des modèles de langage n-grammes dans la partie 2.2. Nous présentons par la suite, dans la partie 2.3, l'adaptation des modèles de langage neuronaux.

2.2 Adaptation des modèles de langage n-grammes

Les systèmes de reconnaissance de la parole actuels sont sensibles à la variation de sujets. Cette sensibilité est principalement due à la nature des données utilisées pour l'apprentissage qui sont inadéquates avec les données du test. L'adaptation des modèles de langage a pour but de résoudre ce problème. Nous présentons dans cette section le principe d'adaptation linguistique des modèles de langage n-grammes ainsi que les techniques d'adaptation les plus utilisées dans l'état de l'art.

2.2.1 Principe de l'adaptation linguistique des modèles de langage n-grammes

Généralement, les modèles de langage sont spécifiques au domaine[1]. Ils sont estimés avec des données du même domaine auquel le modèle sera utilisé. En raison du coût élevé pour collecter des transcriptions manuelles de parole pour chaque domaine traité, plusieurs travaux se sont tournés vers l'utilisation de l'adaptation du modèle de langage des systèmes. L'adaptation d'un modèle de langage à un domaine consiste à réestimer ses probabilités n-grammes de manière à prendre en compte les spécificités linguistiques du nouveau domaine. Le schéma général du processus d'adaptation est présenté dans la figure 2.1. Avec l'adaptation du ML, on peut également enrichir le vocabulaire du système de reconnaissance avec des mots spécifiques au domaine afin d'éviter le problème des mots hors vocabulaire.

2.2.2 Nature des données d'adaptation

L'adaptation des modèles de langage nécessite une étape primordiale qui est l'acquisition des données d'adaptation que l'on nomme corpus d'adaptation ou corpus spécialisé. Les données d'adaptation peuvent être récupérées à partir de plusieurs sources selon trois types d'approches :

- **Utilisation d'un sous-ensemble du corpus d'apprentissage existant :**

1. On désigne par domaine le sujet (sport, politique, science...) que traite le discours à transcrire. De même, un sujet peut contenir plusieurs sous-sujets qui peuvent être considérés comme un domaine

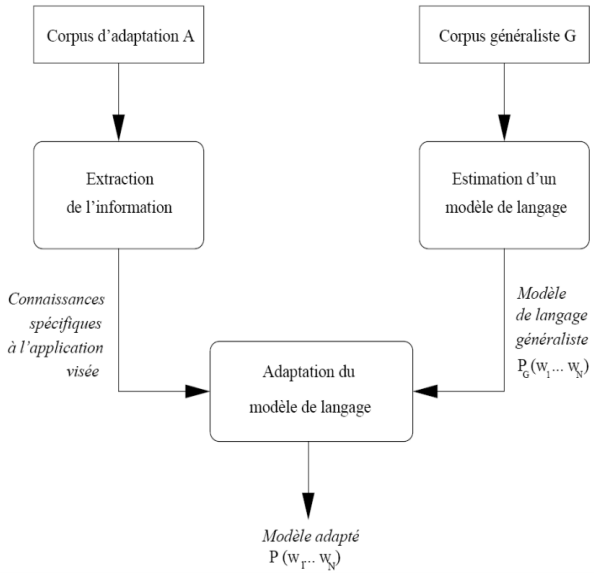


FIGURE 2.1 – Schéma général du processus d'adaptation d'un modèle de langage (ESTÈVE, 2002)

— **Recherche documentaire hors ligne :** L'idée est d'utiliser des archives de documents existants (articles de journaux, livres, etc) ou des corpus provenant des campagnes d'évaluation (ESTER [2], MGB [3], etc) afin de rechercher les parties qui correspondent à la tâche d'adaptation parmi ce corpus. La limite de ce type d'acquisition est qu'on n'est pas toujours sûr que ces corpus contiennent des informations pertinentes pour le nouveau thème et que la quantité de données est suffisante pour l'adaptation.

- **Génération automatique de données artificielles :**

— **Liste N-best ou N-meilleures hypothèses (SOUVIGNIER et al., 2000) :** Dans le cas où aucun corpus d'adaptation existe, il est possible de créer un corpus à partir du processus de reconnaissance de la parole. Il s'agit d'utiliser la liste de N-best hypothèses calculée sur le document à transcrire (Section 1.2.7) comme données d'adaptation. Ce modèle est appris avec l'hypothèse que les séquences de mots bien reconnues apparaissent plusieurs fois dans la liste de n-best hypothèses ce qui implique plus d'incidence de ces séquences par rapport aux séquences erronées.
Ce type des donnée a comme inconvénient qu'il n'apporte aucun nouveau mot par rapport au modèle initial issu du corpus d'apprentissage tout entier.

— **Grammaires génératrices :** Cette approche consiste à construire une grammaire

2. http ://www.afcp-parole.org/camp_eval_systemes_transcription/
3. http ://www.mgb-challenge.org/

pour un domaine particulier et à générer un corpus artificiel à partir de cette grammaire (RAUX et al., 2003 ; KELLNER, 1998). Une petite quantité de données d'apprentissage permet de pondérer les règles de grammaire et de générer ce corpus d'adaptation. Les MLs basés sur la grammaire ont l'inconvénient qu'ils ne seront pas aussi précis que des modèles construits à partir de données réelles car ils se basent sur des estimations artificielles. De plus, ils présentent le même problème que l'acquisition avec la liste N-best en ce qui concerne les nouveaux mots.

- **Recherche en ligne de données :**
  — **Recherche Web en ligne :** Suite à l'émergence de l'Internet, le web est devenu une source très riche de données textuelles. L'avantage de cette technique est que la collecte de ces données peut se faire pour un coût réduit, rapide, avec une quantité de textes généralement satisfaisante et accessible pour de nombreuses langues. Les données web ont été beaucoup utilisés pour l'acquisition de données d'adaptation de modèles de langage. Citons à titre d'exemple les travaux de (SARIKAYA, GRAVANO et GAO, 2005 ; LECORVÉ, GRAVIER et SÉBILLOT, 2008 ; OGER et al., 2008 ; OGER, POPESCU et LINARES, 2009 ; OGER, 2011 ; ABDULLAH, ILLINA et FOHR, 2018) où ils ont apporté des améliorations importantes dans les résultats en WER et en perplexité.
  Une technique souvent utilisée avec l'adaptation du ML à partir du web est la sélection de données. La méthode la plus classique est de conserver toutes les phrases exclusivement avec des mots du vocabulaire spécifique à la tâche définie au préalable. Une autre méthode proposée dans (MOORE et LEWIS, 2010 ; ROUSSEAU, 2013) est d'utiliser un algorithme de filtrage de texte basé sur l'entropie croisée.

2.2.3 Techniques d'adaptation

Les différentes techniques proposées pour l'adaptation des modèles n-grammes peuvent être classées en trois grandes catégories (BELLEGARDA, 2004) : (1) interpolation, (2) spécification de contraintes et (3) extraction de méta-informations.

1. **Interpolation :** L'interpolation nécessite au moins deux corpus : un corpus générique et un corpus d'adaptation. Elle consiste à combiner les probabilités au niveau des modèles ou au niveau des fréquences de mots.
   - **Interpolation linéaire :** l'interpolation linéaire (KLAKOW, 1998) des modèles de langage a été largement étudiée pour l'adaptation de modèles de langage en reconnaissance de la parole. Deux ou plusieurs modèles de langage individuels sont estimés à partir de corpus de différents domaines. Ces modèles de langage sont ensuite combinés comme indique l'équation 2.1. Des coefficients d'interpolation (λ_i) sont utilisés

et ajustés en minimisant la perplexité sur des données similaires au domaine cible. Ces coefficients indiquent l'utilité de chaque source pour une tâche particulière.

$$P(w|h) = \sum_i \lambda_i P_i(w|h) \tag{2.1}$$

avec

$$\sum_i \lambda_i = 1 \tag{2.2}$$

et h correspond à l'historique du mot w, λ est le coefficient d'interpolation et P_i est la probabilité du modèle n-gramme.

- **Interpolation dépendante du contexte :** L'interpolation dépendante du contexte (HSU, 2007) ou en anglais *generalized linear interpolation* est une extension de l'interpolation linéaire. Elle est définie comme suit :

$$P_{GLI}(w|h) = \sum_i \lambda_i(h) P_i(w|h) \tag{2.3}$$

avec $\lambda_i(h)$ représente les poids d'interpolation($\sum_i \lambda_i(h) = 1$). Ces poids sont dépendants de l'historique h.

- **Modèles dynamiques à mémoire cache :** lorsqu'un mot apparaît pour la première fois dans un document, il est beaucoup plus susceptible de réapparaître. Les modèles cache (KUHN et DE MORI, 1990) exploitent cette hypothèse pour améliorer les modèles de langage n-grammes en capturant des dépendances à longue distance dans les documents. Plus précisément, ces modèles ont un composant appelé "mémoire cache", qui contient les mots apparaissant dans l'historique (une fenêtre de taille N). Les modèles à mémoire cache sont généralement utilisés pour l'adaptation en les combinant avec des modèles à base de classes, sous la forme suivante :

$$P(w_i|h_i) = \sum_{\{c_i\}} P_{cache}(w_i|c_i) P(c_i|h_i) \tag{2.4}$$

où $\{c_i\}$ représente l'ensemble des classes pouvant être associées au mot w_i. La composante n-gramme de classes $P(c_i|h_i)$ est indépendante de la tâche, alors que la composante d'affectation de classe $P(w_i|c_i)$ est soumise à une adaptation dynamique par mémoire cache : le calcul de cette composante prend en compte les informations contenues dans cette mémoire. On a alors :

$$P(w_i|c_i) = (1-\lambda)P(w_i|c_i) + \lambda P(w_i|c_i) \tag{2.5}$$

avec λ est le coefficient d'interpolation.

Ces modèles à mémoire cache ont l'avantage de la simplicité de leur mise en oeuvre et apportent souvent des améliorations. Néanmoins, les erreurs sont propagées dans le système : si le système décode de manière incorrecte un mot, ce mot est placé dans la mémoire cache. Ceci nuit à la reconnaissance ultérieure en augmentant les chances de la répétition de la même erreur (GOODMAN, 2001).

- **Adaptation par Maximum a posteriori (MAP) :** les techniques d'adaptation présentées précédemment permettent de combiner les informations au niveau des modèles. L'adaptation par Maximum a posteriori (BACCHIANI et al., 2006) consiste à combiner les informations au niveau des fréquences de mots au lieu de les combiner au niveau des modèles.

$$P^{MAP}(w_i|h_i) = \begin{cases} \frac{\varepsilon C_A(h_i w_i)+C_G(h_i w_i)}{\varepsilon C_A(h_i)+C_G(h_i)} & si C_A(h_i w_i) + C_G(h_i w_i) = 0 \\ 0 & sinon \end{cases} \quad (2.6)$$

où $C_A(h_i w_i)$ et $C_G(h_i w_i)$ sont les nombres d'occurrences de la séquence de mots $h_i w_i$ dans, respectivement, le corpus d'adaptation A et le corpus générique G et ε représente un facteur constant estimé empiriquement afin d'optimiser l'influence du corpus d'adaptation.

2. **Adaptation par spécification de contraintes :** dans les approches d'adaptation basées sur la spécification de contraintes, le corpus d'adaptation est utilisé pour extraire les caractéristiques que le modèle de langage adapté doit satisfaire. En d'autres termes, au lieu de combiner des modèles de langage génériques et statiques, l'idée est de réestimer dans le modèle générique seulement les séquences de mots représentatives du thème.
   - **Adaptation par minimum d'information discriminante (MDI) :**
     L'intérêt de l'adaptation MDI (KULLBACK, 1997) est de pouvoir contraindre le modèle adapté à respecter des caractéristiques statistiques explicitement définies pour la tâche d'adaptation, ces caractéristiques étant généralement observées à partir d'un corpus d'adaptation. Pour ce faire, il s'agit de former un modèle adapté en minimisant la divergence de Kullback-Leibler entre le modèle générique et le modèle adapté

   L'inconvénient de ces modèles vient du fait qu'ils sont basés sur des calculs exponentiels et qu'ils nécessitent un temps de calcul important pour leur estimation (ZITOUNI, 2000).
3. **Adaptation par exploitation de méta-informations :** dans ce type d'adaptation, le corpus d'adaptation est utilisé pour extraire des méta-informations. Ces méta-informations peuvent être de types syntaxique, sémantique, thématique, etc.
   - **Triggers :** les modèles de langage basés sur des triggers (LAU, ROSENFELD et ROUKOS, 1993 ; SINGH-MILLER et COLLINS, 2007), appelés aussi modèles amorces, proposent de généraliser la modélisation linguistique en intégrant la distance entre des

paires de mots lors de l'apprentissage du modèle. Supposons que les données d'apprentissage révèlent une corrélation significative entre la paire de mots «chien» et «chat». La présence de «chien» dans le document pourrait alors automatiquement déclencher «chat», ce qui entraîne une modification de l'estimation de la probabilité. En pratique, les paires de mots avec une dépendance sont identifiées en recherchant les co-occurrences significatives de mots dans une fenêtre de taille fixe N. L'intégration de ces relations avec un modèle statique se fait alors par les schémas classiques de l'interpolation linéaire.

- **Adaptation à base de thèmes :** d'autres travaux ont cherché à élargir le concept de triggers en intégrant un mécanisme plus performant pour gérer la sélection de paires de mots. L'adaptation à base de thèmes s'exprime de la manière suivante :

$$P(w|h) = \sum_{t} P(w|t)P(t|h) \tag{2.7}$$

avec t est une variable latente qui fait référence à tous les thèmes, $P(w|t)$ est la probabilité d'un mot étant donné le thème et $P(t|h)$ est la probabilité d'un thème étant donné un historique.

L'intégration avec un modèle statique n-gramme se fait aussi par les schémas classiques de l'interpolation linéaire.

Plusieurs modèles à base de thèmes peuvent être utilisés à savoir les Bi-grammes thématiques (BIGI, DE MORI et THIERRY, 2000), LSA (BELLEGARDA, 2000), pLSA (HOFMANN, 1999), LDA (BLEI, NG et JORDAN, 2003), DM (SADAMITSU, MISHINA et YAMAMOTO, 2007).

- **Modèles de langage structurés :** les modèles de langage structurés ont été introduits dans le travail de (CHELBA et JELINEK, 2000). L'idée est de prendre en compte la structure syntaxique du langage parlé dans le calcul des probabilités conditionnelles. Ces modèles attribuent une probabilité $P(W,T)$ où W est une séquence de mots et T est l'analyse syntaxique de W représentée sous la forme d'un arbre binaire. Les symboles terminaux de l'arbre sont les mots de W accompagnés de leur classe syntaxique, alors que les symboles non terminaux correspondent au mot principal du syntagme (figure 2.2). Les syntagmes sont des sous-séquences de mots liés par des contraintes syntaxiques. Dans les travaux de (CHELBA et JELINEK, 2000 ; CHELBA et al., 1997), les auteurs définissent alors le calcul d'une probabilité de ce modèle comme :

$$P(w_i|\overline{h}_i) = \frac{1}{Z(\overline{h}_i)} \sum_{\{\pi_i\}} P(w_i|\overline{h}_i, \pi_i)P(\overline{h}_i, \pi_i) \tag{2.8}$$

où $\overline{h}_i$ représente l'historique qui commence à partir du premier mot w_1 de la phrase jusqu'au mot actuel w_i, $\{\pi_i\}$ est l'ensemble des analyses grammaticales partielles possibles de w_1 à w_i, $Z(\overline{h}_i)$ est un facteur de normalisation pour que la valeur de la sommation soit égale à 1.
En pratique, afin de simplifier la modélisation de ce type de modèle, l'ensemble π_i est remplacé par les racines de l'arbre syntaxique p_i. Alors, le calcul des probabilités peut se simplifier en appliquant deux réductions markoviennes en prenant en compte que les $n-1$ racines précédentes. Le modèle qui en résulte est de la forme :

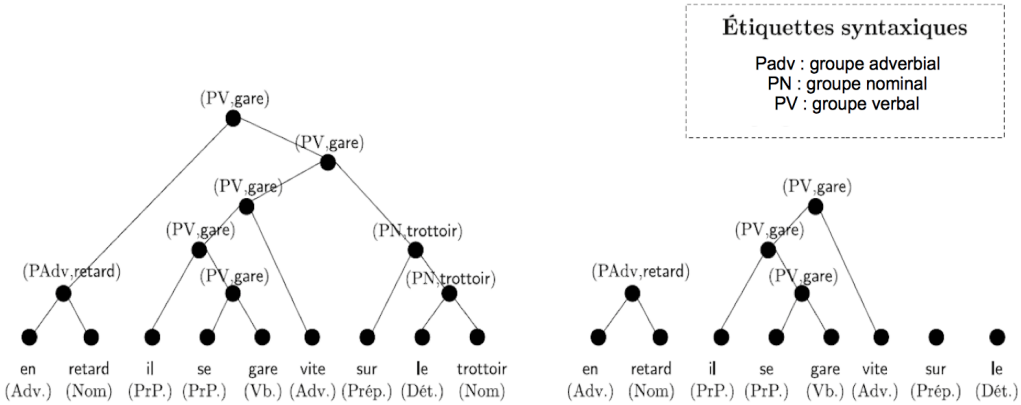
$$P(w_i|\overline{h}_i, \pi_i) = P(w_i|h_i, p_i) \tag{2.9}$$

Avec $h_k = w_{i-n+1}...w_{i-1}$
L'adaptation avec des modèles structurés considère que les données générique et les données du domaine dans une tâche spécifique sont soumis aux mêmes contraintes syntaxiques. Alors, l'idée d'adaptation avec des données structurées consiste à utiliser le corpus généraliste pour extraire les structures syntaxiques. Les informations issues du corpus d'adaptation sont alors combinées à celles portées par ces structures.
Les modèles de langage structurés présentent certaines limites : (1) la complexité du calcul, (2) l'ambiguïté vu qu'il existe plusieurs représentations syntaxiques possibles, (3) la nécessité d'une grande intervention manuelle par des linguistes experts en particulier lorsque la technique doit être appliquée à d'autres langues (MIKOLOV, 2012).

2.2.4 Adaptation du vocabulaire

L'adaptation du modèle de langage permet de ré-estimer les probabilités du modèle de langage seulement pour les mots qui existent dans le vocabulaire (Section 1.2.5). Ceci ne permet pas de résoudre les erreurs de transcription dues à l'absence de mots du vocabulaire. Un autre aspect de l'adaptation thématique consiste à ajouter des mots qui ne sont pas présents dans le vocabulaire généraliste appelés mots hors vocabulaire (*Out-Of-Vocabulary* en anglais (OOV)) du système. Parfois, il est très important d'intégrer ces mots dans le système de reconnaissance de la parole car ce sont les mots qui représentent le nouveau thème abordé. L'adaptation du vocabulaire suit en général les étapes suivantes :

1. La recherche de mots OOV pertinents : ces mots peuvent être recherchés selon plusieurs critères à savoir : le critère phonétique (PALMER et OSTENDORF, 2005), le critère morphologique (MARTINS, TEXEIRA et NETO, 2006), le critère thématique (LECORVÉ, 2010), le critère temporel (AUZANNE et al., 2000 ; FEDERICO et BERTOLDI, 2004), etc.



(a) Arbre syntaxique binaire complet

(b) Analyse syntaxique partielle

FIGURE 2.2 – Représentation sous forme d'arbres binaires d'une analyse syntaxique complète (a) et partielle (b) de la séquence de mots "en retard il se gare vite sur le trottoir" (LECORVÉ, 2010)

2. La phonétisation des mots OOV : les nouveaux mots doivent être phonétisés (section 1.2.5)

3. L'intégration des mots OOV : les mots OOV sélectionnés sont ensuite ajoutés au dictionnaire d'origine.

4. La rééstimation du modèle de langage.

2.3 Adaptation des modèles de langage neuronaux

Pour une étude détaillée des différentes techniques d'adaptation des modèles de langage neuronaux, il est intéressant de se reporter à (DEENA et al., 2019) et (HENTSCHEL et al., 2019b). Brièvement, ces techniques d'adaptation sont décrites ci-dessous :

2.3.1 Adaptation fondée sur les modèles "*Model-based adaptation*"

Ce type d'adaptation consiste à ajuster les paramètres d'un modèle de langage estimé sur de larges quantités de données à un domaine spécifique. On distingue deux types d'adaptation fondée sur les modèles. Ces deux types sont présentés ci-dessous.

2.3.1.1 Réglage fin "*Fine-tuning*" des modèles

Le *fine-tuning* des modèles est un moyen d'adapter un modèle neuronal à un domaine spécifique. Il implique un apprentissage supplémentaire du modèle de langage avec des données spécifiques à un domaine, ce qui donne un modèle spécifique au domaine.

2.3.1.2 Adaptation par couche linéaire cachée (*Linear Hidden Layer*(LHN))

Une couche d'adaptation peut être mise en cascade dans un modèle neuronal appris sur des données génériques, puis paramétrée en ne mettant à jour que les poids reliant la couche d'adaptation et la couche suivante en utilisant des données du domaine. Ce type d'adaptation a été utilisé pour les modèles de langage *feedforward* dont la couche d'adaptation a été mise entre la couche de projection et la couche cachée (PARK et al., 2010). Cette technique a été effectuée aussi pour les modèles récurrents (DEENA et al., 2016) où la couche d'adaptation s'ajoute entre la couche cachée et la couche de sortie. La figure 2.3 illustre l'adaptation par couche linéaire cachée pour les modèles de langage *feedforward* et pour les modèles récurrents.

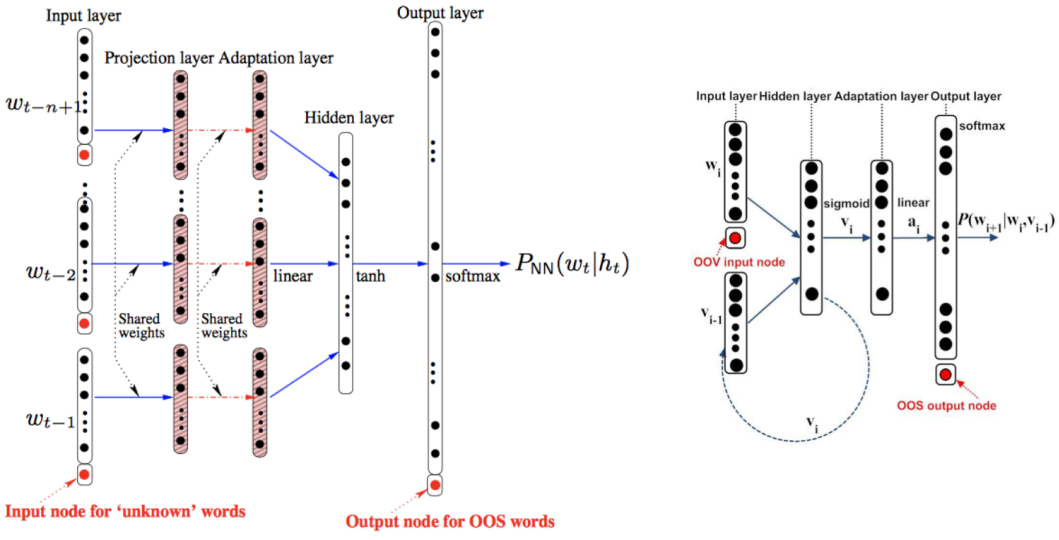


FIGURE 2.3 – Adaptation car couche linéaire cachée pour les modèles de langage *feedforward* (PARK et al., 2010) et pour les modèles de langage récurrents (DEENA et al., 2016)

2.3.2 Adaptation fondée sur des caractéristiques auxiliaires "*Feature-based adaptation*"

Feature-based adaptation a été utilisée pour adapter les modèles de langage neuronaux (MIKOLOV et ZWEIG, 2012 ; ARANSA, SCHWENK et BARRAULT, 2015). Contrairement à l'adaptation fondée sur les modèles qui nécessite de réapprendre le modèle générique avec des données du domaine, l'adaptation fondée sur des caractéristiques auxiliaires consiste à apprendre un nouveau modèle qui est conçu dés le départ pour tenir en compte des informations du domaine. Il s'agit d'ajouter un vecteur auxiliaire de contexte f qui représente l'information du domaine avec les entrées du réseau comme illustre la figure 2.4. Ces vecteurs peuvent être des :

- Vecteurs "one-hot" : les domaines sont représentés par un vecteur one-hot dont la dimension du vecteur est égale au nombre de thèmes présents dans les données d'apprentissage. Quelques corpus comme celui de la compagne MGB fournissent l'information de domaine (comédie, documentaire, drame, etc) avec les données (CHEN et al., 2015 ; ARANSA, SCHWENK et BARRAULT, 2015).
- Représentation dans un espace discret : le vecteur auxiliaire peut correspondre à un vecteur qui représente le thème tel que des vecteurs LDA (MIKOLOV et ZWEIG, 2012 ; SOUTNER et MÜLLER, 2013). L'adaptation d'un modèle de langage avec des vecteurs LDA nécessite l'apprentissage d'un modèle LDA (BLEI, NG et JORDAN, 2003) qui sera appris indépendamment du modèle de langage. Ceci nécessite un prétraitement du texte et une segmentation des données d'apprentissage en documents. Cependant, la segmentation est une information qui est difficile à avoir dans un corpus. (HENTSCHEL et al., 2019a) ont proposé d'utiliser des vecteurs de contexte "sequence summary network" (VESELỲ et al., 2016) et qui sont appris au même moment que l'apprentissage du modèle de langage.

2.4 Conclusion

Ce chapitre décrit un état de l'art sur les techniques d'adaptation des modèles de langage. La première partie présente les travaux proposés pour la récupération de données d'adaptation, l'adaptation des modèles de langage n-grammes et l'adaptation du vocabulaire. La deuxième partie a été consacrée à l'adaptation des modèles de langage neuronaux. L'introduction des modèles de langage neuronaux a amélioré les performances de SRAP. Les modèles de langage neuronaux peuvent être combinés avec des modèles de langage n-grammes classiques tels que dans les travaux de (CHEN et al., 2015 ; XIONG et al., 2016). Bien que ces diverses études rapportent dans l'ensemble des améliorations des taux de reconnaissance et

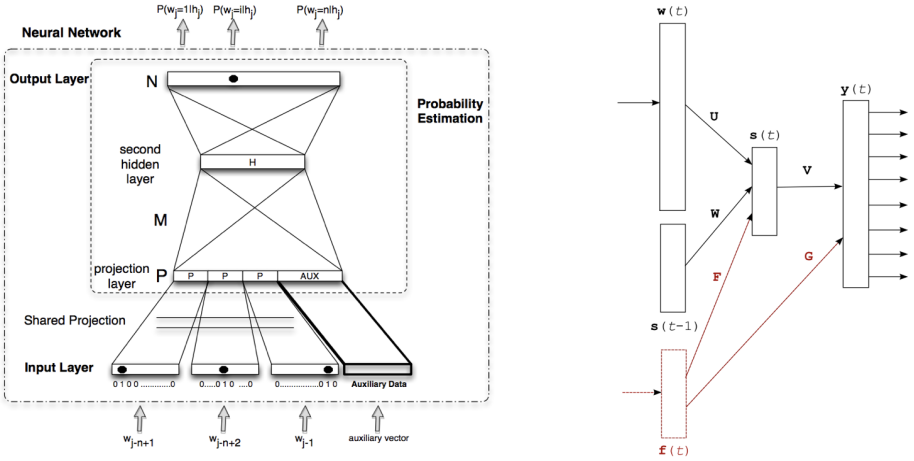


FIGURE 2.4 – Adaptation par contexte pour les modèles de langage *feedforward* (ARANSA, SCHWENK et BARRAULT, 2015) et les modèles de langage récurrents (MIKOLOV et ZWEIG, 2012)

des diminutions de la perplexité des modèles de langage adaptés, les systèmes de reconnaissance de la parole sont toujours confrontés à la difficulté d'ajout de mots hors vocabulaire. Dans le cadre de cette thèse, où il s'agit en premier temps de fournir la transcription des cours magistraux portant sur divers domaines, nous cherchons à proposer un schéma d'adaptation qui consiste à ré-estimer les probabilités du ML et à intégrer les mots relatives au domaine de chaque cours afin d'obtenir de meilleures transcriptions.

Chapitre 3

État de l'art : structuration automatique de la transcription

Contents

3.1 Introduction

Bien que la transcription soit largement utilisée dans de nombreux domaines tels que la transcription d'émissions télévisées, de réunions, etc., les utilisateurs ont souvent du mal à trouver des connaissances spécifiques au sein de la transcription en raison de leur nature non structurée (absence de paragraphes, de ponctuations, etc.). La structuration automatique telle que la segmentation thématique a été proposée dont le but est de structurer une transcription et d'offrir aux utilisateurs une meilleure visualisation.

L'intérêt de la structuration thématique ne se restreint pas que à des fins de visualisation. En effet, cette tâche constitue aussi le premier pas vers diverses applications telles que la na-

vigation, la recherche d'informations, le résumé automatique, etc. La segmentation thématique a pour objectif de diviser la transcription en des unités thématiques cohérentes.

Ce chapitre décrit l'état de l'art des méthodes de segmentation thématique les plus utilisées dans littérature. La section 3.2 présente quelques méthodes de segmentation automatique. La section 3.3 décrit les métriques d'évaluation les plus utilisées dans l'état de l'art de la segmentation thématique.

3.2 Les méthodes de segmentation automatique

L'analyse de la structure d'un texte comprend plusieurs plans d'organisation de l'information (FAUCONNIER et al., 2014 ; HERNANDEZ, 2004) : la structure logique (segmentation en titres, paragraphes, sections, chapitres, etc.), la structure visuelle marquée par la mise en forme et la mise en page (segmentation en pages, blocs visuels, etc.), la structure discursive (segmentation en unités élémentaires et complexes du discours) et la structure thématique (segmentation en unités thématiques cohérentes).

Dans le contexte d'étude de ce travail de thèse, les transcriptions automatiques peuvent servir pour la recherche automatique de matériaux pédagogiques et d'informations complémentaires. En travaillant à la structuration automatique du discours du tuteur, par exemple en découpant ce discours en segments thématiques, il est possible d'extraire de ces segments des mots-clés, ou de caractériser ces segments par d'autres moyens, de manière telle qu'il soit possible de lier un segment thématique à un ensemble de sources d'informations disponibles par ailleurs et non produites par le tuteur. Dans ce contexte, nous nous intéressons dans le cadre de ce travail à la structure thématique que nous considérons comme pertinente pour les transcriptions de cours magistraux.

Dans cette section, nous donnons un aperçu général des techniques de segmentation thématique les plus utilisées. Les techniques de segmentation de texte peuvent être groupées en deux différentes approches : (1) les méthodes de segmentation non-supervisée et (2) les méthodes de segmentation supervisée.

3.2.1 Les méthodes de segmentation thématique non supervisée

La majorité des méthodes de segmentation non supervisée s'appuient sur la cohésion lexicale. Nous définissions dans cette section la cohésion lexicale ainsi que quelques techniques de segmentation basées sur celle-ci.

3.2.1.1 Définition de la cohésion lexicale

La segmentation fondée sur la cohésion lexicale repose sur l'hypothèse que la répétition de mots similaires est un indice de cohérence locale dans un texte et que, de la même manière, un changement local de vocabulaire correspond à la présence de frontières entre des segments de texte. Cette segmentation est non-supervisée, elle ne nécessite donc pas de phase d'apprentissage et peut directement s'appliquer à tout document textuel.

(STOKES, CARTHY et SMEATON, 2004) illustrent les différents types de cohésion lexicale pouvant être présents dans le texte selon les catégories suivantes :

- **Répétition de mots :** il s'agit de la répétition de la même forme de surface d'un mot dans les parties ultérieures du texte.
- **Répétition par synonymie :** il s'agit de l'apparition des mots ayant la même signification mais présents sous des formes syntaxiques différentes.
- **Association de mots par spécialisation / généralisation :** il s'agit de l'utilisation d'une forme spécialisée d'un mot généralisé utilisé précédemment, par exemple les deux mots "animal" et "chien".
- **Association de mots par le biais de relations partielles et entières :** il s'agit de la présence dans le texte d'une relation partielle / entière entre deux mots, par exemple les deux mots "comité" et "membre".
- **Associations statistiques entre les mots :** ce type de relation se produit lorsque la nature de l'association entre deux mots ne peut pas être définie selon les types de relations exprimées ci-dessus par exemple "Obama" et "États-unis".

3.2.1.2 Les techniques de segmentation thématique fondées sur la cohésion lexicale

Dans cette section, nous donnons un aperçu général des algorithmes de segmentation les plus utilisés : l'algorithme TextTiling (HEARST, 1997), l'algorithme C99 (CHOI, 2000), l'algorithme U00 (UTIYAMA et ISAHARA, 2001), l'algorithme Dot Plotting (REYNAR, 1998), l'algorithme Topic Tiling (RIEDL et BIEMANN, 2012b), l'algorithme WLL (SITBON et BELLOT, 2007), et l'algorithme MinCut (MALIOUTOV, 2006).

— **L'algorithme TextTiling :**
TextTiling (HEARST, 1997) est l'un des premiers algorithmes de segmentation thématique. Cet algorithme consiste en 4 étapes principales : (1) le pré-traitement, (2) la tokenization, (3) le calcul des scores lexicaux et (4) l'identification de frontières thématiques. Le pré-traitement consiste à supprimer les mots vides non porteurs de sens (*stopwords* en anglais) et à lemmatiser (*i.e.* attribuer la forme canonique) les mots restants. La lemmatisation permet de réduire la taille du vocabulaire de sorte à faire apparaître la

répétition cachée des mots (due à la conjugaison, aux accords, etc.).
L'étape de tokenization consiste à segmenter le texte en phrases de longueur prédéfinie de N mots.
Par la suite, à l'aide d'une fenêtre glissante, la similarité entre des blocs adjacents est calculée tout au long du texte (figure 3.1). Les blocs sont constitués des pseudo-phrases obtenues lors de la phase de tokenization. Généralement, la similarité entre deux blocs est calculée en utilisant la mesure de cosinus. Étant donnés deux blocs b_1 et b_2 de taille k pseudo-phrases, la mesure de cosinus se calcule de la manière suivante :

$$sim(b_1, b_2) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}} \tag{3.1}$$

avec t représentant tous les mots dans un bloc et w_{t,b_i} le poids assigné au mot t dans le bloc b_i.
Une valeur de similarité proche de 1 signifie que les deux blocs sont similaires. Par contre, une valeur faible indique une faible similarité entre les blocs.

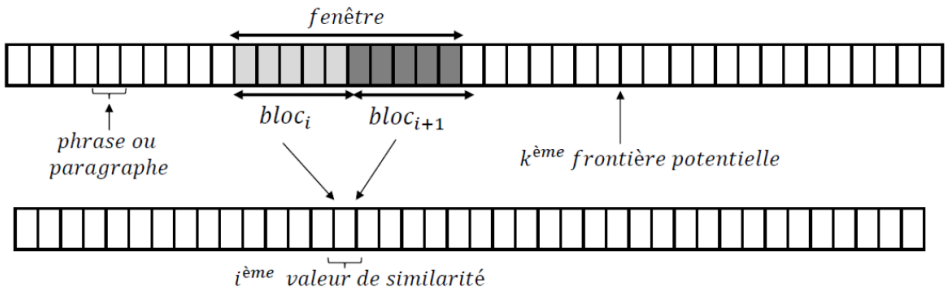


FIGURE 3.1 – Calcul de la cohésion lexicale avec le principe de la fenêtre glissante (figure extraite de (BOUCHEKIF, 2016))

Ces valeurs de similarité permettent de tracer une courbe de cohésion lexicale comme celle de la figure 3.2. Les minimums locaux dans cette courbe sont considérés comme des frontières potentielles.
L'étape d'identification de frontières consiste d'abord à déterminer les pics et les vallées à partir de la courbe de cohésion lexicale. Un pic correspond donc à deux blocs fortement liés thématiquement alors qu'une vallée correspond à une rupture de thèmes. Chaque vallée est donc considérée comme une rupture thématique candidate et correspond à une limite entre deux blocs thématiquement différents.
La profondeur de chaque vallée est calculée comme la somme des différences entre les pics à gauche et à droite par rapport à la vallée en question. Pour une vallée v_j, la

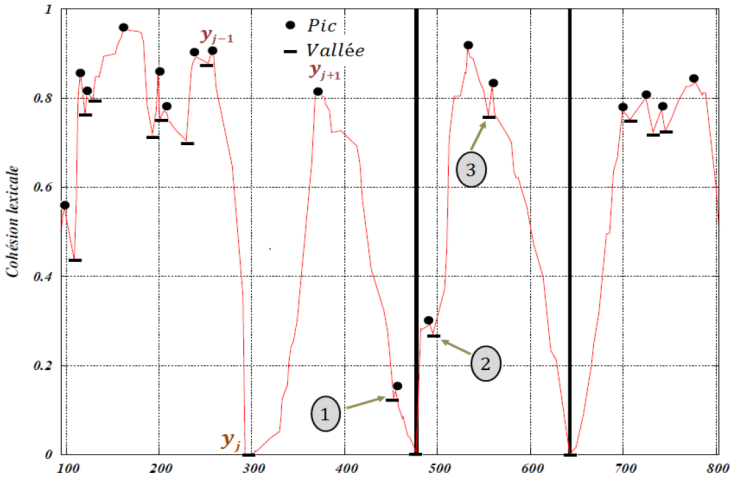


FIGURE 3.2 – Courbe de la cohésion lexicale (figure extraite de (BOUCHEKIF, 2016))

profondeur est donnée par :

$$depth(v_j) = (y_{j-1} - y_j) + (y_{j+1} - y_j) \tag{3.2}$$

avec y_c correspondant à la valeur de similarité à la position c.
En utilisant ces scores de profondeur, l'algorithme est en mesure de sélectionner un seuil permettant d'éliminer les petites vallées et de garder que les vallées qui dépassent le seuil comme segments thématiques.
(BOUCHEKIF, 2016) a adapté l'algorithme TextTiling aux spécificités des documents oraux. Pour cela, la similarité est calculée entre deux blocs de groupe de souffle. Les vallées sont ensuite détectées par un mécanisme récursif de détection de minimum local. Une deuxième passe de l'algorithme peut être faite, en considérant la cohésion sémantique (BOUCHEKIF et al., 2015) entre les fenêtres au lieu de la cohésion lexicale.

— **L'algorithme C99 :**
L'algorithme C99 (CHOI, 2000) s'appuie sur des calculs de similarité entre phrases et non pas entre blocs. À partir d'un texte pré-traité, l'algorithme construit une matrice de taille $n \times n$ (n représente le nombre de phrases dans le texte) qui contient les valeurs de similarité entre tous les couples de phrases.
Ensuite, au lieu d'utiliser directement la matrice de similarité telle quelle, l'algorithme construit une nouvelle matrice, appelée matrice de rang. La matrice de rang détermine un classement local de chaque case de la matrice de similarité vis-à-vis des cases voisines au sein d'un masque défini par l'utilisateur (figure 3.3). Le rang est calculé de

la manière suivante :

$$rang = \frac{Nombre\ d'éléments\ ayant\ une\ similarité\ inférieure\ dans\ le\ masque}{Nombre\ d'éléments\ réellement\ présents\ dans\ le\ masque} \quad (3.3)$$

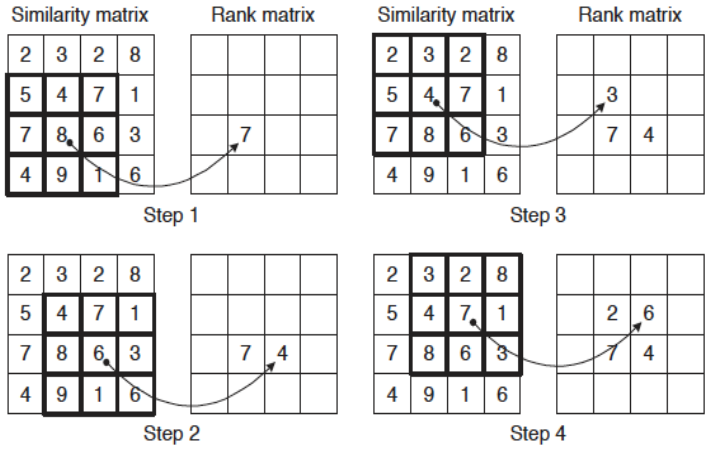


FIGURE 3.3 – Calcul de la matrice de rang à partir de la matrice de similarité (CHOI, 2000)

Les segments thématiques sont ensuite identifiés en utilisant un processus de regroupement (clustering) inspiré de l'algorithme de maximisation de Reynar (REYNAR, 1994).

— **L'algorithme U00** :
L'algorithme U00 (UTIYAMA et ISAHARA, 2001) propose d'utiliser une approche statistique fondée sur des modèles de Markov cachés pour trouver la segmentation la plus cohérente possible dans un document. La cohésion lexicale dans son approche est mesurée de manière classique à l'aide de la modélisation du langage. Un texte est représenté sous forme de graphe linéaire (figure 3.4). Chaque segment est défini par deux noeuds : un noeud de début et un noeud de fin.

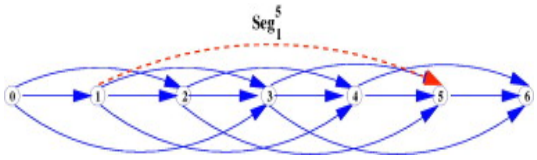


FIGURE 3.4 – Noeuds et segments dans l'algorithme U00 (figure extraite de (MISRA et al., 2009))

La probabilité d'une segmentation S est définie par la règle de Bayes :

$$P(S|W) = \frac{P(W|S)P(S)}{P(W)} \tag{3.4}$$

L'objectif est de trouver la meilleure segmentation parmi toutes les segmentations possibles. Notons que la probabilité $P(W)$ est constante pour le texte W. Alors il s'agit de maximiser la formule suivante :

$$\widehat{S} = argmax\ P(W|S)P(S) \tag{3.5}$$

— **Dot Plotting :**
L'algorithme Dot Plotting (REYNAR, 1998) s'appuie sur une représentation graphique du texte contenant les occurrences de mots du texte à segmenter. Lorsqu'un mot apparaît à deux positions du texte x et y, quatre points seront représentés sur le graphe : $(x, x), (x, y), (y, x)$ et (y, y). Ceci permet de visualiser les zones du texte où les répétitions sont nombreuses (figure 3.5).

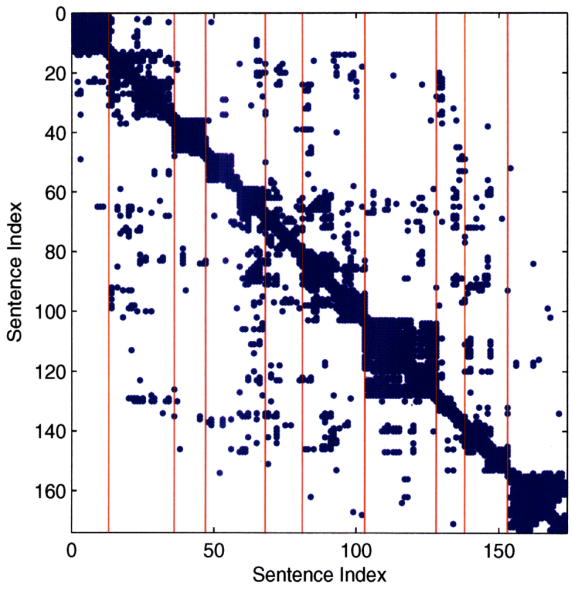


FIGURE 3.5 – Graphe de similarité pour un cours magistral (MALIOUTOV, 2006)

Les limites des segments thématiques correspondent aux positions de début et de fin des zones les plus denses du graphe (les lignes en rouge sur le graphe 3.5). La densité est calculée pour chaque unité d'aire en divisant le nombre de points contenus dans une région par l'aire de cette région. À partir des résultats de la densité, deux manières sont possibles pour déterminer les frontières thématiques. La première consiste à identifier

les limites en maximisant la densité au sein des segments. La deuxième consiste à repérer la configuration qui minimise la densité des zones entre les segments.
L'algorithme Dot Plotting a été utilisé pour la segmentation de cours magistraux par (MALIOUTOV, 2006).

— **TopicTiling :**
TopicTiling (RIEDL et BIEMANN, 2012b) est un algorithme qui repose sur les thèmes et non directement sur des mots ; un thème étant décrit par un ensemble de mots. Cela a l'avantage de réduire la dispersion de données puisque l'espace des mots est réduit à un espace de thèmes de dimension beaucoup plus basse. Ces thèmes sont déterminés avec des inférences LDA (Latent Dirichlet Allocation). Cependant, pour obtenir des inférences de thèmes pertinents, le modèle de thème doit être appris avec des documents dont le contenu est similaire à celui des documents de test (documents à segmenter). La figure 3.6 illustre l'architecture de l'algorithme TopicTiling et la figure 3.7 illustre un exemple de texte avec des inférences LDA.

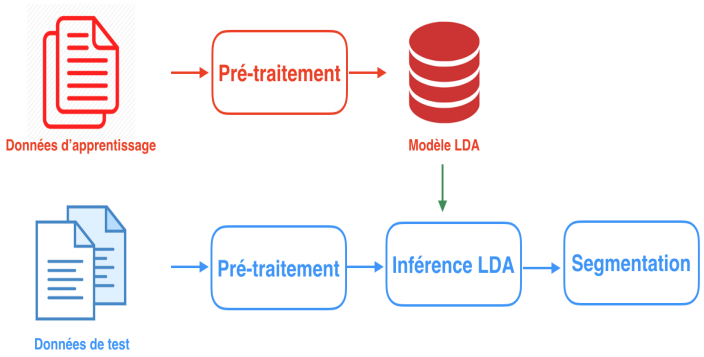


FIGURE 3.6 – Architecture de la segmentation de l'algorithme TopicTiling

Une fois les thèmes sont inférés à partir d'un corpus d'apprentissage et reconnus dans le texte à segmenter, des algorithmes comme TextTiling et C99 peuvent être appliqués pour la segmentation, en utilisant l'identifiant de thème de chaque mot au lieu du mot lui-même (RIEDL et BIEMANN, 2012a).

— **L'algorithme WLL (Weighted Lexical Links) :**
Dans l'algorithme WLL (SITBON et BELLOT, 2007), les frontières sont déterminées non seulement à partir de la répétition des mots mais aussi à partir de leurs positions. L'algorithme commence par l'extraction des chaînes lexicales. Une chaîne lexicale relie les mots ayant la même forme écrite. Une chaîne lexicale commence à la première occurrence et se termine à la dernière occurrence de ce mot. La chaîne sera divisée en sous-chaînes dans le cas où la distance entre deux occurrences consécutives dépasse

Mr:62 .:97 Pohs:2 ,:2 previously:4 executive:2 vice:2 president:2 and:17 chief:2 operating:2 officer:2 ,:72 was:2 named:2 interim:2 president:2 and:73 chief:2 executive:2 officer:2 after:17 David:2 M:27 .:36 Harrold:65 ,:2 a:84 company:2 founder:2 ,:26 resigned:2 from:91 the:34 posts:2 for:62 personal:61 reasons:2 in:84 August:2 .:58 Cellular:70 said:54 Robert:2 J:61 .:42 Lunday:2 Jr:18 .:31 ,:44 its:57 chairman:2 and:73 another:25 founder:2 ,:31 resigned:2 from:91 the:57 company:2 's:24 board:2 to:10 pursue:2 the:10 sale:55 of:67 his:28 telephone:31 company:42 ,:74 Big:10 Sandy:50 Telecommunications:31 Inc:2 .:74
APARTHEID:37 FOES:37 STAGED:41 a:37 massive:37 anti-government:37 rally:37 in:40 South:37 Africa:37 .:19 More:29 than:34 70:45 ,:26 000 **people:37** *filled:17 a:22 soccer:37 stadium:88 on:46 the:34 outskirts:37 of:93 the:24 black:37 township:37 of:45 Soweto:37 and:37 welcomed:11 freed:37 leaders:37 of:98 the:57 outlawed:37 African:37 National:45 Congress:87 .:72 It:79 was:55 considered:37 South:37 Africa:37 's:33 largest:90 opposition:67 rally:37 .:37*

FIGURE 3.7 – Extrait d'un document du corpus WSJ de Galley (GALLEY et al., 2003). Chaque mot est suivi de l'identifiant qui représente le thème (RIEDL et BIEMANN, 2012a)

un certain seuil. La figure 3.8 illustre le processus de construction des chaînes lexicales. Les lignes verticales indiquent les frontières de phrases. Plus les répétitions des mots sont proches, plus le mot est significatif pour le thème. Un poids, défini comme suit, sera calculé pour chaque chaîne lexicale :

$$w(chain, m) = cat_m \times nb_m \times log(\frac{L_{text}}{L_{chain}}) \tag{3.6}$$

où cat_m est un poids syntaxique (un poids est attribué à chaque lemme en considérant que certaines catégories grammaticales sont plus importantes que d'autres), et nb_m est le nombre d'occurrences du lemme m.

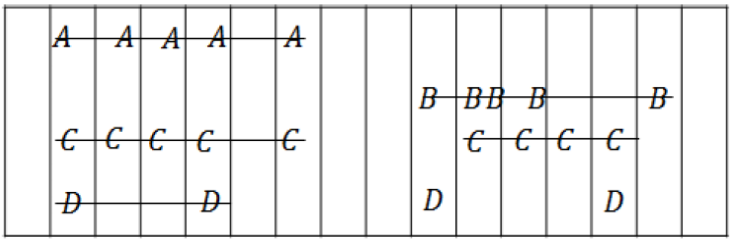


FIGURE 3.8 – Construction des chaînes lexicales pour les mots A, B, C et D dans un document (SITBON et BELLOT, 2007)

La similarité entre les phrases adjacentes est calculée en utilisant le principe général de TextTiling et tout au long du texte par le biais d'une fenêtre glissante de taille n. La

similarité est alors donnée par :

$$cosine(A,B) = \frac{\sum_i w(A,t_i) \times w(B,t_i)}{\sqrt{\sum_i w^2(A,t_i) \times \sum_i w^2(B,t_i)}} \tag{3.7}$$

où A et B sont les ensembles des vecteurs représentant les poids des chaînes lexicales présentes dans les n phrases précédentes et suivantes.

— **L'algorithme MinCut :**

Le principe de MinCut (MALIOUTOV, 2006) repose sur le principe des graphes. Soit G un graphe non-orienté $G = \{V, E\}$ où V est l'ensemble de sommets correspondant à des phrases dans le texte et E est l'ensemble des poids des arêtes. Le but est de trouver une coupe minimale. La coupe minimale est une partition du graphe en deux ensembles disjoints A et B de noeuds qui minimise le critère de la coupure, définie par $NCut(A,B)$. Elle correspond à la somme normalisée des coûts associés à chacun des arcs reliant les segments. $NCut(A,B)$ est définie par :

$$NCut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \tag{3.8}$$

où $cut(A,B) = \sum_{u \in A, v \in B} w(u,v)$ et $vol(A) = \sum_{u \in A, v \in V} w(u,v)$.

La figure 3.9 illustre deux exemples de partitionnement. Les valeurs de critère de coupure se calculent de la manière suivante.

Pour le graphe à gauche :

$cut(A,B) = 0.1, vol(A) = 1.7, vol(B) = 0.5$ et $NCut = 0.26$.

Pour le graphe à droite :

$cut(A,B) = 0.5, vol(A) = 2.1, vol(B) = 0.5$ et $NCut = 1.2$.

La partition à gauche prend la valeur la plus faible, c'est donc celle qui sera considérée comme étant la meilleure segmentation.

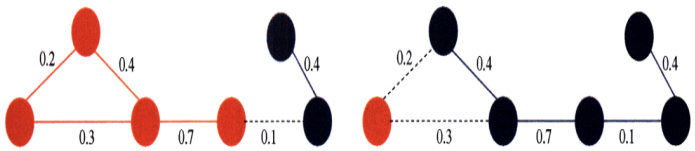


FIGURE 3.9 – Exemple d'une coupure d'un graphe binaire (MALIOUTOV, 2006)

Dans la segmentation de texte, les textes se composent généralement de plus de deux segments. Par conséquent, par extension, il faut s'intéresser non seulement aux coupes binaires mais aux coupes multi-voies sur les graphes. Pour un graphe (V) contenant

$A_1, .., A_k$ partitions, le découpage normalisé est donné par :

$$NCut_k(V) = \frac{cut(A_1, V - A_1)}{vol(A_1)} + ... + \frac{cut(A_k, V - A_k)}{vol(A_k)} \tag{3.9}$$

avec $V - A_k$ est la différence entre la $k^{ème}$ partition et le graphe entier.
Pour le découpage, il s'agit d'utiliser la programmation dynamique.

$$C[i,k] = \min_{j<k}[C[i-1,j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]}] \tag{3.10}$$

$$B[i,k] = \underset{j<k}{argmin}[C[i-1,j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]}] \tag{3.11}$$

où $C[i,k]$ est la valeur de coupure optimale pour les k premières phrases en i segments, $A_{j,k}$ est l'ensemble de noeuds commençant par le $j^{ème}$ noeud et se terminant par le $k^{ème}$ noeud et $B[i,k]$ est une table contenant la séquence optimale de la segmentation thématique.

Les algorithmes de segmentation présentés dans cette section ont été explorés dans le cadre de la segmentation de transcription automatique, que se soit en utilisant les algorithmes tels qu'ils sont ou en les adaptant aux spécificités des documents oraux (GUINAUDEAU et HIRSCHBERG, 2011 ; BOUCHEKIF, 2016). Ces travaux présentés visent à donner un poids à chaque mot selon son degré d'importance, en pénalisant les mots apparaissant tout au long de la transcription et en favorisant les mots importants dans chaque thème. Nous retenons, dans le cadre de ce travail de thèse, l'algorithme TextTiling et nous proposons dans le chapitre 8 d'intégrer l'information de changement de diapositives dans le calcul de la similarité.

3.2.2 Les méthodes de segmentation thématique supervisée

La segmentation supervisée peut-être vue comme une classification binaire. Il s'agit de prédire si une unité de texte contient, ou non, un changement thématique. Une unité de texte peut être un paragraphe, une phrase, un groupe de souffle, etc. La figure 3.10 schématise le principe de la segmentation supervisée.

Divers classifieurs tels que les arbres de décision (TÜR et al., 2001), les champs aléatoires conditionnels (Condional Random Fields - CRF) (WANG et al., 2012), les machines à vecteur de support (Support Vector Machines - SVM) (GEORGESCUL, CLARK et ARMSTRONG, 2006), etc. ont été testés pour la segmentation supervisée de texte.

Les caractéristiques les plus populaires sont notamment des caractéristiques linguistiques (TF-IDF, mots déclencheurs, etc. (JOTY et al., 2011)) ou des caractéristiques du signal audio et de la parole (indicateurs acoustiques, longueur des pauses, variation du débit de la parole,

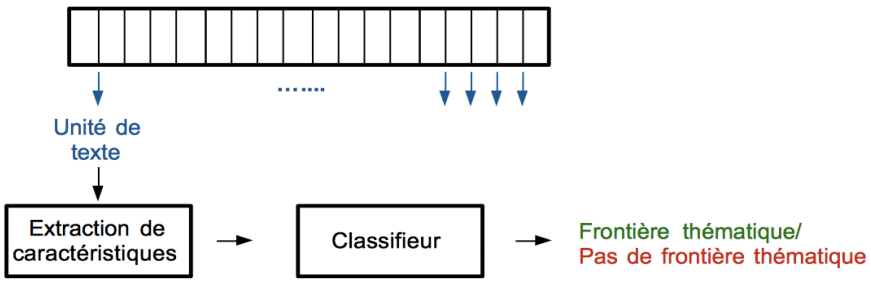


FIGURE 3.10 – Principe de la segmentation supervisée

silences, chevauchements, changement de locuteur, etc. (GALLEY et al., 2003 ; WANG et al., 2012)).

Plus récemment, les réseaux de neurones ont été introduits dans le domaine de segmentation thématique (WANG et al., 2017 ; BADJATIYA et al., 2018 ; KOSHOREK et al., 2018 ; ARNOLD et al., 2019). En particulier, on relève les travaux de (WANG et al., 2017) qui ont abordé la segmentation en formant un réseau de neurone convolutif (Convolutional Neural Network - CNN) pour apprendre les scores de cohérence entre des paires de textes. (BADJATIYA et al., 2018) ont introduit une architecture neuronale pour la segmentation s'appuyant sur un modèle d'attention LSTM bidirectionnel, dans lequel les représentations de phrases sont apprises à l'aide de CNN et les segments sont prédits en fonction des informations contextuelles.

L'inconvénient de ces méthodes de segmentation supervisées est qu'elles nécessitent la présence de gros corpus d'apprentissage dont l'annotation manuelle est coûteuse en temps et en ressources humaines qualifiées (et par conséquent aussi pécuniairement parlant).

3.3 Évaluation de la segmentation thématique

L'évaluation de la segmentation thématique est une tâche très difficile. Une première difficulté vient du fait que les juges humains ne sont pas toujours en accord sur la position exacte où les frontières doivent être placées. De ce fait, il est difficile de choisir une segmentation de référence pour la comparaison.

Afin d'évaluer la segmentation, il est nécessaire de disposer d'un corpus de textes pré-segmentés qui servira de référence, d'un corpus de textes segmentés automatiquement par notre système (hypothèse) et d'une mesure d'évaluation.

Nous présentons dans cette section quelques métriques d'évaluation utilisées pour l'évaluation de la segmentation.

3.3.1 Rappel et précision

En traitement automatique du langage naturel, la précision et le rappel sont des mesures standards pour évaluer les résultats et la performance des algorithmes utilisés. Ces mesures ont été essentiellement utilisées en classification et en recherche d'informations. Le rappel représente le nombre de documents pertinents qui sont correctement trouvés par rapport au nombre total de documents pertinents à trouver pour une requête. La précision représente le nombre de documents pertinents correctement trouvés par rapport au nombre de documents étiquetés comme pertinents par un système.

Dans le contexte de la segmentation thématique la précision et le rappel sont identifiés comme :

$$Précision = \frac{nombre\ de\ frontières\ correctes\ proposées\ par\ le\ système}{nombre\ total\ de\ frontières\ correctes} \tag{3.12}$$

$$Rappel = \frac{nombre\ de\ frontières\ correctes\ proposées\ par\ le\ système}{nombre\ total\ de\ frontières\ à\ trouver} \tag{3.13}$$

La moyenne harmonique du rappel et de la précision correspond à une mesure nommée F-mesure qui a pour but de donner une appréciation de la performance globale du système.

$$F\_mesure = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \tag{3.14}$$

Le problème de ces métriques dans la tâche de la segmentation automatique est que les frontières détectées par le système de segmentation doivent coïncider exactement avec celles du corpus de référence pour être comptabilisées comme juste. Une frontière prédite, même très proche de la frontière de référence, est tout simplement comptabilisée fausse. De ce fait, les valeurs de rappel, de précision et de F-mesure vont être très faibles. Afin de résoudre ce problème, une marge de tolérance entre les frontières d'hypothèse et de référence peut être utilisée. Par conséquent, si une frontière identifiée par l'algorithme se trouve très proche de la frontière de référence et elle respecte la marge de tolérance, elle sera comptabilisée comme juste.

3.3.2 Beeferman p_k

Pour résoudre le problème avec précision, rappel et F-mesure, la mesure p_k a été introduite (BEEFERMAN, BERGER et LAFFERTY, 1997). L'idée est de vérifier si deux phrases du corpus sont correctement identifiées comme appartenant au même segment ou n'y appartenant pas. Pour ce faire, il s'agit d'utiliser une fenêtre glissante de taille fixe k sur le document et de vérifier si les phrases sur les bords de la fenêtre appartiennent au même segment ou à des segments

différents. Ceci est fait séparément pour le corpus de référence ainsi que pour le corpus de l'hypothèse.

Plus formellement, étant donné deux segmentations, référence (ref) et hypothèse (hyp), pour un corpus contenant N unités (une unité peut correspondre, par exemple, à une phrase dans un texte, à un groupe de souffle dans une transcription, etc.), la mesure p_k est donnée par l'expression suivante :

$$p_k(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} f(f(ref_i, ref_{i+k}), f(hyp_i, hyp_{i+k})) \tag{3.15}$$

avec ref_i et hyp_i correspondant respectivement à la $i^{ème}$ unité de segmentation de la référence et de l'hypothèse.

La fonction f est égale à 1 si ses deux arguments de la fonction f sont égaux. Sinon, elle est égale à 0 (opérateur XOR).

Bien que cette métrique ait permis de résoudre les problèmes de précision, rappel et F-mesure, (PEVZNER et HEARST, 2002) mettent en évidence plusieurs failles avec la métrique p_k :

— la pénalisation injuste des faux négatifs par rapport aux faux positifs,
— l'incapacité de pénaliser les erreurs qui se trouvent dans les k unités de la fenêtre,
— la sensibilité aux variations de la taille des segments,
— les erreurs de proximité trop pénalisées.

3.3.3 WindowDiff

Afin de pallier certaines des limites citées de la mesure p_k, (PEVZNER et HEARST, 2002) proposent une mesure d'évaluation alternative qui est inspirée de la mesure p_k, nommée la mesure WindowDiff (WD). WD s'appuie aussi sur le principe d'une fenêtre glissante. À chaque position de la fenêtre, le nombre de frontières dans la fenêtre glissante est déterminé à la fois pour la segmentation référence et pour la segmentation hypothèse. Si le nombre de frontières n'est pas le même, une pénalité est attribuée. Formellement, WindowDiff se calcule de la manière suivante :

$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0) \tag{3.16}$$

Où $b(i, j)$ représente le nombre de nombre de frontières entre les positions i et j dans le texte et N représente le nombre de phrases dans le texte. Si $b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}) > 0$, alors une pénalité de 1 est attribué, sinon si $b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}) = 0$, alors pas de

pénalisation.

Les auteurs ont montré que cette mesure permet de pénaliser les faux positifs et les faux négatifs plus équitablement, de réduire la sensibilité à la variance de la taille du segment et est capable de prendre les erreurs dans les petits et les grands segments. Cependant, (LAMPRIER et al., 2007) ont montré que les erreurs près des deux extrémités d'un texte sont pénalisées moins que celle du milieu en utilisant la métrique WindowDiff.

3.4 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur les méthodes de segmentation thématique. Deux principales approches ont été détaillées. La première approche est applicable sur n'importe quel corpus du fait qu'elle dépend de la cohésion lexicale dans un texte. La deuxième approche nécessite beaucoup de données d'apprentissage pour obtenir de meilleures performances. L'existence d'un gros corpus de segmentation annoté n'est pas toujours facile. Dans un second temps, nous avons introduit les métriques d'évaluation les plus utilisées dans le domaine de la segmentation thématique.

La segmentation thématique a été utilisée dans la littérature avec différents genres de données telles que les émissions de télévision, les vidéos d'actualités, les réunions, etc. Nous nous intéressons dans cette thèse à la segmentation thématique de cours magistraux. Nous présentons dans le chapitre suivant un aperçu sur les travaux de traitement automatique dans le contexte de cours magistraux dans lequel nous présentons une partie sur les travaux faites pour la structuration thématique des cours magistraux.

Chapitre 4

Aperçu sur le traitement automatique de la parole dans le contexte de cours magistraux

Contents

4.1 Introduction

Récemment, les applications à vocation pédagogique ont acquis une grande visibilité de la part de la communauté de traitement automatique des langues naturelles. Un certain nombre d'études s'est intéressé à développer et à améliorer des applications dans un contexte éducatif. Ces applications visent à améliorer la qualité de l'enseignement, à aider les étudiants et à faciliter la compréhension et la lecture pour différents lecteurs (SHADIEV et al., 2014) tels que les étudiants souffrant de déficits cognitifs ou physiques, les locuteurs et les étudiants non natifs, les étudiants distanciel, les étudiants dans des environnements d'apprentissage traditionnels, etc.

Nous présentons dans ce chapitre un état de l'art sur les travaux de traitement automatique de la parole dans le contexte de l'éducation. La section 4.2 présente un aperçu sur les projets en traitement automatique de la parole pour des cours magistraux. La section 4.3 décrit quelques possibles applications dans le cadre d'éducation. Les sections 4.4 et 4.5 présentent respectivement la reconnaissance de la parole et la structuration automatique de la transcription dans le contexte de cours magistraux.

4.2 Historique des projets en traitement automatique de la parole pour des cours magistraux

Au cours des dernières décennies, un certain nombre d'équipes de recherche se sont intéressées au traitement automatique de la parole lors des cours magistraux, explorant progressivement de multiples techniques et approches. De nombreux projets un peu partout dans le monde ont été financés dans le but d'encourager la recherche en traitement automatique de la parole pour des cours magistraux. Nous parcourons ici quelques uns des principaux projets qui se sont intéressées à cette problématique.

- **Le projet japonais "Spontaneous Speech Corpus and Processing Technology" (FURUI, MAEKAWA et ISAHARA, 2000 ; FURUI et al., 2001) :** ce projet vise à créer un corpus de parole spontanée à grande échelle et à proposer une technique de reconnaissance et de compréhension de la parole spontanée.
- **Le projet "Liberated learning project", (BAIN, BASSON et WALD, 2002) :** l'objectif de ce projet est de fournir une transcription automatique de la parole en temps réel et de haute qualité pour aider les étudiants malentendants.

- **Le projet européan "Computers In the Human Interaction Loop (CHIL)" (LAMEL et al., 2005) :** l'objectif du projet CHIL est d'inclure l'ordinateur dans l'interaction homme-homme. Le cadre applicatif du projet concerne des séminaires, des réunions et des cours dans des salles équipées avec des caméras et des microphones. Le but est de développer un système informatique qui agit en fonction des besoins des utilisateurs en proposant des services appropriés tout en étant le moins intrusif possible. Plusieurs campagnes d'évaluation ont été organisées par ce projet, à savoir : "NIST RT06s Speech-to-Text" (HUANG et al., 2006 ; FÜGEN et al., 2006b) et deux campagnes internes au projet CHIL, "CLEAR06" (NICKEL et al., 2006), "CLEAR07" (MOREAU et al., 2008 ; STIEFELHAGEN et al., 2008).
- **Le projet américain "The MIT spoken lecture processing project" (GLASS et al., 2005 ; GLASS et al., 2007)** du laboratoire CSAIL [1] (Computer Science and Artificial Intelligence Laboratory) au États-Unis : l'objectif du projet MIT Spoken Lecture Processing est d'améliorer l'accès aux enregistrements audiovisuels en ligne des cours universitaires en développant des outils de traitement, de transcription, d'indexation, de segmentation, de synthèse, de récupération d'informations et de navigation.
- **Le projet "LECTRA" (TRANCOSO, NUNES et NEVES, 2006 ; TRANCOSO et al., 2008) :** LECTRA est un projet portugais axé sur la production de contenu de cours multimédia pour les applications de e-learning dont le but est de produire la transcription automatique de la parole sous forme de légende.
- **Le projet "REPLAY" (SCHULTE, WUNDEN et BRUNNER, 2008) :** le but recherché de ce projet est d'automatiser les enregistrements des cours et l'indexation du contenu, ce qui permet de créer des interfaces utilisateur permettant d'accéder aux méta-données.
- **Le projet "New technologies for Voice-converting in barrier-free learning environments (Net4voice)" (LUPPI et al., 2009) :** le projet Net4Voice vise à présenter une méthodologie centrée sur l'utilisation de la reconnaissance de la parole pour l'enseignement. Le but est de générer des transcriptions électroniques synchronisées avec l'audio et la vidéo. Ce projet examine l'impact de la technologie sur des cours dans trois des principales langues européennes telles que l'italien, l'anglais et l'allemand, suivis par différents types d'étudiants, tels que les étudiants handicapés ou les apprenants de langue seconde.
- **Le projet "SpokenMedia" (MURAMATSU et al., 2009) :** le projet SpokenMedia a pour objectif d'améliorer l'efficacité des supports de cours mis sur le Web en ajoutant des briques pour la recherche de segments audio spécifiques et pertinents. La figure 4.1 donne un aperçu du moteur de recherche développé durant ce projet.
- **Le projet "Speech technology integrated learning modules for Intercultural Dia-**

1. https ://www.csail.mit.edu/

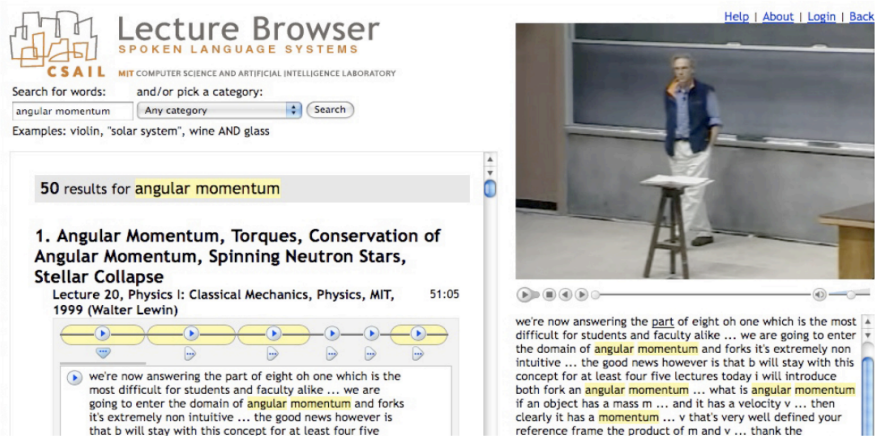


FIGURE 4.1 – Le moteur de recherche du projet "SpokenMedia" (MURAMATSU et al., 2009)

logue" (GELAN, 2010) : l'objectif du projet est d'augmenter la disponibilité et la qualité des supports de cours en ligne pour les langues européennes les moins répandues et les moins enseignées, à l'aide d'un environnement d'apprentissage convivial et hautement accessible et l'intégration des nouveaux technologies Text-to-Speech afin d'aider les apprenants ayant des difficultés de lecture (par exemple, les apprenants dyslexiques ou préférant l'apprentissage auditif).

- **Le projet "APEINTA" (IGLESIAS, MORENO et JIMÉNEZ, 2010 ; IGLESIAS et al., 2016) :** APEINTA est un projet espagnol qui vise le domaine d'éducation. Ce projet propose deux initiatives principales : (1) une transcription en temps réel et de la synthèse vocale en salle de classe et (2) une plate-forme d'apprentissage Web accessible en dehors de la salle de classe avec des ressources numériques accessibles.
- **Le projet "Transcription and Translation of Video Lectures (TransLectures)" (CERDÀ et al., 2012) :** le projet TransLectures a pour objectif de développer des solutions novatrices pour produire des transcriptions et des traductions de cours magistraux.
- **Le projet "Accessing Dynamic Networked Multimedia Events (inEvent)[2]" (BOURLARD et al., 2013) :** l'objectif du projet est de développer de nouveaux moyens de structurer, de récupérer et de partager de grandes archives d'enregistrements multimédias, qui changent de façon dynamique. Cela concerne principalement les réunions, les vidéo-conférences et les cours magistraux.
- **Le projet "Amara" (ABDELALI et al., 2014 ; JANSEN, ALCALA et GUZMAN, 2014) :** le but du projet Amara est d'ajouter des légendes, des sous-titres et des traductions à des vidéos Web. Les vidéos de cours sont parmi les vidéos visées de ce projet.

2. http ://www.inevent-project.eu/

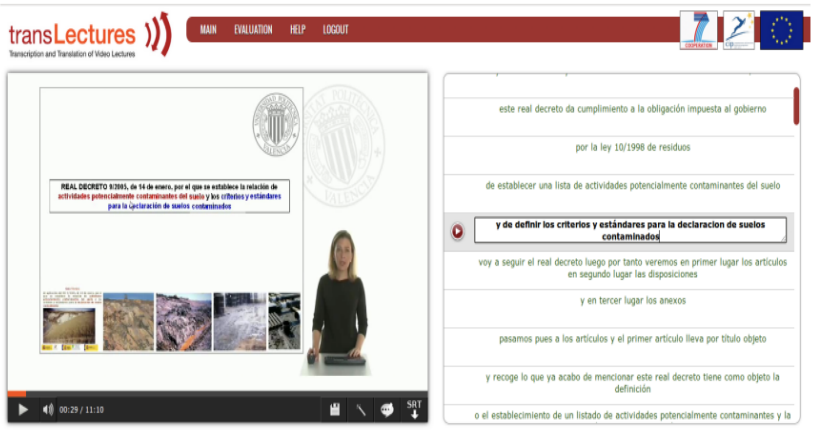


FIGURE 4.2 – Plate-forme du projet "TransLectures" (VALOR MIRÓ et al., 2014)

- **Le projet "European Multiple MOOC Aggregator (EMMA)[3]" (MIRÓ, 2017) :** EMMA a pour objectif de fournir un système permettant de déposer des cours en ligne ouverts dans plusieurs langues à toutes les universités européennes. Le but est ainsi de préserver la richesse du patrimoine culturel, éducatif et linguistique de l'Europe, en promouvant un véritable apprentissage en ligne interculturel et multilingue. La transcription et la traduction automatiques ont été ajoutées pour toutes les vidéos de cours afin de permettre aux apprenants d'accéder aux MOOCs qui ne sont pas dans leur langue maternelle et de les comprendre.
- **Le projet "Translation for Massive Open Online Courses (TraMOOC)[4]" (KORDONI et al., 2016 ; BEHNKE et al., 2018) :** le projet TraMOOC s'intéresse à la traduction automatique en ligne pour fournir une traduction précise et cohérente de supports de cours textuels multi-genres et hétérogènes inclus dans les MOOCs de l'anglais vers onze langues (bulgare, tchèque, allemand, grec, croate, italien, néerlandais, polonais, portugais, russe, chinois).

Ces projets ont été destinés pour plusieurs langues (anglais, japonais, portugais, espagnol, etc.) visant plusieurs domaines d'applications (séminaires, réunions, cours). Un grand nombre de corpus ont été développés dans le cadre de ces projets. Quelques modalités ont été capturé avec ces corpus (audio, vidéo et le support de cours). Le projet PASTEL est destiné pour la langue française. À part les modalités de vidéo, audio et support de cours proposé dans le cadre des projets présentés dans cette section, les données développées dans le projet PASTEL inclut des annotations manuelles supplémentaires : transcriptions manuelles des cours magistraux, annotation des mots du domaine, segmentation thématique dans deux niveaux de

3. https ://platform.europeanmoocs.eu/
4. http ://tramooc.eu

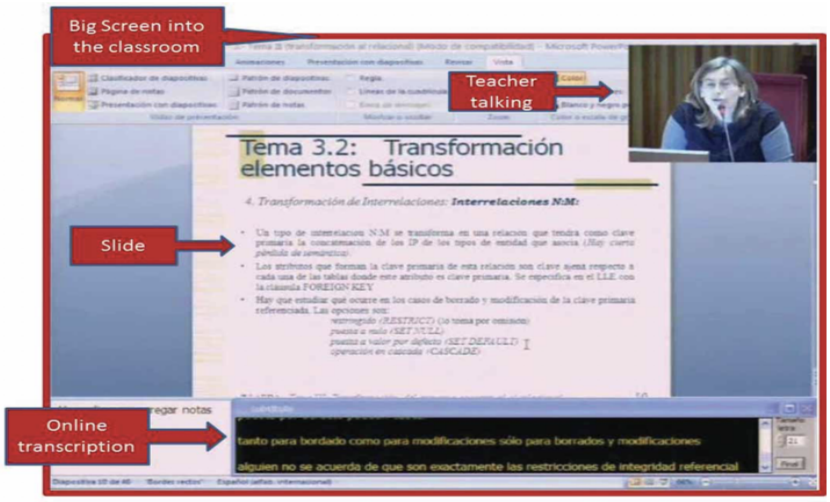


FIGURE 4.3 – Plate-forme du projet "APEINTA" (IGLESIAS et al., 2016)

granularité et alignement des supports du support de cours (diapositives) et la présentation orale pendant les cours. Le projet PASTEL s'intéresse à la transcription automatique, la structuration automatique et l'enrichissement de la transcription avec des documents pédagogiques disponibles dans une base de connaissances extérieure.

4.3 Problématiques de recherche en éducation

Les études et les recherches menés dans le cadre de l'éducation ont pour but de faciliter l'apprentissage en intégrant divers applications. Parmi ces applications on trouve :

- **La reconnaissance de la parole :** La transcription automatique des cours magistraux consiste à convertir le discours de l'enseignant durant une séance de cours en texte. Ces transcriptions peuvent aider les étudiants pour la prise de notes individuelle et collaborative, ou la rédaction de compte-rendu. Cependant, une transcription n'est souvent qu'une première étape pour atteindre d'autres objectifs. Sur la base de transcriptions automatiques, il est possible d'envisager d'autres tâches de traitements automatiques du langage naturel à l'aide de procédés généralement éprouvés sur du langage écrit.
- **La structuration automatique :** généralement, les transcriptions produites par un système de reconnaissance de parole ne sont pas structurées en phrases ou en paragraphes (pas de ponctuation, pas de majuscule). En travaillant à la structuration automatique du discours de l'enseignant, par exemple en découpant ce discours en segments thématiques, il est possible de faciliter la navigation au sein du cours. Plusieurs travaux dans la littérature se sont intéressés à la segmentation de la transcription des

cours magistraux à savoir les travaux de (YAMAMOTO, OGATA et ARIKI, 2003 ; LIN et al., 2004 ; FÜGEN et al., 2006a).

- **La recherche d'information :** au cours des dernières années, l'ensemble des matériels pédagogiques, notamment les présentations et les enregistrements de cours, a connu une croissance considérable dans de nombreux établissements universitaires du monde entier. Le matériel enregistré est souvent mis à disposition en ligne sur un site Internet, généralement accompagné de méta-informations telles que les mots-clés, les catégories, le titre, l'auteur ou les liens. Ces méta-informations aident les étudiants à trouver le cours dont ils ont besoin. Cependant, au vu de la longue durée d'un cours (généralement plus d'une heure), la recherche des informations pertinentes au sein d'un cours (passage, partie,etc) est difficile et fastidieuse. Une recherche automatique est capable d'augmenter l'accessibilité à des parties de cours enregistrés. Un nombre considérable de navigateurs de cours a été développé (FUJII, ITOU et ISHIKAWA, 2006 ; SZÖKE et al., 2010 ; RIEDHAMMER, GROPP et NÖTH, 2012).
- **Le résumé automatique :** la quantité de supports de cours existants en ligne est énorme. Ces cours ne comportent pas tous des méta-informations qui aident l'utilisateur à comprendre le contenu du cours. Si nous pouvons résumer la transcription des cours, le matériel sera beaucoup plus facile à référencer et plus accessible. Le résumé automatique des cours a été exploré dans quelques travaux (FUJII et al., 2008 ; ZHANG et FUNG, 2009 ; KIM et KIM, 2016).
- **La traduction automatique :** les étudiants étrangers dans les universités ont souvent du mal à suivre les cours. Grâce au développement rapide de la technologie de communication, beaucoup de travaux se sont intéressés à la traduction de discours de l'enseignant afin de faciliter la compréhension de cours pour ces étudiants. Citons à titre d'exemple les travaux de (FÜGEN, 2008 ; CHO et al., 2013 ; MÜLLER et al., 2016).
- **L'extraction de phrases ou mots clés :** de nos jours, de nombreuses universités proposent des cours en ligne. Les vidéos de cours sont parfois accompagnés de méta-informations. Parcourir la vidéo de conférence en entier devient un travail fastidieux pour trouver le passage intéressant. Afin de faciliter le travail, plusieurs travaux proposent des systèmes automatiques qui génèrent automatiquement les phrases ou mots clés pertinentes (BALAGOPALAN et al., 2012 ; CHEN et al., 2010 ; BALASUBRAMANIAN, DORAISAMY et KANAKARAJAN, 2016).

4.4 Reconnaissance de la parole dans le contexte de cours magistraux

Cette section présente les travaux de reconnaissance de la parole faites dans le contexte de cours magistraux.

4.4.1 Adaptation des modèles de langage dans le contexte de cours magistraux

Les cours sont généralement spécialisés et spécifiques à un domaine. Ce contexte est difficile pour les systèmes de reconnaissance de la parole car ils sont sensibles à la variabilité des thèmes.

De ce fait, une large partie des travaux de recherche de traitement automatique de la parole dans le contexte de cours magistraux a été consacré à l'amélioration de l'efficacité du modèle de langage. Les méthodes d'adaptation du modèle de langage dans le contexte de cours magistraux peuvent être classifiées en trois niveaux (MARQUARD, 2012) :

- Adaptation "macro" : il s'agit d'adapter une seule fois le modèle de langage pour tous les cours afin de prendre en compte les spécificités des cours magistraux, en considérant que les cours couvrent des spécificités qui ne se trouvent pas dans d'autres types d'applications, tels que les émissions télévisées, les articles de journaux, etc. (CERVA et al., 2012).
- Adaptation "mesco" : il s'agit d'adapter le modèle de langage, pour chaque cours séparément, puisque chaque cours possède une terminologie propre à lui.
- Adaptation "micro" : il s'agit d'adapter le modèle de langage pour chaque partie du cours en utilisant des informations provenant des segments ou des transitions à l'intérieur des cours (YAMAZAKI et al., 2007).

Les données utilisées pour réaliser l'adaptation ont été de différents genres à savoir les articles de journaux (CERVA et al., 2012), les livres et les documents techniques (CERVA et al., 2012), les diapositives de cours (YAMAZAKI et al., 2007 ; KAWAHARA, NEMOTO et AKITA, 2008 ; MIRANDA, NETO et BLACK, 2013 ; AKITA, TONG et KAWAHARA, 2015) ou encore les textes web (MASUMURA, HAHM et ITO, 2011).

Les techniques d'adaptation sont aussi multiples, à savoir l'interpolation linéaire (MARTÍNEZ-VILLARONGA et al., 2013 ; YAMAZAKI et al., 2007 ; MAERGNER, WAIBEL et LANE, 2012 ; AKITA, TONG et KAWAHARA, 2015), les modèles caches (AKITA, TONG et KAWAHARA, 2015) ou encore les modèles de thèmes (HSU et GLASS, 2006 ; GLASS et al., 2007).

4.4.2 Évaluation des systèmes de reconnaissance de la parole dans le contexte de cours magistraux

La performance de la plupart des systèmes de reconnaissance de parole dans le contexte de cours magistraux a été évaluée en utilisant la métrique WER (CERVA et al., 2012 ; BELL et al., 2013 ; YAMAZAKI et al., 2007 ; KAWAHARA, NEMOTO et AKITA, 2008).

Quelques travaux, tels que (PARK, HAZEN et GLASS, 2005 ; YAMAZAKI et al., 2007), ont considéré que cette métrique est insuffisante et difficile à interpréter et ils ont utilisé les métriques standards de recherche d'informations (Précision, Rappel, F-mesure) comme alternative pour calculer la performance de la transcription. L'évaluation du système par les mesures rappel et précision nécessite de définir tout d'abord les mots-clés sur lesquels ces métriques vont être appliquées.

4.5 Structuration automatique de la transcription dans le contexte de cours magistraux

Cette section présente les travaux de la structuration automatique de la transcription faites dans le contexte de cours magistraux.

4.5.1 Structure générale d'un cours

Un cours magistral contient un ensemble de thèmes. Chaque thème peut être divisé en sous-thèmes. Un thème ou un sous-thème peut contenir un ou plusieurs éléments tels qu'une ou plusieurs diapositives, des explications sur le tableau et le discours de l'enseignant... La figure 4.4 schématise cette structure générale d'un cours.

La structuration de la transcription d'un cours est très utile car cela permet d'extraire de précieuses informations visuelles pour l'utilisateur. Cette structuration peut être effectuée à plusieurs niveaux dans le cadre de cours magistraux. Nous détaillons dans cette section les types de structuration possibles, après avoir donner un aperçu sur la difficulté de la tâche.

4.5.2 Difficultés de la structuration automatique des cours magistraux

(LIN, 2006) a essayé de lister les difficultés de la structuration thématique de cours magistraux.

— **Le non-professionnalisme de l'enseignant pour des enregistrements vidéos :** ceci signifie que le caméra-man n'est pas préalablement formé en avance à ce travail. Ces vidéos ne passent pas par un travail de montage. Un non professionnel signifie également

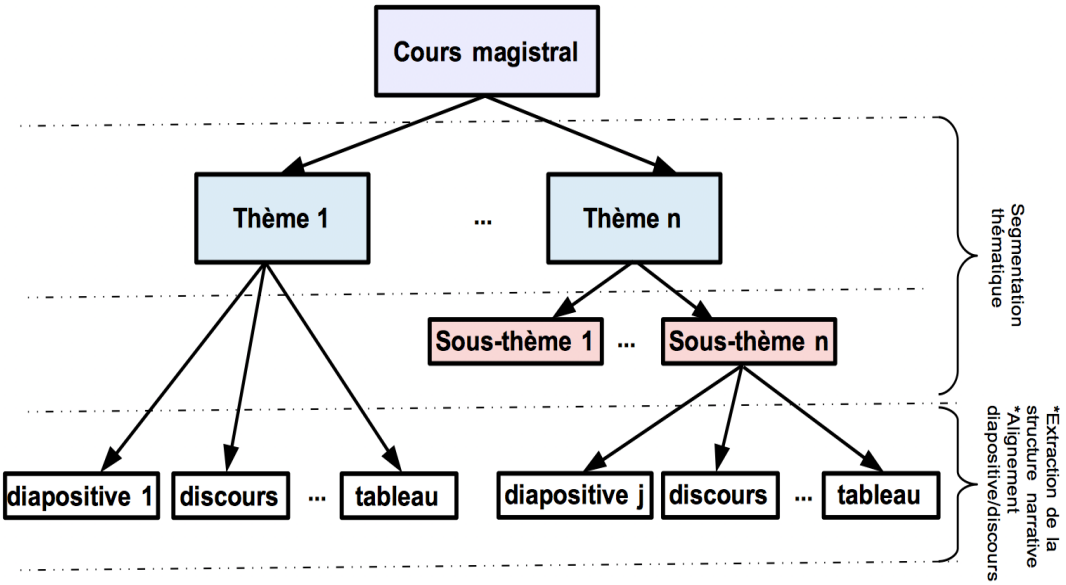


FIGURE 4.4 – Structure générale d'un cours

que les orateurs du cours (les enseignants) ne sont pas formés professionnellement, contrairement aux orateurs de la diffusion de télévisions ou de radios.

— **L'absence d'une structure syntaxique :** les vidéos de cours magistraux manquent généralement de changements de scènes. Les changements de scènes ne sont pas significatifs et ne correspondent pas à la structure thématique des vidéos de cours. De plus, il n'y a pas de notion d'histoire (*story* en anglais) comme dans la segmentation des vidéos d'actualités où le segment est bien défini et associé à une durée relativement courte. Dans le contexte de cours magistraux, la durée et le nombre de segments thématiques peuvent varier entre les différents cours et les enseignants.

— **L'hétérogénéité :** contrairement au corpus de vidéos d'actualités radiodiffusées, les vidéos de cours magistraux peuvent être assez hétérogènes. Par exemple, dans un cours, un enseignant utilise uniquement des diapositives de cours et les suit à la lettre. Lors d'un autre cours, l'enseignant peut utiliser à la fois le tableau et les diapositives. Chaque enseignant a son propre style qui varie d'un enseignant à un autre. Ceci indique que les méthodes de segmentation conçues pour un enseignant ou un type de cours risquent de ne pas fonctionner avec un autre enseignant ou un autre type de cours.

— **La spontanéité du discours :** le discours dans un cours magistral tend à être plus spontané que celui dans une vidéo d'actualité. Le discours dans un cours magistral se caractérise par la présence de mots de remplissage («d'accord», «bien», etc.) et des pauses non lexicales («euh» ou «hum»). Les silences et les événements sans discours ont tendance à être plus longs entre les histoires dans les vidéos d'actualité.

— **Les erreurs due à la transcription automatique :** les erreurs de transcription automatique sont susceptibles de dégrader la qualité de la segmentation automatique.

4.5.3 Segmentation thématique dans le cadre de cours magistraux

La structure d'un cours présenté dans la figure 4.4 consiste à une structure hiérarchique. Plusieurs travaux se sont intéressés au niveau 3 (niveau contenant des sous-thèmes) pour détecter les frontières de zones homogènes au niveau du contenu de la transcription : la segmentation thématique (chapitre 3).

Comparée à d'autres types de vidéos, tels que les films, les émissions radiodiffusées et les textes, la segmentation des cours magistraux est particulière (LIN, 2006).

(LIN et al., 2005) a utilisé deux types de caractéristiques pour effectuer la segmentation : des caractéristiques basées sur le contenu et des caractéristiques basées sur le discours. Les caractéristiques basées sur le contenu sont composées de structures linguistiques telles que les phrases nominales, les classes de verbes, les racines de mots. Les caractéristiques basées sur le discours sont composées des pronoms et des phrases de repère. Ces caractéristiques permettent la création d'un espace vectoriel en utilisant le poids de chaque entité dans des fenêtres de taille fixe de phrases de transcription. Ensuite, la similarité est calculée en utilisant une fenêtre et la segmentation finale est obtenue avec le même principe de l'algorithme TextTiling (section 3.2). La performance de cette méthode atteint 70% de F-mesure. Le résultat avec TextTiling est 56%. (BALAGOPALAN et al., 2012) ont apporté quelques modifications à l'algorithme TextTiling en utilisant dans leur espace vectoriel les phrases clés détectées automatiquement. Ceci a permis d'améliorer les résultats.

(ALHARBI et HAIN, 2015) comparent beaucoup d'algorithmes de segmentation pour structurer des cours de physique et des cours d'économie. Ces algorithmes sont le C99, U00, LCSeg et une méthode de segmentation bayésienne (section 3.2). Les deux mesures Pk et le WindowDiff ont été utilisées pour comparer la performance de ces algorithmes. La combinaison de la segmentation bayésienne et des indices de discours a obtenu le meilleur score.

L'algorithme MinCut a été utilisé dans les travaux de (GLASS et al., 2007 ; MALIOUTOV, 2006). (MALIOUTOV, 2006) a comparé MinCut avec l'algorithme U00 (section 3.2) et il a montré dans ces résultats expérimentaux que MinCut est plus performant dans les environnements vocaux bruyants.

Malgré tous les efforts faites dans ce domaine, les performances sont encore loin d'être optimales et ne peuvent être comparées aux performances humaines.

4.5.4 Alignement du discours de l'enseignant avec les diapositives

Les résultats de segmentation de l'état de l'art montrent bien la difficulté de la tâche dans le cadre de cours magistraux. Pour remédier à ces difficultés, beaucoup de travaux ont décidé de structurer les cours à un niveau plus fin (figure 4.4) qui est l'alignement de la transcription avec les diapositives du support de cours .

(JUNG, SHIN et KIM, 2018), les auteurs représentent les diapositives et les phrases prononcées sous forme de vecteurs de sacs de mots, chaque mot étant pondéré par un score TF-IDF. Ensuite, la similarité cosinus est calculée entre les vecteurs. Enfin, ils attribuent à chaque phrase parlée le texte de diapositive le plus similaire.

Cette technique a été également utilisée dans le travail de (YAMAMOTO, OGATA et ARIKI, 2003) en ajoutant une étape de post-traitement pour corriger l'ordre des alignements possibles. Les corrections ont été implémentées manuellement, sous la forme de règles, telles que le premier segment de cours et la première section de manuel doivent être alignées ensemble, l'alignement de segment de parole doit être conforme au contexte (c'est-à-dire si les segments de cours précédents et suivants sont alignés sur la même section de manuel, la section actuelle doit donc l'être). La performance est de 67,7% avant l'étape de post-traitement et atteint 89% après le post-traitement.

(LU et al., 2014) ont proposé un ensemble de trois méthodes pour aligner les diapositives sur la transcription, à savoir l'alignement de mots, le SVM structurée et l'intégration du score. Les méthodes peuvent être supervisées ou non supervisées, mais reposent toujours sur une grande quantité de données. La meilleure performance a été obtenu en utilisant une méthode supervisée (73,15% en utilisant une transcription automatique et 77,16% en utilisant une transcription manuelle)

4.5.5 Extraction de la structure narrative d'un cours

Un autre type de structuration est l'extraction de la structure narrative d'un cours. La structure narrative résulte du fait de l'utilisation de différents types de contenu de présentation tels que des diapositives, des pages Web et des textes écrits sur tableau (figure 4.4). L'extraction de cette structure est utile pour segmenter une vidéo éducative afin de faciliter l'accès au contenu et la navigation non linéaire au sein du matériel présenté. Beaucoup de travaux se sont intéressés à cette problématique (LIU et al., 2006 ; DORAI, ORIA et NEELAVALLI, 2003).

4.6 Conclusion

Dans ce chapitre nous avons dressé un bref historique des principaux travaux de traitement automatique dans le cadre de cours magistraux. Ceci nous a permis de tracer l'origine

du domaine et de suivre les évolutions qu'il connu ces dernières années. Dans le cadre de cette thèse, nous nous focalisons sur la transcription automatique des cours magistraux et la structuration de la transcription automatique.

Afin de réaliser ces tâches, il est nécessaire d'avoir un corpus et d'avoir une idée des caractéristiques des données afin de choisir les techniques les mieux appropriées. Ceci est l'objectif du chapitre suivant.

DEUXIÈME PARTIE

Contributions

Chapitre 5

Cadre expérimental : corpus PASTEL et système pour la reconnaissance de la parole

Contents

5.1 Introduction

Afin de construire et d'évaluer un système de reconnaissance de la parole et de structuration de la transcription, il est indispensable de disposer d'un jeu de données adapté à la tâche. Depuis les années 2000, un grand nombre de travaux ont été publiés sur le traitement automatique de cours magistraux, faisant de lui un domaine de recherche très actif (section 4.2). En revanche, les corpus existants ne sont pas nombreux et la plupart des corpus existants ne sont pas accessibles à la communauté scientifique. À notre connaissance, il n'existe pas de corpus de cours magistraux transcrits et segmentés en thèmes pour la langue française. Dans ce chapitre, nous présentons un nouveau corpus de cours magistraux intitulé PASTEL. Ce corpus est composé de vidéos de cours, de supports de cours et de transcriptions manuelles du discours en situation de cours magistraux. Il s'accompagne d'une segmentation manuelle et d'une annotation en mots clés. La segmentation manuelle va également servir aux développement et à l'expérimentation des techniques pour la structuration automatique. L'annotation en mots clés vise, d'une part, à soutenir la tâche d'enrichissement de la transcription avec des ressources extérieures relatives aux concepts difficiles à comprendre. Ces expressions clés vont servir à l'évaluation des outils développés afin de les repérer automatiquement. D'autre part, cette annotation va permettre d'évaluer les erreurs du système de reconnaissance de parole pour ces expressions. Nous décrivons dans un premier temps le schéma d'annotation adopté puis nous dressons quelques statistiques et caractéristiques du corpus. Dans un deuxième temps, nous détaillons le système de reconnaissance de la parole, avec lequel toutes les expériences ont été menées dans cette thèse et la chaîne de traitements de notre processus d'adaptation des modèles de langage.

5.2 Sources du corpus

Le corpus consiste en une collection de cours de différents domaines informatique (traitement automatique des langues, introduction à l'informatique, etc) en première année de licence d'informatique à l'Université de Nantes. Il est constitué de cours qui proviennent de deux sources : le projet COCo et la plateforme Canal-U. Les cours dont la source est COCo ont été filmés et mis en ligne sur une plateforme web [1] suite au projet Comin Open Courseware "COCo" qui a eu lieu de 2013 jusqu'au 2016. Le but principal de ce projet a été de mobiliser les annotations vidéo dans des contextes pédagogiques et de recherche ainsi que de promouvoir les ressources éducatives ouvertes et les licences ouvertes (AUBERT, PRIÉ et CANELLAS,

1. http ://www.comin-ocw.org/

2014 ; AUBERT et JAEGER, 2014 ; MOUGARD et al., 2015 ; CANELLAS, AUBERT et PRIÉ, 2015). Six cours ont été téléchargés à partir de la plateforme de ce projet. Trois autres cours ont été téléchargés à partir de la plateforme Canal-U [2]. Canal-U est un site contenant des ressources audiovisuelles de l'enseignement supérieur et de la recherche.

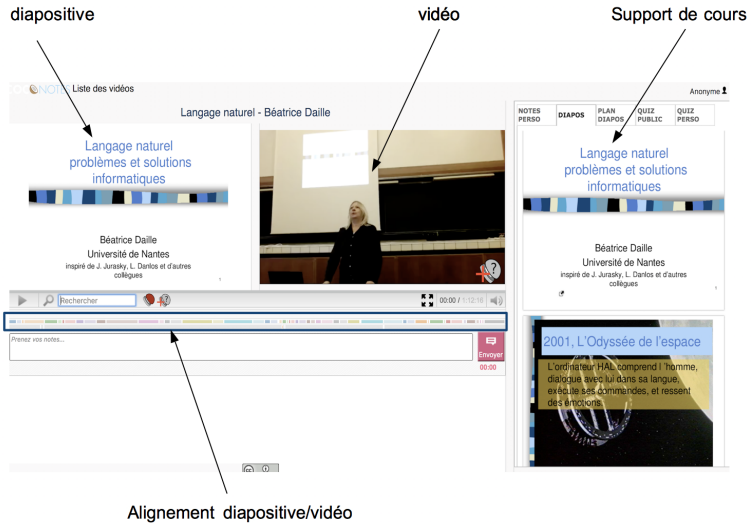


FIGURE 5.1 – Plateforme COCo

La figure 5.1 représente une capture d'écran de la plateforme COCo [3]. Chaque cours dispose de la vidéo du cours, du support de cours et d'un alignement entre la vidéo et les diapositives du support de cours. Les cours dont la source d'origine est Canal-U ne disposent pas du support de cours et par conséquent d'aucune information d'alignement.

5.3 Guide et processus d'annotation pour le corpus PASTEL

La mise en place d'un guide d'annotation consiste à rédiger un ensemble de règles définies, de sorte que n'importe quel annotateur qui utilise ces règles puisse annoter le document de la même façon qu'un autre annotateur qui utilise ce même guide. Nous présentons, dans cette section, les règles d'annotation établies pour le corpus PASTEL ainsi que le processus d'annotation.

5.3.1 Transcription manuelle

Cette opération a été réalisée en deux passes. La première passe a consisté à effectuer la transcription du cours par le biais d'un système générique de reconnaissance de la parole

2. https ://www.canal-u.tv/
3. https ://www.comin-ocw.org/contents/infol1/20131115/

(dont le modèle de langage ne contient pas de données du domaine). L'expert humain est ensuite intervenu, lors de la seconde passe, pour corriger manuellement les erreurs dans la transcription automatique.

Le logiciel Transcriber[4] a été utilisé (BARRAS et al., 1998 ; BARRAS et al., 2001) pour effectuer l'annotation. Transcriber est un logiciel optimisé pour la transcription et l'annotation de corpus volumineux. Les conventions généralement utilisées pour les transcriptions de campagnes d'évaluation (GRAVIER et al., 2004) ont servi de guide pour transcrire les cours enregistrés.

5.3.2 Segmentation thématique

Un des objectifs applicatifs du projet PASTEL consiste à découper le cours en une suite de séquences homogènes de quelques minutes, de manière à ce que chaque séquence puisse être considérée comme une partie complète de cours. Ceci a pour but de permettre aux étudiants de naviguer aisément dans une vidéo du cours, c'est-à-dire dans le discours transcrit de l'enseignant. Nous nous sommes intéressés, d'une part, à identifier les séquences insécables qui constituent des unités minimales d'information dans une vidéo de cours "scènes", et d'autre part, à identifier les séquences qui constituent des tout homogènes pouvant se visionner indépendamment ou tout au moins avec la possibilité de marquer des pauses entre chaque "épisode".

Nous qualifions ce type de découpage par une segmentation thématique. Alors, «*Qu'est-ce qu'un segment thématique dans un cours qui est de base monothémathique (l'objet principal du cours) ?*».

Nous avons décidé de répondre à cette question en étudiant la possibilité de nous appuyer sur les diapositives des supports de présentations de cours. Un segment thématique dans un cours peut ainsi être le discours lié à :

— la structuration logique en sections hiérarchiques : dans le cadre de la segmentation d'articles scientifiques, des livres, etc, une section ou une sous-section peut être considérée comme un segment thématique. Toutefois, au sein d'un support de cours présenté en diapositives, la notion de section n'est pas toujours présente. Lorsque cette information est présente, voire explicitée par un sommaire, celle-ci se révèle être une source d'information fiable pour segmenter thématiquement le discours. Ceci dépend ainsi fortement de la technique de préparation d'un cours de l'enseignant, qui n'est pas toujours la même.

— la séquence de diapositives avec un même titre : les diapositives ayant le même titre appartiennent souvent au même segment thématique. Toutefois, au sein d'un support de cours présenté en diapositives, l'enseignant peut utiliser différents titres pour présenter

4. http ://trans.sourceforge.net/en/presentation.php

la même notion.

— une diapositive : une diapositive semble correspondre à l'unité minimale utilisée pour la transmission d'une idée ou notion. Nous relevons le fait que, parfois, plusieurs idées peuvent être traitées dans une diapositive mais que leurs pertinences ne permettent pas de les considérer comme indissociables les unes des autres. Nous relevons également que certaines idées peuvent être développées en plus d'une diapositive.

Nous avons donc considéré ces informations (découpage en diapositives, marquage du découpage en sections, présence d'un sommaire et homogénéité dans les titres) pour soutenir le travail de segmentation des vidéos. Une frontière thématique ne peut se situer qu'au voisinage d'un changement de diapositive pendant le cours. Par conséquent, à chaque changement de diapositive, il est nécessaire d'annoter en cas de changement thématique :

— l'instant exact du changement thématique défini comme étant positionné entre deux mots,
— la granularité du changement thématique.

L'analyse des vidéos et diapositives nous a conduit à définir deux niveaux d'annotation en accord avec nos objectifs applicatifs. La granularité 1 est utilisée pour marquer le fait qu'une nouvelle notion du cours est abordée tout en restant dans le même sous-thème (scènes). La granularité 2 est utilisée lorsqu'il y a un changement sous-thématique plus général qui permet d'arrêter l'apprentissage à ce moment-là et de reprendre plus tard l'apprentissage d'autres notions (épisode). À partir des segments thématiques de granularité 2, il est possible d'obtenir un découpage en vidéos adaptées pour la création de MOOCs (Massive Open Online Course)[5].

Nous avons aussi ajouté la notion d'interruption à l'annotation. Il s'agit de marquer les segments temporels qui sont relatifs à la gestion en direct d'un groupe d'étudiants et d'un matériel : gestion des interruptions, d'une panne matérielle, du bavardage, etc.

L'annotation a été effectuée par deux annotateurs étudiants en master en linguistique à l'aide de l'outil ELAN (WITTENBURG et al., 2006 ; AUER et al., 2010). ELAN[6] (EUDICO Linguistic Annotator) est un outil d'annotation linguistique conçu pour la création d'annotations textuelles pour les fichiers audio et vidéo. Il contient une interface adaptée et il consiste à écouter l'enregistrement sonore pour effectuer les annotations nécessaires.

Concernant la segmentation des cours dont la source d'origine est Canal-U, les changements de diapositive ont été visibles via la vidéo, ce qui n'a pas gêné le processus de segmentation thématique manuelle.

5. https ://fr.wikipedia.org/wiki/Massive_Open_Online_Course
6. https ://tla.mpi.nl/tools/tla-tools/elan/

5.3.3 Extraction manuelle des expressions clés

Les motivations de ce travail d'annotation sont doubles. La première est de déterminer dans quelle mesure ces expressions clés ont été bien reconnues par le système de reconnaissance de la parole. La seconde motivation vise à soutenir l'étude d'un cas d'usage du projet, à savoir l'enrichissement de la transcription à l'aide de ressources extérieures (documents extérieurs issus de bases de données spécifiques, d'encyclopédies en ligne comme Wikipédia, ou d'autres sources du web). Ces expressions clés devraient permettre d'évaluer la précision de nos outils à les repérer automatiquement.

Nous avons considéré, comme expressions clés du domaine, les expressions linguistiques faisant référence à des concepts, les objets ou les entités étant essentiels à la compréhension de la diapositive actuelle ou d'une transcription donnée. Nous avons inclus tous les termes scientifiques et techniques ainsi que les acronymes et expressions nous permettant d'aller plus loin dans le sujet du cours.

L'annotation en expressions clés a été effectuée par un annotateur étudiant en licence 3 d'informatique. L'annotation a été faite pour la transcription automatique et pour le support de cours (les diapositives). La figure 5.2 correspond à un exemple d'annotation manuelle pour les diapositives du cours *"Les fonctions"*. La figure 5.3 correspond à l'annotation manuelle pour la transcription relative à la diapositive de la figure 5.2.

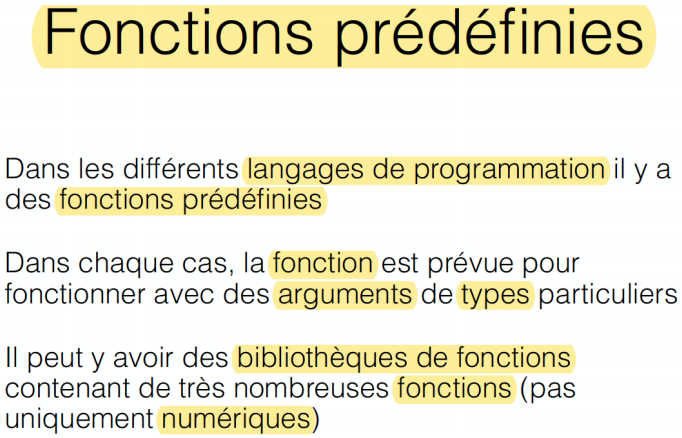


FIGURE 5.2 – Exemple d'annotation manuelle, à partir du support de cours. Extrait du cours de Jérémie Bourdon : *"Les fonctions"*

1367.018 1369.415 on va maintenant parler des fonctions prédéfinies
1369.415 1374.946 puisque dans tous les langages de programmation, il y en a plus ou moins des fonctions prédéfinies, mais il y en a
1374.946 1379.280 donc je vous ai dit : il y a racine carrée, il y a de grandes chances qu'elle soit prédéfinie dans le langage
1379.280 1384.433 que vous allez manipuler ; javascript, elle est prédéfinie.
1384.433 1389.219 valeur absolue, elle est prédéfinie donc il y en a il y en a tout un tas.
1389.219 1393.570 dans chacun de ces de ces prédéfinitions,
1393.570 1397.723 la fonction est prévue pour fonctionner avec des types particuliers qu'il faudra connaître.
1397.723 1400.691 et ça, c'est donné dans la documentation
1400.691 1402.474 du javascript.
1402.474 1408.412 ça se trouve sur internet, sur plein de sites qui vous sont indiqués.
1408.412 1409.325 voilà
1409.325 1412.371 la plupart du temps, ces fonctions
1412.371 1413.849 elles sont groupées
1413.849 1418.006 les fonctions qui font des calculs mathématiques sont regroupées dans une bibliothèque qui s'appelle maths
1418.006 1419.558 maths point quelque chose donc
1419.558 1424.293 racine carrée, elle s'appelle pas racine carrée, elle s'appelle maths point sqrt
1424.293 1425.544 en javascript.
1425.544 1428.430 valeur absolue s'appelle maths point abs.
1428.430 1432.752 et il y a un certain nombre de bibliothèques comme ça, qui permettent de manipuler des fonctions
1432.752 1435.511 qui ne sont pas uniquement numériques.

FIGURE 5.3 – Exemple d'annotation manuelle à partir de la transcription correspondant à la figure 5.2. Extrait du cours de Jérémie Bourdon : *"Les fonctions"*

5.4 Statistiques du corpus

Le corpus annoté global comprend 9 cours magistraux (MDHAFFAR, LAURENT et ESTÈVE, 2018b ; MDHAFFAR et al., 2020). La durée totale du corpus est d'environ 10 heures. Le tableau 5.1 présente quelques statistiques de notre corpus. Les deuxième, troisième et quatrième colonnes du tableau représentent, respectivement, le nombre de segments de "granularité 1", de "granularité 2" et de "interruption". Les colonnes 5 et 6 représentent le nombre d'expressions-clés annotées pour les transcriptions et les diapositives, respectivement. La colonne 7 représente le nombre de diapositives dans chaque cours et la colonne 8 contient la durée de chaque cours. Enfin, la dernière colonne indique la source du cours (notez que les 3 cours ne contenant pas de diapositives proviennent de Canal-U). Le nombre de locuteurs dans ce corpus est 7. Les cours *Introduction à l'informatique*, *Introduction à l'algorithmique* et *Les fonctions* sont donnés par le même enseignant.

7. http ://www.comin-ocw.org/contents/infol1/20140911/
8. http ://www.comin-ocw.org/contents/infol1/20140912/
9. http ://www.comin-ocw.org/contents/infol1/20140925/
10. http ://www.comin-ocw.org/contents/infol1/20131010/
11. http ://www.comin-ocw.org/contents/infol1/20131128/
12. http ://www.comin-ocw.org/contents/infol1/20131115/
13. https ://www.canal-u.tv/video/universite_rennes_2_crea_cim/l_architecture_de_la_republique_en_france_au_xixe_et_au_xxe_siec
14. https ://www.canal-u.tv/video/les_amphis_de_france_5/une_methode_traditionnelle_le_cours_mauger_2.3014

TABLE 5.1 – Statistiques du corpus (I : Interruption, G : Granularité, EC : expression clé, t : transcription manuelle, d : diapositive)

Nom du cours	**G. 1**	**G. 2**	**I**	$\mathbf{EC_t}$	$\mathbf{EC_d}$	#d	**Durée**	**Source**
Introduction à l'informatique[7]	31	2	2	47	38	75	1h 04mn 42s	COCo
Introduction à l'algorithmique[8]	38	10	3	25	35	62	1h 17mn 28s	COCo
Les fonctions[9]	35	3	3	109	78	137	1h 14mn 29s	COCo
Réseau sociaux et graphes[10]	43	7	7	53	65	64	1h 05mn 51s	COCo
Algorithmique distribuée[11]	72	5	3	232	146	73	1h 16mn 30s	COCo
Langage naturel[12]	52	5	5	106	66	55	1h 09mn 35s	COCo
Architecture de la république[13]	49	7	0	-	-	-	1h 21mn 14s	Canal-U
Méthode traditionnelle[14]	12	7	1	-	-	-	0h 41mn 02s	Canal-U
Imagerie	57	0	1	-	-	-	1h 08mn 14s	Canal-U
Total	389	46	25	572	428	466	10h 19mn 05s	-

5.5 Analyse des annotations du corpus

L'analyse des annotations représente le cœur des études linguistiques sur les données. Dans cette section, nous nous intéressons au traitement et à l'analyse des annotations de segmentation thématique et d'extraction d'expressions clés.

5.5.1 Analyse des segments thématiques

Un segment thématique est constitué du discours lié à un ensemble de diapositives. La durée d'un segment peut aller de quelques secondes à des dizaines de minutes. Les tableaux 5.2, 5.3 et 5.4 présentent les statistiques de durée des segments, pour chaque cours, et pour chaque type de segment (granularité 1, granularité 2 et interruption). La deuxième colonne représente le nombre de segments du type considéré. La troisième colonne représente la durée moyenne des segments et les colonnes 5 et 6 représentent, respectivement, la durée minimale et la durée maximale parmi les segments de chaque cours.

Les statistiques dans les tableaux 5.3, 5.2 et 5.4 mettent en évidence la forte disparité de la taille des segments entre les différents cours mais aussi au sein d'un même cours.

5.5.2 Analyse des expressions clés

Dans le corpus, les expressions clés peuvent constituer un avantage pour plusieurs systèmes de traitement automatique de langage naturel tels que la segmentation automatique ou le résumé automatique. Il est alors intéressant de connaître, dans le corpus :

— le nombre d'occurrences des expressions clés : le nombre d'occurrences désigne le nombre d'apparitions d'une expression quelconque dans le corpus. Il est également

TABLE 5.2 – Nombre et durée (moyenne, minimale et maximale) des segments "granularité 1"

Nom du cours	**G. 1**	**durée moy**	**durée min**	**durée max**
Introduction à l'informatique	31	123,0	16,3	307,8
Introduction à l'algorithmique	38	106,6	18,7	248,4
Les fonctions	35	124,3	42,2	393,8
Réseau sociaux et graphes	43	85,3	11,6	475,6
Algorithme distribuée	72	53,8	6,4	204,4
Langage naturel	52	80,3	5,2	215,2
Architecture république	49	92,3	14,4	317,6
Méthode traditionnelle	12	187,2	17,4	724,5
Imagerie	57	63,0	4,0	224,1

TABLE 5.3 – Nombre et durée (moyenne, minimale et maximale) des segments "granularité 2"

Nom du cours	**G. 2**	**durée moy**	**durée min**	**durée max**
Introduction à l'informatique	2	1941,3	455,6	3427,01
Introduction à l'algorithmique	10	476,1	129,9	1041,5
Les fonctions	3	1036,5	584,6	1672.5
Réseau sociaux et graphes	5	960,8	285,6	1871,1
Algorithme distribuée	5	1114,7	466,4	1824,4
Langage naturel	5	960,8	285,6	1871,1
Architecture république	7	696,3	350,5	1179,6
Méthode traditionnelle	7	340,6	45,4	874,3
Imagerie	0	-	-	-

TABLE 5.4 – Nombre et durée (moyenne, minimale et maximale) des segments "interruption"

Nom du cours	**I**	**durée moy**	**durée min**	**durée max**
Introduction à l'informatique	2	21,8	16,8	26,9
Introduction à l'algorithmique	3	60,4	46,0	80,8
Les fonctions	3	39,6	22,5	51,5
Réseau sociaux et graphes	7	13,8	7,3	28,1
Algorithme distribuée	3	17,2	6,3	38,7
Langage naturel	5	14,7	6,3	28,5
Architecture république	0	-	-	-
Méthode traditionnelle	1	77,2	77,2	77,2
Imagerie	1	24,9	24,9	24,9

important de savoir quelles expressions clés, annotées dans les diapositives, sont présentes ou absentes dans la transcription manuelle, et inversement.

— la répartition des expressions clés : la répartition d'une expression clé dans un corpus correspond à l'ensemble des emplacements où cette expression apparaît : apparaissent-elles de manière uniforme dans la totalité du corpus ou au contraire sont-elles plutôt

localisées dans quelques diapositives ?

Nous étudions ces différents points dans les sections 5.5.2.1 et 5.5.2.2.

5.5.2.1 Occurrence des expressions clés

Les figures 5.4, 5.5 et 5.6 présentent le nombre d'occurrences des expressions clés, *annotées à partir du support de cours*, dans la transcription manuelle et dans le support de cours pour trois cours du corpus. Les figures 5.7, 5.8 et 5.9 illustrent les mêmes calculs mais pour les expressions clés, *annotées à partir de la transcription manuelle*. Le nombre d'occurrences des expressions clés dans la transcription manuelle est représenté par les bâtons rouges. Le nombre d'occurrences des expressions clés dans le support de cours est représenté par les bâtons bleus.

Ces figures montrent que les expressions clés sont différentes en termes de nombre d'occurrences et d'omniprésence. On observe que le nombre d'occurrences des expressions clés n'est pas similaire dans les diapositives et dans le discours de l'enseignant et que cette différence est plus ou moins importante selon les enseignants. On note que certaines expressions clés apparaissent autant dans le discours de l'enseignant que dans les diapositives.

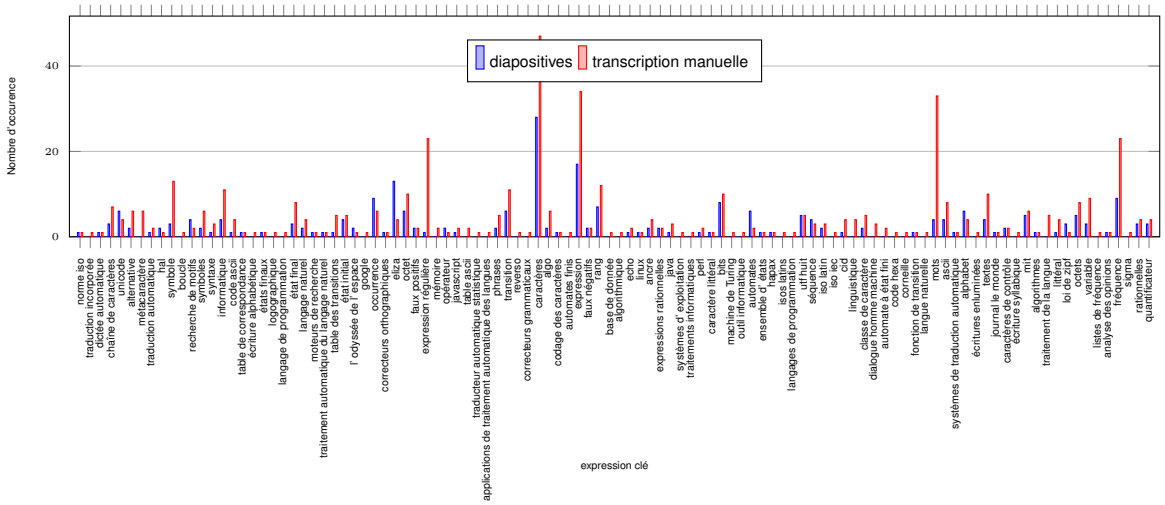


FIGURE 5.4 – Occurrence des expressions clés annotés à partir de la transcription manuelle pour le cours *"Langage naturel"*

5.5.2.2 Répartition des expressions clés dans les diapositives

Comme nous l'avons annoncé dans la section 5.3.2, une diapositive est considérée ici comme une unité textuelle "atomique", à l'échelle de laquelle notre corpus a été segmenté. Ainsi, la structure interne des diapositives est importante. Une donnée importante au sein

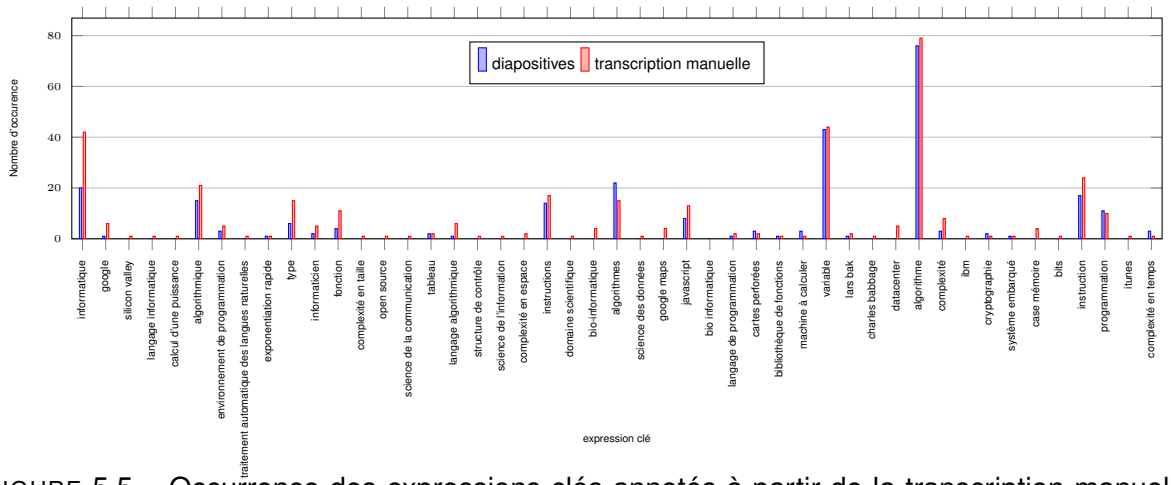


FIGURE 5.5 – Occurrence des expressions clés annotés à partir de la transcription manuelle pour le cours *"Introduction à l'informatique"*

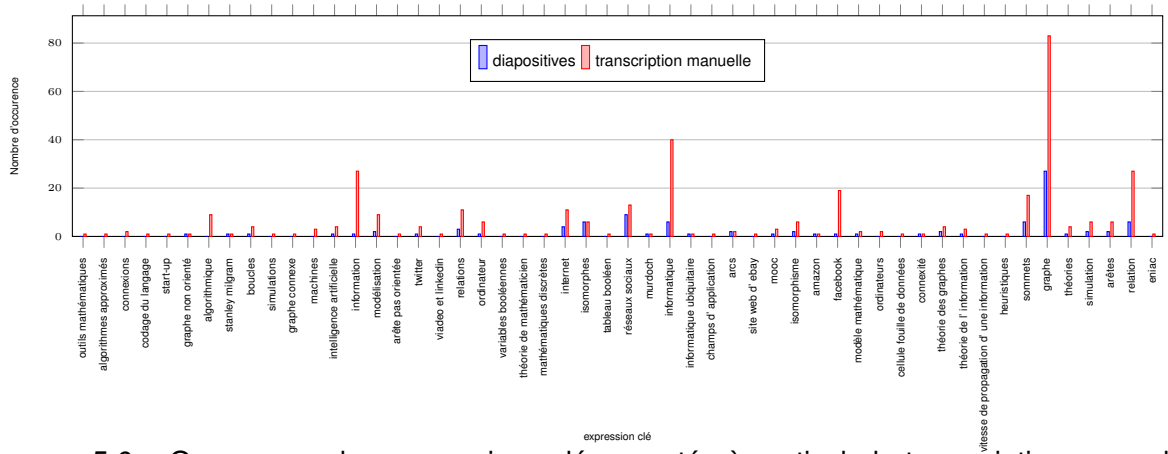


FIGURE 5.6 – Occurrence des expressions clés annotés à partir de la transcription manuelle pour le cours *"Réseaux sociaux et graphes"*

des diapositives est le nombre d'occurrences des expressions clés dans chaque diapositive, autrement dit la répartition des expressions clés du corpus sur ses diapositives. La répétition des expressions clés peut constituer un avantage pour les techniques de segmentation basées sur la cohésion lexicale. Nous poursuivons notre analyse en réalisant un focus sur la répartition des expressions clés par diapositive.

Les figures 5.12, 5.11 et 5.10 présentent, sur l'axe des abscisses, les expressions clés et, sur l'axe des ordonnées, les numèros des diapositives. Les points indiquent la présence d'une ou plusieurs expressions clés dans la diapositive.

On observe que nos données souffrent d'un manque de répétition des expressions clés dans les diapositives successives (la répétition d'une expression clé dans plusieurs diaposi-

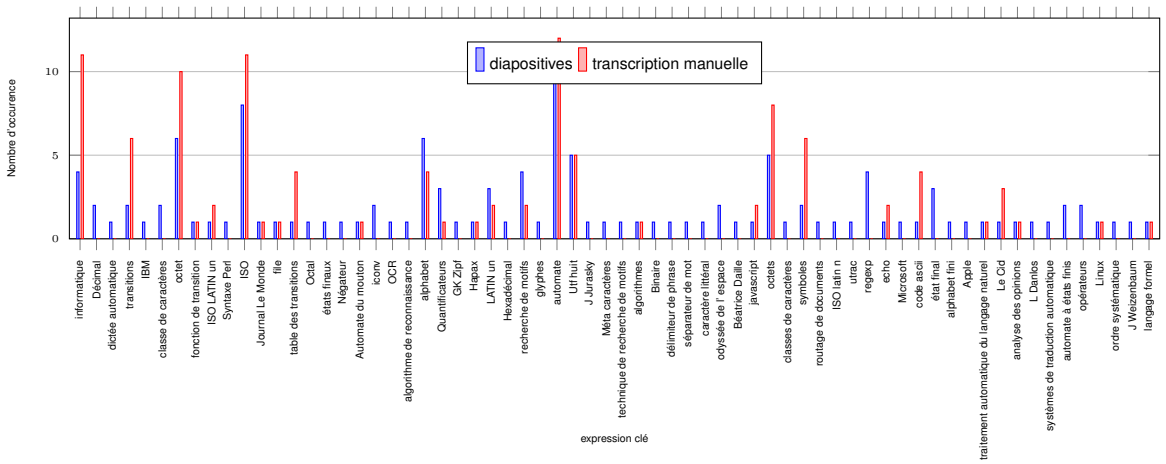


FIGURE 5.7 – Occurrence des expressions clés annotés à partir du support de cours pour le cours *"Langage naturel"*

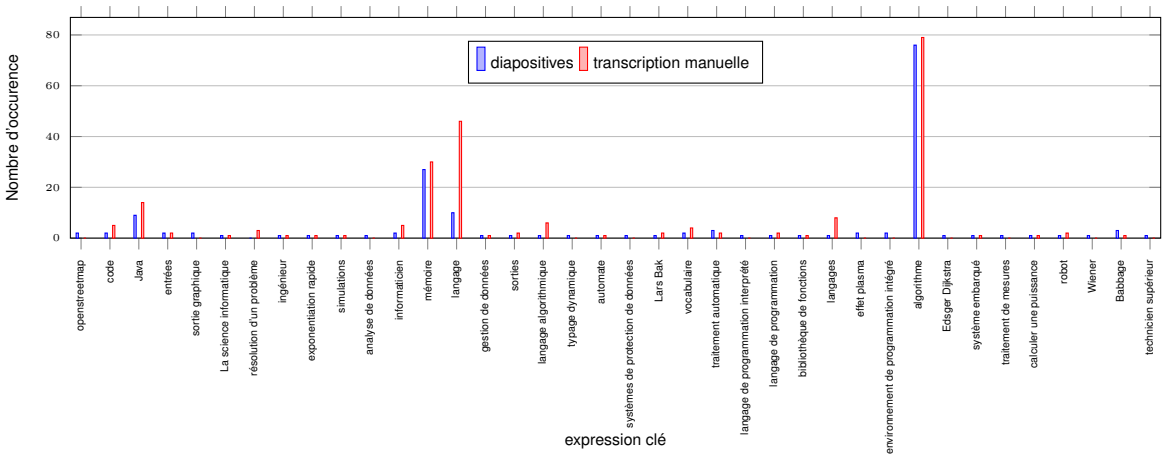


FIGURE 5.8 – Occurrence des expressions clés annotés à partir du support de cours pour le cours *"Introduction à l'informatique"*

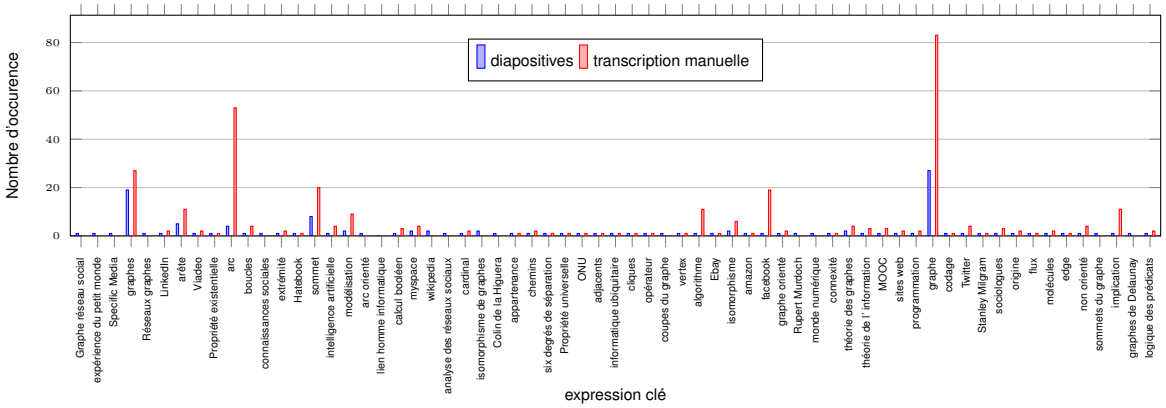


FIGURE 5.9 – Occurrence des expressions clés annotés à partir du support de cours pour le cours *"Réseaux sociaux et graphes"*

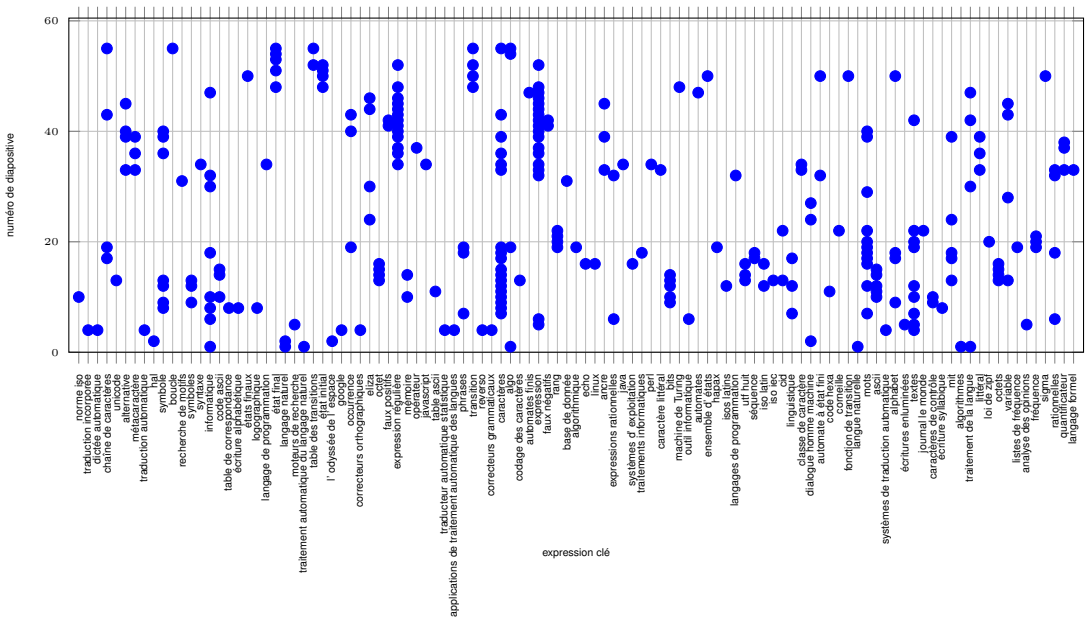


FIGURE 5.10 – Répartition des expressions clés annotés à partir de la transcription manuelle pour le cours *"Langage naturel"*

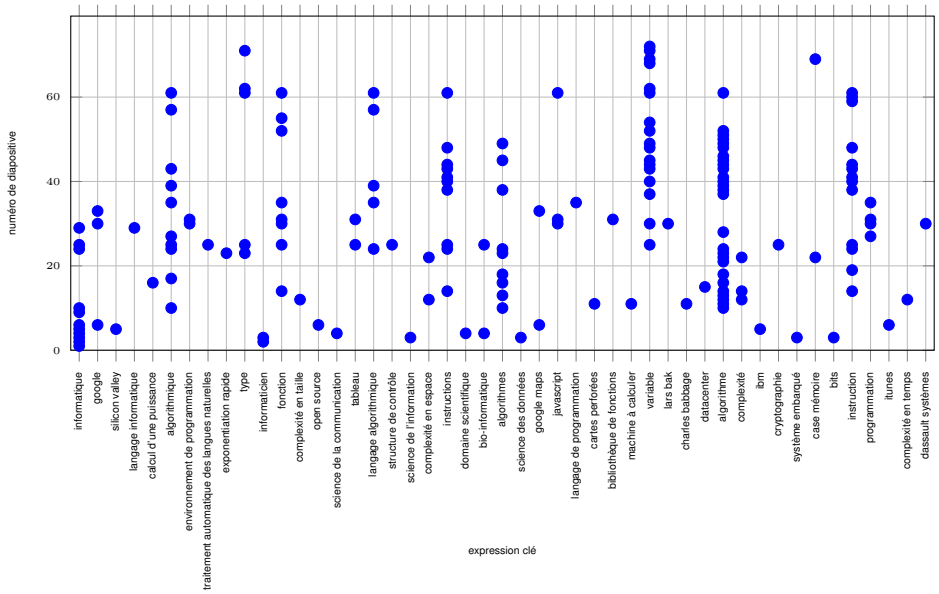


FIGURE 5.11 – Répartition des expressions clés annotés à partir de la transcription manuelle pour le cours *"Introduction à l'informatique"*

tives successives correspond à une barre verticale, dans les figures). Ce manque de répétition peut avoir un impact négatif sur la tâche de segmentation thématique. L'utilisation d'autres informations utiles pour la tâche de segmentation, hormis la transcription, devient alors une priorité

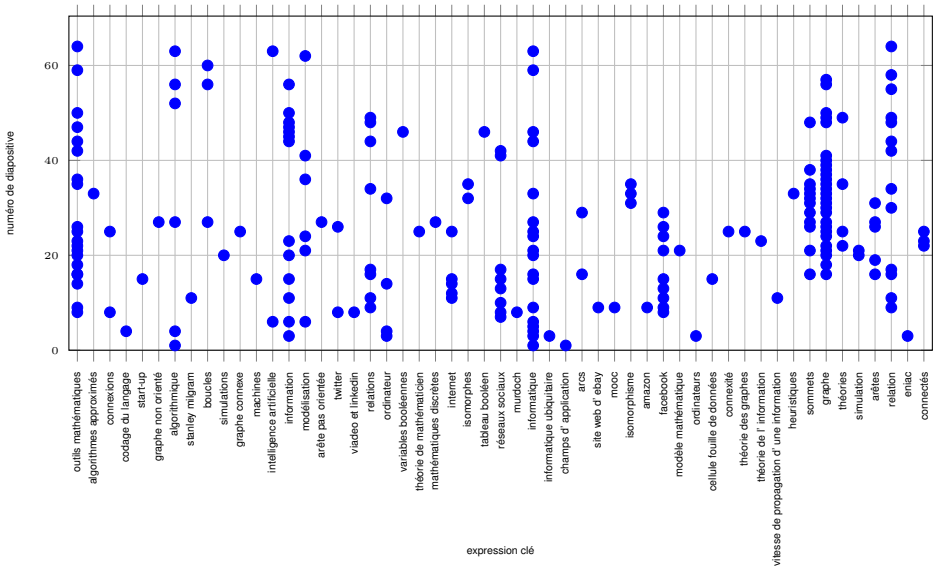


FIGURE 5.12 – Répartition des expressions clés annotés à partir de la transcription manuelle pour le cours *"Réseaux sociaux et graphes"*

à laquelle nous nous attachons dans le chapitre 8.

5.6 Système de base

L'état de l'art présenté précédemment dans ce manuscrit décrit plusieurs techniques de récupération de données d'adaptation et d'adaptation de modèles de langage dont plusieurs ont été mises en œuvre dans le contexte de cours magistraux. Parmi ces techniques, nous avons choisi d'adapter certaines d'entre elles à notre tâche.

5.6.1 Système de reconnaissance de la parole

Dans cette section, nous décrivons le système de reconnaissance de la parole utilisé pour les expériences menées dans cette thèse. Des expériences dans un cadre d'expérimentation réel (en direct dans un cours avec un vrai enseignant et de vrais étudiants) utilisant ce système de transcription ont été réalisées. Ces expériences seront décrites dans l'annexe A.

5.6.1.1 Modèles acoustiques

Les modèles acoustiques ont été appris sur environ 220 heures de données audio annotées. Les données d'apprentissage utilisées pour la modélisation acoustique sont représentées dans le tableau 5.5. Toutes les expériences ont été conduites en utilisant un modèle de type

chain-TDNN avec de l'apprentissage discriminant du type sMBR (POVEY et al., 2016). La quantité de données a été artificiellement multipliée par 3 en faisant varier la vitesse des données d'entraînement. Cette technique elle a montré des performances pour la reconnaissance vocale (KO et al., 2015).

TABLE 5.5 – Source des données d'apprentissage du modèle acoustique

	Durée
ESTER1 (GALLIANO et al., 2005)	85h35m
ESTER2 (GALLIANO, GRAVIER et CHAUBARD, 2009)	89h4m
REPERE (KAHN et al., 2012)	38h16m
VERA	5h44m
Total	≃219h

5.6.1.2 Vocabulaire

Le vocabulaire comprend environ 160K mots et correspond au vocabulaire qui a été utilisé par le LIUM pour la compagne d'évaluation REPERE (GALIBERT et KAHN, 2013), construit à partir de plusieurs sources, également utilisées pour l'apprentissage des ML. Le tableau 5.6 présente le nombre de mots de chaque source.

TABLE 5.6 – Source des données du modèle de langage

Source	**Taille**
Transcriptions manuelles des corpus d'apprentissage utilisés pour entraîner les modèles acoustiques	8M
Articles de sites de télévision	5M
Google News	204M
French Gigaword	1015M
Journaux	366M
Sous-titres de journal télévisé	11M
Total	160K

5.6.1.3 Modèles de langage

Le modèle de langage du système de base a été construit par interpolation linéaire de plusieurs modèles dont les sources (des articles du journal LeMonde, des Google News, un corpus Gigaword, des sous-titres de journaux, etc.) (ROUSSEAU et al., 2014) sont présentées dans le tableau 5.6. Le vocabulaire est composé de 160K mots 5.6.1.2.

5.6.1.4 Système de segmentation et de regroupement en locuteurs

Le système de segmentation et regroupement en locuteurs (SRL) utilisé a été développé au LIUM : *LIUM_SpkDiarization* (MEIGNIER et MERLIN, 2010). Ce système a été développé pour la campagne d'évaluation française ESTER 2 (GALLIANO, GRAVIER et CHAUBARD, 2009), où il a obtenu les meilleurs résultats dans la tâche de SRL (10.8% DER (Diarization Error Rate)) sur des émissions journalistiques. Le système repose sur une segmentation acoustique de type GLR (Generalized Likelihood Radio) (SIU, YU et GISH, 1992) suivi d'une classification hiérarchique utilisant le critère BIC (Bayesian Information Criterion) comme mesure de similarité entre les locuteurs (GISH, SIU et ROHLICEK, 1991). Il se compose de quatre principales étapes :

1. Le signal est découpé en petits segments acoustiquement homogènes,
2. Les segments sont ensuite regroupés en clusters de locuteurs sans changer les frontières,
3. Les frontières sont ajustées en utilisant un décodage avec l'algorithme de Viterbi (VITERBI, 1967),
4. Les régions correspondant à des zones de non-parole (musique, bruit, et silence) sont éliminées en utilisant un nouveau décodage Viterbi.

5.6.1.5 Données de test

Les données de test sont issues du corpus français PASTEL. Ces données sont décrites en détail dans le chapitre 5. Puisque la méthode d'adaptation proposée repose sur l'utilisation des supports de cours, nous avons évalué nos performances en utilisant les six cours dont on dispose de diapositives (les cours dont la source est COCo (tableau 5.1)).

5.6.2 Performance du système préliminaire

Le tableau 5.7 présente le résultat du système générique en utilisant la métrique WER.

	SRAP sans adaptation
Taux d'erreurs sur les mots (WER)	19,46

TABLE 5.7 – Résultat du système préliminaire en WER

5.6.3 Adaptation du modèle de langage

Comme nous l'avons vu précédemment au chapitre 1, les enjeux concernant l'adaptation des modèles de langage sont nombreux. Nous cherchons à proposer une adaptation des modèles de langage qui réponde aux spécificités et caractéristiques des cours magistraux. Le but de la procédure d'adaptation que nous proposons est d'adapter le modèle de langage et d'enrichir le vocabulaire générique d'un système de reconnaissance de la parole initial pour chaque nouveau cours à transcrire, en vue d'en fournir la meilleure transcription automatique possible. Cette section donne une vue globale du processus d'adaptation que nous avons retenu pour mener l'adaptation linguistique.

5.6.3.1 Extraction des requêtes

Dans le cadre de l'adaptation des modèles de langage dans un contexte de cours magistraux, l'exploitation des textes contenus dans le support de cours a montré son utilité dans beaucoup d'études (YAMAZAKI et al., 2007 ; KAWAHARA, NEMOTO et AKITA, 2008 ; MIRANDA, NETO et BLACK, 2013 ; AKITA, TONG et KAWAHARA, 2015).

Sur la base de ces travaux, nous utilisons les diapositives des cours pour extraire des données pertinentes par rapport à chaque cours. Généralement, une extraction de mots clés s'effectue à partir du support de cours en utilisant des techniques de recherche d'information s'appuyant sur les valeurs TF-IDF (section 8.2.2). Or, il a été démontré dans l'analyse du corpus dans la section 5.5.2 que certaines expressions clés souffrent d'un manque de répétitions.

Une information importante qui se trouve dans les cours correspond aux titres des diapositives. En effet, c'est souvent l'information principale sur laquelle un lecteur se base pour rechercher et se positionner dans une partie du cours. L'idée est alors d'utiliser les titres des diapositives comme requêtes pour récupérer des documents web.

Les titres de diapositives sont parfois trop génériques et peuvent désigner des thématiques qui n'ont aucun rapport avec le cours à traiter. Prenons l'exemple d'une recherche web avec le titre "les variables", les résultats de la recherche peuvent contenir des variables bancaires, des variables informatiques, des variables statistiques, etc. Or notre but est que les résultats soient précis pour retourner des pages pertinentes, ici traitant des variables informatiques.

Afin de résoudre ce problème, chaque titre de diapositive sera concaténé avec le titre du cours de manière à consolider le fait que les requêtes soient liées au contexte du cours.

5.6.3.2 Recueil des données pertinentes

Comme nous l'avons souligné dans la section 2.2.2, il existe plusieurs manières pour collecter des données d'adaptation spécifiques à la tâche. Dans le cadre de cours magistraux où chaque cours comporte un sujet spécifique, les documents thématiques à récupérer doivent

venir d'une ressource textuelle couvrant le plus grand nombre de sujets possibles pour assurer la réussite de l'adaptation thématique dans un maximum de cas.

Pour cela, l'approche retenue repose sur le Web qui correspond à une ressource linguistique ouverte, facilement accessible grâce aux moteurs de recherche.

Le moteur de recherche "Google" a été utilisé pour exécuter les requêtes telles que formulées dans la section 5.6.3.1. Les pages pointées par les liens renvoyés sont téléchargées. La recherche web a été limitée à 100 pages web par requête. Le contenu textuel principal de ces pages a été extrait et nettoyé.

5.6.3.3 Vocabulaire

Le vocabulaire générique présenté à la section 5.6.1.2 a été enrichi avec les 30K mots les plus fréquents extraits des données du domaine récupérées du web. Les phonétisations des mots ajoutés sont obtenues à l'aide de l'outil LIA-PHON (BÉCHET, 2001).

5.6.3.4 Adaptation du ML

Les données collectées du web (section 5.6.3.2) servent à construire un modèle de langage du domaine. Le vocabulaire de ce modèle est celui décrit dans la section précédente.

Le modèle de langage adapté est calculé par l'interpolation linéaire de notre modèle de langage généraliste présenté dans la section 5.6.1.3 avec un modèle de langage d'adaptation appris sur le corpus récupéré du web. L'interpolation passe par l'optimisation d'un coefficient d'interpolation λ en utilisant un corpus de développement.
Dans le cadre de cette thèse, on ne dispose pas d'un corpus volumineux pour le segmenter simplement en deux parties : corpus de test et corpus de développement. Pour fixer cette valeur de λ :

1. nous avons appliqué la validation croisée (*leave-one-out cross-validation*). Pour un cours donné, la valeur de λ est égale à la moyenne des λ optimisé en perplexité sur tous les autres cours sauf sur le cours traité. On a appliqué cette opération pour tous les cours. Les valeurs de λ obtenues sont tous entre $0,65$ et $0,75$.
2. nous avons calculé les valeurs de λ en l'optimisant sur les transcriptions générées par le modèle de langage générique du cours visé.

Nous avons également calculé les valeurs de λ en l'optimisant directement sur les transcriptions manuelles du cours visé. Les valeurs obtenues ne sont pas tout à fait les mêmes que celles obtenues par la validation croisée ou par les transcriptions générées par le modèle générique du cours visé. Vu la difficulté de calculer la valeur exacte de λ, nous avons utilisé un coefficient constant $\lambda = 0,7$ qui semble un compromis à travers les valeurs obtenues de ces

expériences. Dans un contexte applicatif similaire, (LECORVÉ, 2010) ont obtenu des résultats optimaux avec un coefficient situé entre 0,7 et 0,8 pour le modèle générique. En utilisant ce coefficient constant de λ, nous générons alors les transcriptions en décodant de nouveau le graphe en utilisant le modèle de langage avec adaptation.

5.6.3.5 Résultats de l'adaptation du ML

Le tableau 5.8 présente les résultats d'adaptation du modèle de langage en utilisant la métrique WER (section 1.4). Nous remarquons une réduction relative de 15,62% en WER (de 19,46% à 16,42%) avec un système adapté au domaine.

TABLE 5.8 – Résultats d'adaptation du modèle de langage en WER

	SRAP sans adaptation	**SRAP avec adaptation**
Taux d'erreurs sur les mots (WER)	19,46	16,42

5.7 Conclusion

Dans ce chapitre, nous avons présenté dans un premier temps le corpus sur lequel sont réalisées toutes les expériences décrites dans ce manuscrit. Ce corpus a été créé dans le cadre du projet PASTEL. Ce corpus se compose de transcriptions manuelles du discours d'enseignement en situation de cours magistraux, de segmentations manuelles en deux granularités et de l'annotation d'expressions clés. Il comprend plus de neuf heures de parole. Ce corpus va aider au développement et à l'expérimentation d'une application dans un cadre pédagogique, va permettre l'évaluation des systèmes développés et des approches proposées, et va apporter à la communauté de recherche en TALN un corpus dédié au domaine éducatif. Les données et les annotations seront distribuées, à la communauté, sous licence libre. Dans un deuxième temps, nous avons présenté le système de reconnaissance de parole et proposé une technique d'adaptation du modèle de langage. L'approche consiste à extraire les titres des diapositives utilisées comme support de cours. Une requête sur un moteur de recherche est alors construite à partir de ces titres afin de collecter des données du domaine. La partie expérimentale a montré une amélioration en terme de WER par rapport à un SRAP générique.

Chapitre 6

Nouvelles métriques pour l'évaluation qualitative d'un système de reconnaissance automatique de la parole

Contents

6.1 Introduction

Dans le chapitre précédent, la performance du système de reconnaissance automatique de la parole (SRAP) a été évaluée de manière classique, à l'aide du taux d'erreurs sur les mots. Cependant, cette mesure fournit un score global à partir duquel on ne peut pas interpréter la performance pour les mots du domaine considéré, comme il apparaît difficile d'estimer la performance pour certaines tâches exploitant la transcription.

Considérant que le taux d'erreurs sur les mots (WER) n'est pas suffisamment pertinent pour comparer la performance du SRAP et que ces erreurs peuvent avoir une forte incidence sur la précision de plusieurs tâches de traitement automatique du langage naturel, nous explorons, dans ce chapitre, l'utilisation de trois nouvelles mesures d'évaluation plus pertinentes pour comparer l'apport de l'adaptation du modèle de langage pour un SRAP.

Ce chapitre est organisé comme suit. Nous exposons dans un premier temps les limites des métriques d'évaluation existantes dans la littérature. Puis, les deux sections suivantes présentent nos propositions d'évaluation ainsi que les résultats sur des données réelles.

6.2 Les mesures d'évaluation des SRAP : limites et motivations

Comme nous avons vu dans la section 1.4, le WER mesure directement la qualité d'un système de transcription de la parole, en comptant le nombre d'erreurs entre la sortie de ce système et la transcription de référence fournie par un expert humain. Un inconvénient majeur du WER dans l'évaluation de l'adaptation du ML, dont le but est d'adapter le système à un domaine particulier, est que tous les mots de la transcription (mots du domaine / mots génériques) partagent la même importance pendant l'évaluation et sont évalués équitablement.

Pour pallier ce problème, certains travaux tels que (PARK, HAZEN et GLASS, 2005 ; YAMAZAKI et al., 2007) ont utilisé les métriques d'évaluation de recherche d'informations standard (Précision, Rappel, F-mesure) pour évaluer la performance pour un ensemble de mots défini (les mots du domaine dans le cas d'adaptation du ML). Or, il a été démontré dans l'étude de (MAKHOUL et al., 1999) que la combinaison de la précision et du rappel avec la moyenne harmonique diminue l'importance des erreurs de suppression et d'insertion par rapport aux erreurs de substitution.

Non seulement, le WER ne permet pas de différencier les mots du nouveau domaine des mots génériques, mais il ne permet pas non plus de prendre en compte la performance pour la tâche visée. En effet, les SRAP sont souvent conçus comme une brique dans d'autres applications de traitement automatique de langage naturel qui utilisent les transcriptions de sortie pour effectuer d'autres tâches. Les transcriptions générées constituent alors une ressource précieuse pour d'autres modules technologiques appliquant des traitements tels que la recherche d'informations, la traduction, l'indexation de documents...

Dans la littérature, plusieurs travaux ont montré que le WER n'a pas toujours une bonne corrélation avec le score d'une tâche plus complexe utilisant des transcriptions automatiques. Les auteurs de (HE, DENG et ACERO, 2011) observent une non corrélation entre les performances de la traduction automatique et le WER. Une autre étude montre cette non corrélation pour la tâche de compréhension de la parole (WANG, ACERO et CHELBA, 2003). De même pour la détection d'entités nommées (BEN JANNET et al., 2015 ; BEN JANNET, 2015) où les

auteurs ont proposé une autre métrique d'évaluation : ETER (BEN JANNET et al., 2014). Cette non corrélation n'est pas du tout observée dans d'autres travaux tels que (MUNTEANU et al., 2006) pour la tâche de question-réponse, (SANDERS, LE et GAROFOLO, 2002) pour la traduction automatique et (MUNTEANU et al., 2006) pour la tâche de recherche d'informations. Selon ces études, il est important de prendre en compte dans l'évaluation le contexte applicatif dans lequel les SRAP sont utilisés.

6.3 Évaluation intrinsèque

L'évaluation intrinsèque permet d'évaluer la performance de la transcription indépendamment de l'application qui utilise la transcription. L'évaluation intrinsèque proposée dans cette thèse se focalise sur une mesure existante : l'individual word error rate. Dans un premier temps, nous présentons dans cette section la métrique individual word error rate. Nous présentons dans un second temps les modifications apportées à cette métrique pour qu'elle soit adaptée à l'évaluation de l'adaptation des ML dans le contexte de transcription de documents multi-domaines.

6.3.1 Individual word error rate

Dans le but de déterminer quelles erreurs sont difficiles à transcrire par le SRAP, les auteurs dans (GOLDWATER, JURAFSKY et MANNING, 2008 ; GOLDWATER, JURAFSKY et MANNING, 2010) ont proposé une nouvelle métrique d'évaluation : le IWER (Individual Word Error Rate). La spécificité de cette métrique est sa capacité à calculer l'erreur pour un ensemble de mots. Cette métrique s'inspire du principe de WER. Pour les erreurs de suppression et de substitution, le principe est le même que WER : on attribue une valeur binaire 1 ou 0 en comparant l'hypothèse et la référence pour chaque mot. Cependant, pour les erreurs d'insertion, il peut y avoir deux mots de référence adjacents qui pourraient être responsables, et comme nous n'avons aucun moyen de savoir quel mot est responsable, une pénalité partielle sera attribuée. Cette pénalité est égale au nombre de toutes les erreurs d'insertion sur le nombre total des erreurs adjacentes. Donc, pour un mot, l'IWER est calculé comme suit :

$$IWER(w_i) = del_i + sub_i + \alpha.ins_i \tag{6.1}$$

avec $del_i = 1$ si w_i a été supprimé (0 sinon), $sub_i = 1$ si w_i a été modifié (0 sinon) et ins_i correspond aux nombres d'insertions adjacents au mot w_i. Le paramètre α est calculé comme suit :

$$\alpha = \frac{I}{\sum_{w_i} ins_i} \tag{6.2}$$

où I est le nombre d'insertions dans tout le corpus (la pénalité totale pour les erreurs d'insertion est la même que pour le calcul de WER).

Le IWER pour un ensemble de mots correspond à la moyenne de l'IWER pour des mots individuels :

$$IWER(w_1...w_n) = \frac{1}{n}\sum_{i=1}^{n} IWER(w_i) \tag{6.3}$$

Nous proposons d'utiliser le IWER pour évaluer l'adaptation des ML en considérant les mots du domaine comme notre ensemble de mots à évaluer. Dans l'étude de (GOLDWATER, JURAFSKY et MANNING, 2010), l'ensemble de mots ont été sélectionnés en se basant sur des caractéristiques définies. Ces caractéristiques sont présentés dans la figure 6.1. Ces caractéristiques sont génériques à n'importe quelle transcription. Dans le cas de la transcription de cours magistraux, chaque cours magistral possède son propre lexique du domaine qui est diffèrent d'un cours à un autre. Par conséquent, nous proposons d'étendre la métrique afin de permettre d'obtenir un score global sur tout un corpus de transcriptions de cours qui ne partage pas les mêmes mots du domaine et pas seulement sur une seule transcription. Cette proposition sera décrite dans la section suivante.

6.3.2 $\mathbf{IWER_{Average}}$

La métrique IWER offre un moyen pour mesurer la performance de reconnaissance pour une caractéristique donnée sur un document. Nous proposons un nouveau cas d'utilisation de la métrique IWER en considérant les mots dans le domaine comme caractéristique et nous l'étendons à une version plus générale, appelée le $IWER_{Average}$, pour obtenir un score global sur un corpus de transcriptions multi-domaines. Le $IWER_{Average}$ (MDHAFFAR et al., 2019b) est défini par la formule suivante :

$$IWER_{Average} = \frac{1}{\sum_{j=1}^{m} n_j}\sum_{j=1}^{m}\sum_{i=1}^{n_j} IWER(w_i) \tag{6.4}$$

avec m est le nombre de transcriptions de cours magistraux et n_y correspond au nombre de mots de domaine dans la transcription du cours y.

6.3.3 Résultats expérimentaux

Nous utilisons le $IWER_{Average}$ pour évaluer l'adaptation présentée dans le section 5.6.3. Les résultats expérimentaux sont résumés dans le tableau 6.1. Quatre ensembles de caractéristiques ont été utilisés. La première ligne considère tous les mots du lexique et le résultat avec cet ensemble correspond au résultat WER (suite à la normalisation faite par le paramètre α). La ligne 2 et la ligne 3 présentent les résultats sur des ensembles de mots extraits à partir des

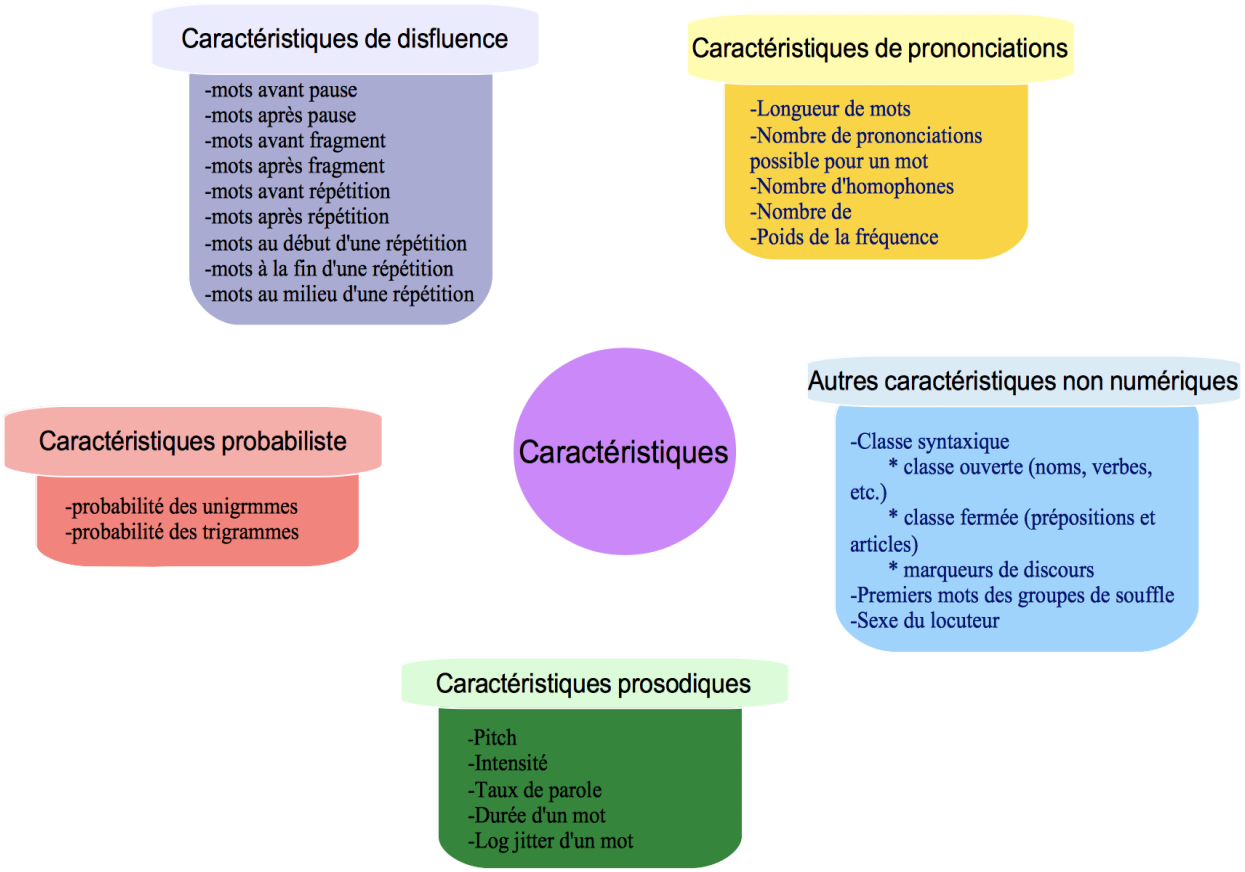


FIGURE 6.1 – Caractéristiques des mots utilisés par (GOLDWATER, JURAFSKY et MANNING, 2010) pour comparer la performance entre deux systèmes de reconnaissance de la parole

diapositives de cours (les titres de diapositives et les mots clés annotés par l'expert à partir des diapositives (section 5.3.3)). La dernière ligne présente le résultat pour les mots clés annotés par l'expert à partir des transcriptions manuelles (section 5.3.3).

TABLE 6.1 – (%) $IWER_{Average}$ scores pour quatre caractéristiques : tous les mots (=WER), mots clés annotés à partir de la transcription, mots clés annotés à partir des diapositives et les titres de diapositives

	SRAP sans adaptation	**SRAP avec adaptation**
Tous les mots (= WER)	19,46	16,42
Titre des diapositives Mots clés annotés à partir des diapositives	29,52 32,31	14,05 14,52
Mots clés annotés à partir de la transcription manuelle	31	17,30

TABLE 6.2 – (%) Taux de mots hors vocabulaire avec et sans enrichissement du vocabulaire

	SRAP sans adaptation	**SRAP avec adaptation**
OOV	0,86	0,11

Les résultats montrent que les mots du domaine sont mal transcrits par le SRAP (31% pour les mots clés annotés à partir de la transcription manuelle). On observe que si un modèle de langage adapté permet de réduire le WER global relatif de 15,62% (de 19,46% à 16,42%), nous montrons que cette réduction atteint 44,2% lorsque elle est calculée uniquement sur les mots pertinents ($IWER_{Average}$ calculé sur des mots-clés manuels extraits de transcriptions manuelles : de 31% à 17,30%).

Comme nous avons présenté dans la section 2.2, l'adaptation du ML comprend l'adaptation des probabilités des n-grammes d'un modèle et/ou l'enrichissement du vocabulaire. L'adaptation présentée dans le tableau 6.1 correspond à l'adaptation des probabilités des n-grammes avec enrichissement vocabulaire (section 5.6.3.3). Le tableau 6.2 présente le taux de mots hors vocabulaire sans et avec adaptation. Nous répétons la même évaluation en n'adaptant que les probabilités d'un modèle n-gramme. Les résultats sont présentés dans le tableau 6.3. Ils montrent que l'amélioration relative en WER est de 1,91% (de 16,74% à 16,42%) alors que l'amélioration relative en $IWER_{Average}$ atteint 19,42% (de 21,47% à 17,30%). Cela montre que la métrique $IWER_{Average}$ est capable de mieux exprimer le gain apporté par l'intégration de mots appartenant à un domaine dans le vocabulaire du SRAP qu'en utilisant le WER.

TABLE 6.3 – (%) $IWER_{Average}$ scores pour l'adaptation du ML avec et sans enrichissement du vocabulaire.

	vocabulaire générique	**vocabulaire enrichi**
Tous les mots (= WER)	16,74	16,42
Mots clés annotés à partir de la transcription manuelle	21,47	17,30

Pour conclure, si on considère que toutes les erreurs ne partagent pas la même importance et que les mots du domaine sont les mots les plus importants d'un cours, le $IWER_{Average}$ exprime mieux le gain apporté par l'adaptation du ML que le WER.

6.4 Évaluation extrinsèque

$IWER_{Average}$ permet d'appréhender autrement la qualité intrinsèque d'un SRAP, comparativement à la métrique classique WER. Il est important de connaître non seulement la précision d'un SRAP, mais également l'impact des erreurs sur d'autres tâches. C'est l'objectif d'une

évaluation extrinsèque, où le système sera évalué sur les tâches s'appuyant sur des transcriptions automatiques. Cette section présente deux métriques pour mesurer la qualité extrinsèque d'une transcription (MDHAFFAR et al., 2019a).

6.4.1 Évaluation de la recherche d'information

L'une des tâches qui peuvent être utiles pour des applications pédagogiques est l'enrichissement automatique (ou sous forme de recommandation) des transcriptions avec des ressources externes. L'objectif est d'offrir aux étudiants des liens externes utiles qui peuvent servir pour réviser ou avoir plus d'explications détaillées concernant les concepts du cours. Dans ce cas, il est important d'évaluer l'impact de la transcription sur une tâche de récupération de documents (SENAY, 2011).

L'évaluation extrinsèque d'une transcription dans la tâche de recherche d'information a été l'objet de quelques études de recherche (CHELBA, HAZEN et SARACLAR, 2008). Généralement, cette évaluation s'effectue en comparant les résultats fournis par le système automatique de recherche d'informations et une liste de référence qui contient un classement fourni par des experts. La performance du système est mesurée par une moyenne arithmétique de précision (MAP - mean average precision) ou des précision et rappel au rang K (Précision@K et Rappel@K).

Dans notre cas, on ne dispose pas d'une référence annotée. Par conséquent, nous allons proposer un autre protocole d'évaluation pour la tâche de recherche d'informations. La section suivante présente le calcul de cette métrique.

6.4.1.1 Méthodologie

Notre évaluation consiste à comparer les résultats de requêtes de recherche pour chaque segment de *"Granularité 1"*. Les requêtes sont construites en utilisant l'approche TF-IDF sur les transcriptions de chacun de ces segments. Ces requêtes sont soumises à un moteur de recherche (dans notre étude, Google). Notre but est de déterminer la pertinence des documents récupérés. Nous avons défini comme documents pertinents (référence) les documents extraits de requêtes constitués à partir de transcriptions manuelles. Sur la base de cette référence, une comparaison avec les documents récupérés à partir de requêtes construites sur des transcriptions automatique, en calculant un taux de couverture. Le taux de couverture se calcule comme suit :

$$Taux\ de\ couverture = \frac{\#(A \cap B)}{\#B} \tag{6.5}$$

avec A correspond aux documents récupérés à partir de requêtes construites des transcriptions automatiques et B correspond aux documents récupérés à partir de requêtes construites des

transcriptions manuelles. Sachant que $\#A$=$\#B$.

La figure 6.4 illustre la méthode décrite dans cette section. On va appliquer cette méthode pour mesurer, dans le cadre de ce travail, l'apport de l'adaptation des ML pour la recherche d'information.

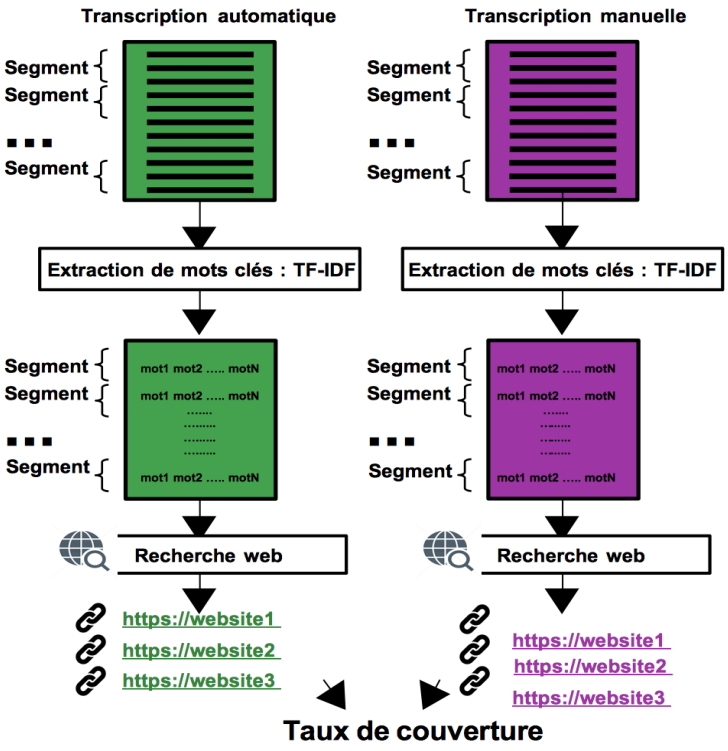


FIGURE 6.2 – Méthode d'évaluation de la recherche d'information

6.4.1.2 Résultats

La figure 6.3 présente le taux de couverture des documents récupérés à partir de requêtes construites sur des transcriptions automatiques, avec (lignes continues) ou sans (pointillés) adaptation des modèles de langage, par rapport aux documents récupérés à partir de requêtes construites sur des transcriptions manuelles. Le taux de couverture est calculé en fonction du nombre de documents visés (de 1 à 20). Nous avons également expérimenté différentes types de requêtes composées de 1 à 5 mots (k dans la figure 6.3) extraits par TF-IDF. On présente, dans la figure 6.3, les résultats pour des requêtes composées de 1 ou 5 mots. Pour 2, 3 et 4 mots, le même comportement qu'avec 1 et 5 mots a été observé.

Les résultats montrent que la transcription avec adaptation surpasse la transcription sans adaptation en termes de récupération des ressources pertinentes, pour toutes les tailles de requêtes.

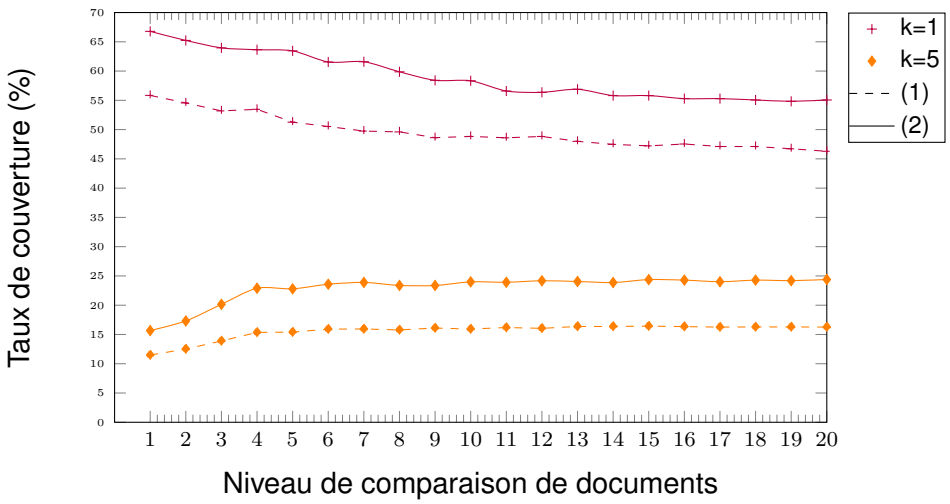


FIGURE 6.3 – Tâche de recherche d'information : comparaison du taux de couverture entre les requêtes construites à partir de segments de transcriptions manuelles et de transcriptions automatiques (1) sans et (2) avec adaptation des modèles de langage.

6.4.2 Évaluation d'indexabilité

La seconde évaluation extrinsèque consiste à évaluer l'indexabilité des transcriptions. En d'autres termes, nous souhaitons déterminer si la qualité des transcriptions joue un rôle dans l'indexation et la récupération des transcriptions, qui sont utiles pour naviguer dans la vidéo du cours et pour atteindre rapidement ce que cherche l'apprenant. De manière identique à l'évaluation de la tâche de recherche d'information, l'évaluation d'indexabilité a été largement étudiée dans la littérature. (SENAY, 2011 ; SENAY, LECOUTEUX et LINARÈS, 2012) ont proposé une mesure d'indexabilité pour la transcription qui consiste, dans un premier temps, à segmenter la transcription en groupes de souffle. Par la suite, l'indexabilité est calculée pour chaque segment. Pour chacune des requêtes du jeu de requêtes, une recherche est effectuée sur l'ensemble de la base de données en utilisant tous les segments transcrits manuellement, excepté pour le segment s dont on utilise la transcription automatique (isolation du segment). Ensuite, les rangs des résultats des recherches sont comparés avec ceux obtenus sur l'ensemble du corpus de référence. L'indexabilité d'un segment s est obtenue en calculant la F-mesure sur les 20 meilleurs rangs où apparaît le segment, relativement aux 20 meilleurs rangs de référence. Cette évaluation permet d'estimer l'indexabilité en ne retirant, à chaque fois, qu'un seul segment. Nous proposons une métrique d'évaluation de l'indexabilité qui se base sur la totalité de la transcription automatique. La section suivante présente le principe de cette métrique.

6.4.2.1 Méthodologie

Les segments de type *"Granularité 1"* ont été indexés en utilisant le moteur de recherche lemur [1]. Lemur permet d'indexer des documents et de faire une recherche en classant les documents correspondants à une certaine requête selon des scores. Trois ensembles de segments ont été considérés : ceux des transcriptions manuelles, ceux de la transcription automatique sans adaptation et ceux de la transcription automatique avec adaptation. Des requêtes vont être soumises au moteur de recherche (lemur) pour récupérer les segments pertinents à partir de chaque ensemble distinct de segments. Les requêtes utilisées sont les mots des titres des diapositives et les mots clés annotés à partir de la transcription manuelle. Chaque requête renvoie une liste ordonnée de segments (résultat de la recherche). Pour estimer la qualité de l'indexabilité, nous avons utilisé le coefficient de Spearman qui mesure la corrélation de rang entre les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription manuelle et les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription automatique (sans et avec adaptation). La figure 6.4 illustre la méthode décrite dans cette section.

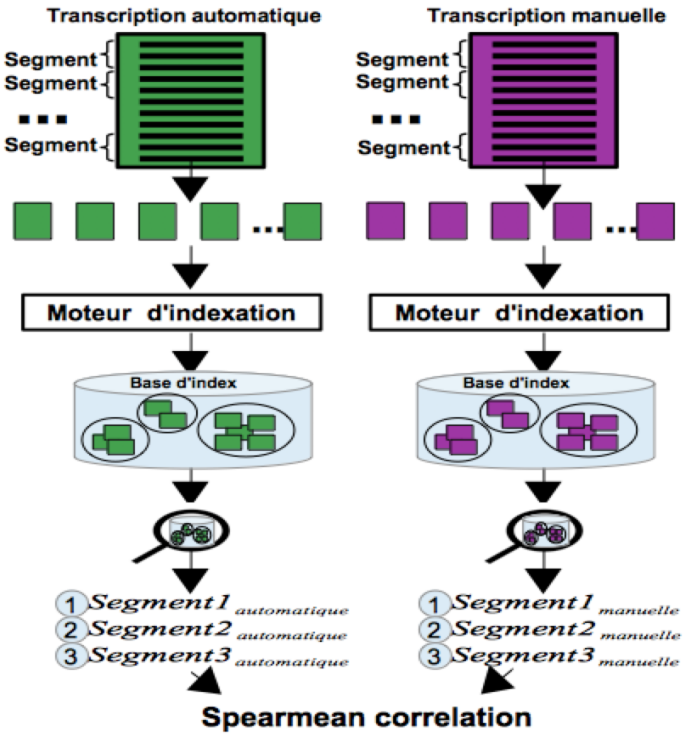


FIGURE 6.4 – Méthode d'évaluation de la tâche d'indéxabilité

1. https ://www.lemurproject.org/

6.4.2.2 Résultats

Le tableau 6.4 présente les scores de corrélation moyens pour l'ensemble du corpus en utilisant deux ensembles de jeux de requête : les mots des titres des diapositives, et les mots-clés de la transcription. Le coefficient de Spearman varie entre -1 et $+1$. Une valeur proche de 1 indique une forte corrélation entre les deux listes de documents renvoyés par la recherche alors qu'une valeur proche de 0 indique une faible corrélation (-1 indique une forte corrélation mais dans un sens opposé). Le tableau 6.4 présente le score moyen de corrélation pour l'ensemble du corpus. Ici également, les résultats indiquent une meilleure indexabilité obtenue après adaptation du modèle de langage du SRAP, pour tous les jeux de requêtes.

TABLE 6.4 – Évaluation de l'indexabilité des transcriptions : comparaison des résultats d'extraction avec le coefficient de corrélation de rang de Spearman, en utilisant différents jeux de requêtes

Jeux de requêtes	**SRAP sans adaptation**	**SRAP avec adaptation**
Les mots des titres des diapositives	0,458	0,588
Les mots clés annotés à partir de la transcription manuelle	0,288	0,516

6.4.3 Discussion

Comme nous l'avons vu dans notre cadre expérimental, l'adaptation automatique de modèles de langage pour la reconnaissance de la parole permet de réaliser environ trois erreurs de moins pour cent mots transcrits (WER passant de 19,46% à 16,42%), ce qui correspond à une réduction du WER d'environ 15,6%. Ces valeurs, bien qu'intéressantes, ne mettent pas en avant certains phénomènes très intéressants liés aux tâches finales pour lesquelles les transcriptions automatiques sont générées.

En termes de recherche d'informations, par exemple, nous constatons une augmentation du taux de couverture des documents retrouvés (par rapport aux documents qui auraient été trouvés à partir de requêtes extraites de transcriptions manuelles) qui peut dépasser 28,5% (k=1, niveau 1, taux de couverture passant de 56% à 67%). Enfin, en termes d'indexabilité, nous montrons que, dans cette étude, le taux de corrélation de Spearman (par rapport à l'indexation obtenue par des transcriptions manuelles) peut augmenter de plus de 79% (de 0,288 à 0,516) pour les termes les plus importants du document grâce à l'adaptation des modèles de langage.

Les résultats présentés montrent que le WER ne permet pas de bien mesurer les performances pour les tâches de recherche d'informations et d'indexation considérées, puisqu'il

masque les gains réels apportés par l'adaptation des modèles de langage sur les tâches visées : cela prouve la nécessité d'utiliser de nouvelles mesures, telles que celles présentées, pour évaluer l'apport réel de l'adaptation des modèles de langage.

6.5 Conclusion

Le taux d'erreurs sur les mots WER n'est pas toujours la meilleure mesure à utiliser pour évaluer les systèmes de reconnaissance de la parole. Une meilleure compréhension de l'impact des erreurs de transcription automatique implique nécessairement le développement de meilleures métriques pour l'évaluation. Dans ce chapitre, nous avons présenté une évaluation qualitative de la qualité de reconnaissance de parole. Nous avons mis en place un protocole d'évaluation intrinsèque qui permet d'évaluer la performance des mots du domaine ($IWER_{Average}$) ainsi qu'un protocole d'évaluation extrinsèque qui permet d'évaluer la qualité de recherche d'informations (le taux de couverture) et l'indexabilité (le coefficient de Spearman). Les résultats obtenus avec les trois différentes évaluations montrent que le taux d'erreur sur les mots est une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

Chapitre 7

Étude diachronique de l'adaptation du modèle de langage

Contents

7.1 Introduction

L'adaptation des modèles de langage a montré son efficacité sur les trois tâches sur lesquelles elle a été évaluée dans le cadre de cette thèse, à savoir la recherche d'informations, l'indexabilité et la performance du SRAP. En effet, des gains en terme de WER et $IWER_{Average}$ ont été observés pour la tâche de transcription, et des gains en terme de taux de couverture et de corrélation de Spearman pour les tâches de recherche d'informations et d'indexabilité. Cette adaptation s'appuie sur des données textuelles extraites à partir du web. Cependant, le web est un espace d'information et de partage en constante évolution. Certaines pages disparaissent quelques heures, voire quelques minutes, après avoir été publiées. D'autres pages web ne changent pas pendant des décennies. L'indexation des moteurs de recherche permet

d'effectuer une exploration de la toile à la recherche de nouveaux contenus, puis à les analyser et à donner un classement aux différentes pages où ils se trouvent. Compte tenu de cette évolution, l'étude présentée dans ce chapitre propose un travail axé sur la reproductibilité des résultats obtenus par les systèmes de reconnaissance de la parole à l'aide de modèles de langage adaptés à partir de données collectées sur Internet dont le but est de répondre sur la question suivante : *la qualité d'un modèle de langage adapté dépend-elle de la période de collecte des données d'adaptation ?*

Ce chapitre est organisé comme suit. Nous commençons par les motivations de cette étude dans la partie 7.2. Nous présentons par la suite, dans la partie 7.3, les résultats obtenus. Enfin, nous proposons une discussion, dans la partie 7.4, par rapport aux expériences menées, ainsi qu'une conclusion à ce chapitre.

7.2 Description de l'étude diachronique

Dans la présente étude diachronique, nous examinons l'impact de données du web collectées à des dates différentes sur les résultats d'adaptation du modèle de langage (ML). Comme précisé dans l'introduction de ce chapitre, de part le caractère évolutif et instable du web, nous nous interrogeons sur la reproductibilité des résultats d'adaptation des modèles de langage. Nous présentons dans cette section les motivations qui nous ont poussées à réaliser cette étude.

7.2.1 Motivations

L'évolution rapide du web est une de ses caractéristiques. Plusieurs études se sont intéressées (NTOULAS, 2006 ; FETTERLY et al., 2004) à étudier cette évolution. La figure 7.1 montre l'évolution du nombre de pages web entre les années 2000 et 2018. On remarque que le nombre de pages web varie d'une année à une autre. On observe aussi que cette variation n'est pas toujours croissante (parfois on observe une augmentation et parfois on observe une diminution du nombre de pages).

Des travaux antérieurs ont étudié le côté diachronique de l'évolution du web dans le but de traiter les mots hors vocabulaire pour les systèmes de reconnaissance de la parole. (ALLAUZEN et GAUVAIN, 2005) ont étudié l'utilisation des sources Internet pour l'adaptation quotidienne du vocabulaire d'un système de transcription de nouvelles radiodiffusées en deux langues : le français et l'anglais britannique.

(SHEIKH, ILLINA et FOHR, 2016) ont comparé l'enrichissement du vocabulaire avec des noms propres sur des périodes temporelles différentes en utilisant plusieurs sources.

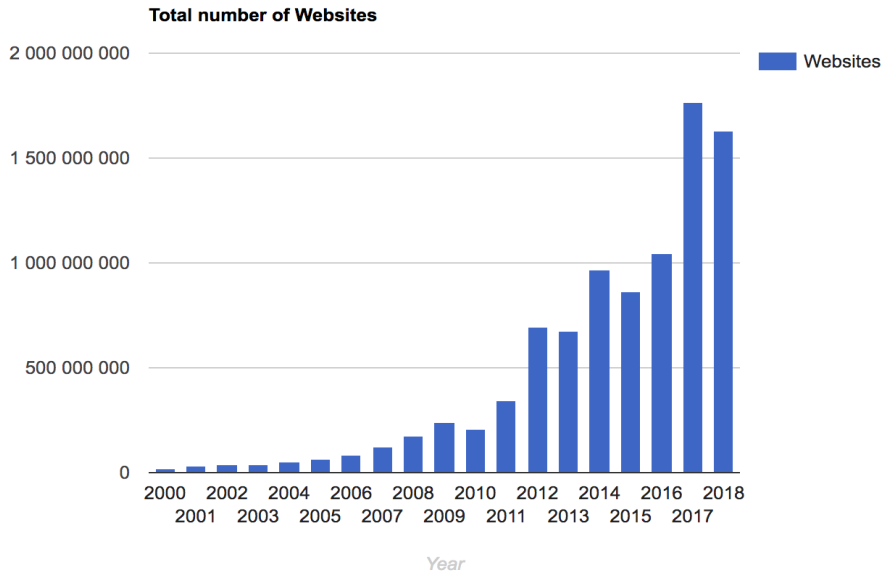


FIGURE 7.1 – Évolution du web entre les années 2000 et 2018 *(Internetlivestats.com)*

Cependant, ces études n'ont jamais été faite dans l'objectif de mesurer la reproductibilité des résultats de l'adaptation des modèles de langage.

7.2.2 Formulation du problème

Nous nous plaçons dans le contexte de l'adaptation des modèles de langage (ML). Nous supposons que la collecte de pages web à deux moments différents produira deux ensembles de pages web différents. Pour chaque mois, pendant une période déterminée, nous répétons le même processus d'adaptation du modèle de langage présenté dans la section 5.6.3. Par la suite, nous calculons le taux d'erreurs sur les mots de l'adaptation du ML avec les données issues de chaque mois. Un intervalle de confiance global sera calculé afin de comparer les taux d'erreur obtenus sur les différentes périodes : si le taux est compris dans cet intervalle, alors nous pouvons considérer que l'adaptation du modèle de langage à cette période produit une performance identique aux autres périodes d'adaptation, et inversement si le taux se trouve en dehors de cet intervalle. Nous utilisons la formule suivante pour calculer un intervalle de confiance :

$$CI \pm z * \sqrt{\frac{ER * (1 - ER)}{N}} \tag{7.1}$$

avec N est le nombre de mots dans la référence, ER représente un taux d'erreurs divisé par 100 et z, issue de la table de Student, correspond à une valeur qui dépend du niveau de confiance. Dans le cas de 95% d'intervalle de confiance, la valeur de z est égale à 1,96.

Les valeurs de la borne inférieure issues des intervalles de confiance (pour chaque mois de la période définie) vont permettre de déterminer une valeur que nous avons nommée $min\_confidence\_interval$. Cette valeur correspond à la valeur maximale entre toutes les valeurs de la borne inférieure issues des intervalles de confiance. Les valeurs de la borne supérieure vont permettre de déterminer une valeur que nous avons nommée $max\_confidence\_interval$. Cette valeur correspond à la valeur minimale entre toutes les valeurs de la borne supérieure issues des intervalles de confiance. Dans le cadre de notre étude sur l'adaptation des ML, nous considérerons alors comme reproductibles nos résultats si, pour toute période considérée, les taux d'erreurs se trouvent inclus dans la borne $min\_confidence\_interval$ et la borne $max\_confidence\_interval$.

7.3 Analyse et résultats

En considérant une année de collecte (octobre 2018 à octobre 2019), la partie 7.3.1 présente une analyse qualitative de l'évolution des pages web d'un mois à un autre. Dans la partie 7.3.2, nous présentons ensuite les résultats obtenus sur la tâche d'adaptation des modèles de langage pour chaque période de temps considérée.

7.3.1 Analyse des données web collectées

Nous commençons notre étude diachronique par une analyse de l'évolution des pages Web au cours d'une année de collecte de documents. Pour ce faire, nous avons calculé le nombre de pages collectées au cours du $n_{ème}$ mois qui sont toujours présentes dans la collecte du premier mois (c'est-à-dire l'intersection de pages web entre deux jeux de données collectées : les pages issus de la collecte du premier mois et les pages issus de la collecte du $n_{ème}$ mois).

La figure 7.2 présente les statistiques pour les différents cours de notre corpus. L'axe horizontal de ce graphique représente le mois et l'axe vertical indique le nombre de pages web. Les barres représentent le nombre de pages du premier mois encore disponibles au cours d'un mois donné.

En examinant le nombre total de pages web explorées dans la figure 7.2, nous constatons que le nombre de pages communes comparé au premier mois diminue rapidement. Par exemple, après un mois de collecte (novembre 2018), pour le cours *"Réseaux sociaux et graphes"*, 80% des pages du premier mois sont encore disponibles et après 12 mois d'exploration (octobre 2019), seulement 45% des pages sont disponibles.

Comme d'autres études l'ont déjà indiqué, les données web de notre étude ont tendance à évoluer rapidement, même sur des sujets qui ne sont pas susceptibles de beaucoup changer (dans le cadre de notre étude il s'agit de cours d'informatique).

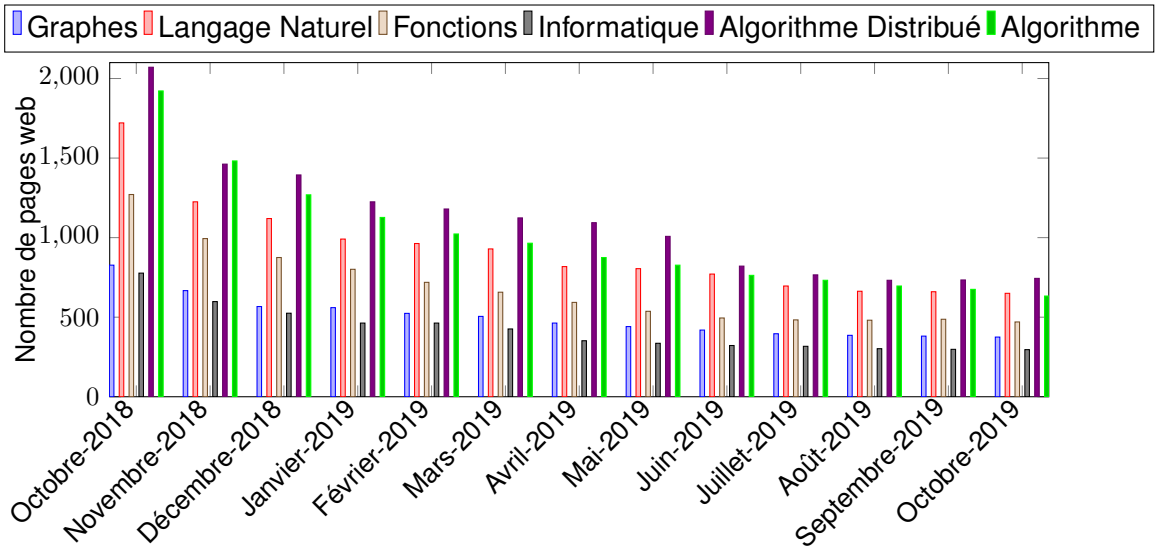


FIGURE 7.2 – Nombre de pages de la collecte d'octobre 2018 qui existent encore après n mois de collecte

7.3.2 Résultats de l'adaptation diachronique

L'analyse effectuée sur le nombre de pages Web (section 7.3.1) collectées montre l'intérêt de l'étude diachronique de l'adaptation de LM.

7.3.2.1 Résultats en WER

Nous commençons notre analyse en nous focalisant sur l'évaluation en taux d'erreur-mot (WER) après adaptation pour chaque collecte mensuelle. La figure 7.3 présente les résultats obtenus en termes de WER (ligne bleue). L'intervalle de confiance, telle que décrite dans la partie précédente, a été reportée sur la figure au moyen de la ligne rouge, pour la borne inférieure ($min\_confidence\_interval$), et la ligne verte, pour la borne supérieure ($max\_confidence\_interval$). On observe que les résultats en WER varient légèrement d'un mois à un autre. Malgré cette variabilité, on remarque que toutes les valeurs de WER sont incluses entre la valeur maximale et la valeur minimale définies à partir des intervalles de confiance. Cette première série d'expériences, avec évaluation en taux d'erreur-mot, montre qu'il n'y a pas de différence statistiquement significative dans les résultats obtenus entre chaque période de temps étudiée.

7.3.2.2 Résultats en $\mathrm{IWER}_{\mathrm{Average}}$

L'étude présentée dans le chapitre 6 a montré que la métrique classique des systèmes de reconnaissance automatique de la parole, le taux d'erreur sur les mots (WER), n'est pas

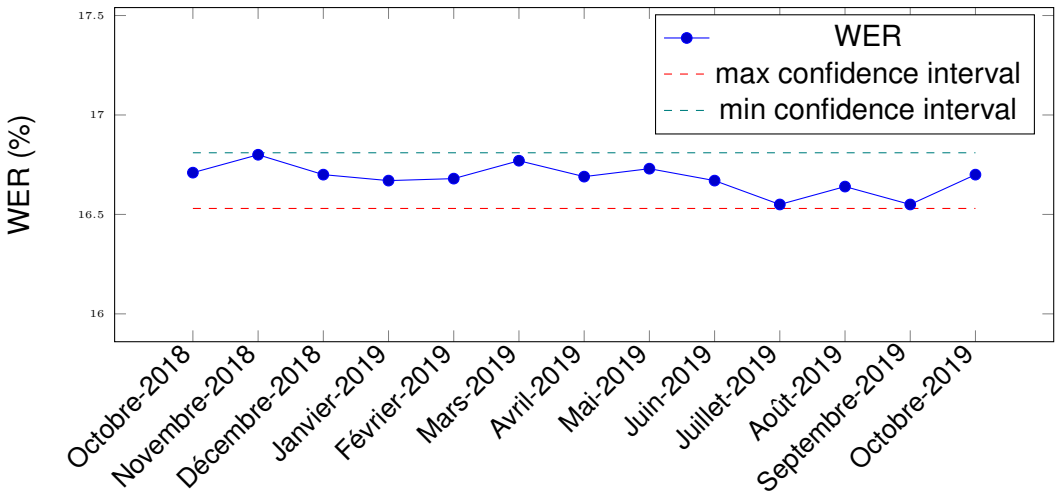


FIGURE 7.3 – (%) WER et intervalles de confiance résultats pour l'adaptation diachronique du modèle de langage : évaluation de Octobre 2018 à Octobre 2019

suffisante pour capturer l'impact de l'adaptation des modèles de langage dans le contexte de transcription de cours magistraux, et qu'il est important d'évaluer l'adaptation en se basant sur des mots du domaine. Partant de cette précédente étude, nous proposons d'évaluer la performance de l'adaptation diachronique en utilisant la métrique $IWER_{Average}$ (MDHAFFAR et al., 2019b) présentée en section 6.3. La figure 7.4 montre les résultats de l'adaptation diachronique en fonction de la métrique $IWER_{Average}$. Les lignes rouge et verte représentent les mêmes informations que la figure 7.3, à savoir l'intervalle de confiance. Tout comme pour l'évaluation en WER, bien que les résultats varient légèrement d'un mois à l'autre, aucune de ces variations n'apparaît statistiquement significative considérant la métrique $IWER_{Average}$.

7.3.2.3 Résultats pour les pages web en commun

Comme nous l'avons présenté dans la partie 7.3.1, il existe un nombre de pages en commun entre deux mois donnés, correspondant alors à l'intersection entre deux corpus. Dans le tableau 7.1, nous présentons le nombre de pages en commun entre tous les mois de collecte (intersection entre les pages de 13 mois) pour tous les cours du corpus [1]. On observe que ce nombre représente un faible pourcentage comparé au nombre total de pages retournées durant un mois donné (19,17% pour le cours *"Introduction à l'informatique"* par rapport au nombre initial de pages d'octobre 2018). Dans la présente étude, nous nous interrogeons sur l'implication de ce sous-corpus de données dans la performance finale de l'adaptation des modèles de langage. Le tableau 7.2 présente les résultats en WER et $IWER_{Average}$ de la performance de

1. (1) Introduction à l'informatique, (2) Introduction à l'algorithme, (3) Les fonctions, (4) Réseaux sociaux et graphes, (5) Algorithme distribué, (6) Langage naturel

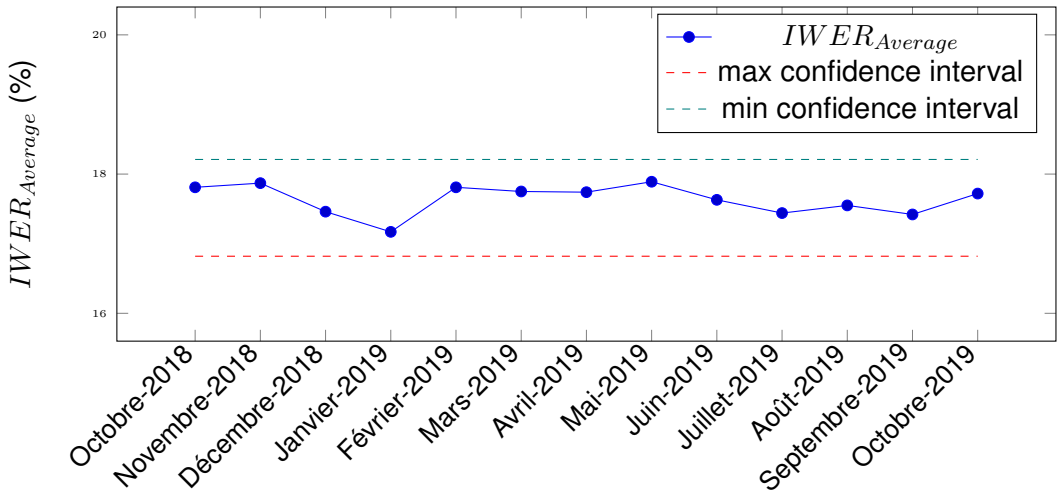


FIGURE 7.4 – (%) $IWER_{Average}$ et intervalles de confiance résultats pour l'adaptation diachronique du modèle de langage : évaluation de Octobre 2018 à Octobre 2019

l'adaptation du ML en n'utilisant que les documents en commun comme corpus d'adaptation. Malgré la faible quantité de pages web, on observe une amélioration de l'adaptation du ML de 38,61% en $IWER_{Average}$ (de 31% à 19,03%) et de 10,53% en WER.

TABLE 7.1 – Nombre de pages web en commun durant 13 mois de collecte pour chaque cours du corpus

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Nombre de pages en octobre 2018	777	1922	1271	827	2071	1721
Nombre de pages commun (13 mois de collecte)	149	301	221	216	333	342

En comparant les résultats obtenus avec les pages en commun (17,41% WER et 19,03% $IWER_{Average}$) et les résultats obtenus en moyennant les résultats mensuelles (16,60% WER et 17,63% $IWER_{Average}$) avec les résultats du modèle de langage générique (19,46% WER et 31% $IWER_{Average}$), on observe que les deux apportent des améliorations. On peut donc penser que ce sous-ensemble de documents commun contient des documents très pertinents pour l'adaptation des modèles, mais que la partie variable de chaque collection a néanmoins un impact positif sur les résultats de l'adaptation.

7.3.2.4 Résultats en variant la taille du nombre de requêtes

Nous avons montré dans les sections précédentes que l'adaptation du ML dans deux périodes différentes donne des changements de résultats non significatifs. Cette adaptation a été faite en collectant 100 pages web par requête. Nous étudions dans cette section l'impact du

TABLE 7.2 – (%) $IWER_{Average}$ et (%) WER en utilisant les pages web en commun durant 13 mois de collecte : de Octobre 2018 jusqu'au Octobre 2019.

	ML générique	**ML adapté moyenne 1-an**	**ML adapté pages en commun**
WER	19,46	16,60	17,41
$IWER_{Average}$	31	17,63	19,03

nombre de pages collectées sur les performances de l'adaptation diachronique.

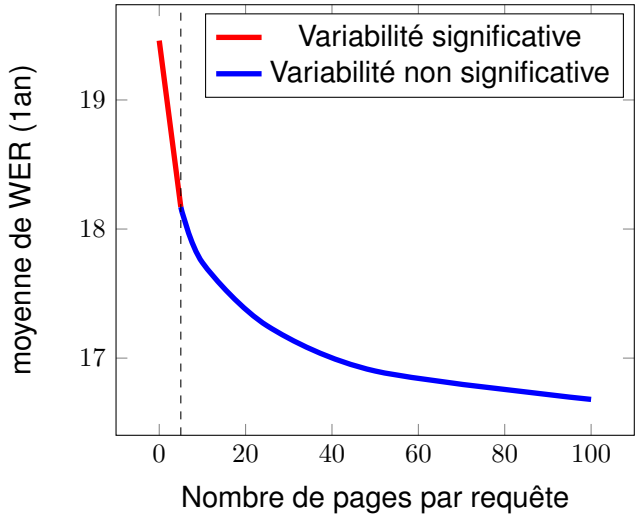


FIGURE 7.5 – WER (moyenne des WER pour 13 mois de collecte : de Octobre 2018 à Octobre 2019) pour différentes valeurs de nombre de page par requête

Les figures 7.5 et 7.6 montrent la performance en WER et $IWER_{Average}$ (moyenne pour les 13 mois de collecte). La ligne bleu correspond à des changements de résultats non significatifs. La ligne rouge correspond à des changements de résultats significatifs. On observe que à partir de 5 pages par requête, les changements en WER et $IWER_{Average}$ ne sont pas significatifs. Mais par contre, ce n'est pas le cas pour un nombre de pages inférieur à 5.

7.4 Discussion

À partir de l'analyse diachronique, il a été démontré que l'adaptation des modèles de langage dans le cadre de notre étude (collecte mensuelle de données pendant un an dans un contexte pédagogique informatique), permet d'obtenir des résultats reproductibles (en WER et $IWER_{Average}$) sans variation significative de la performance du SRAP, quelle que soit la période de collecte considérée.

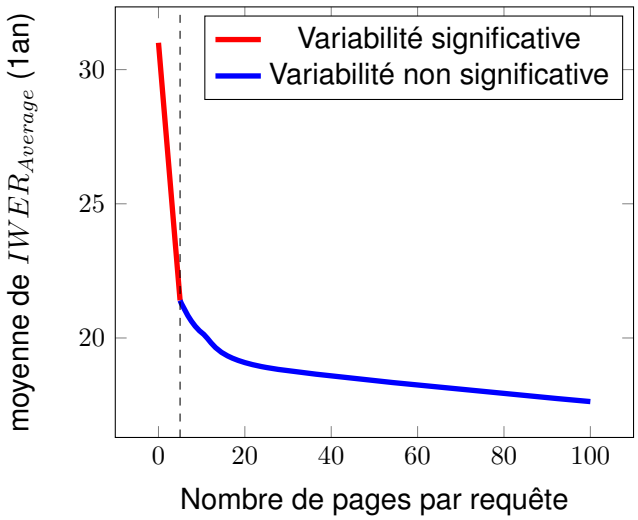


FIGURE 7.6 – $IWER_{Average}$ (moyenne des $IWER_{Average}$ pour 13 mois de collecte : de Octobre 2018 à Octobre 2019) pour différentes valeurs de nombre de page par requête

Dans un deuxième temps, nous avons également souligné qu'en considérant le sous-ensemble des documents (les documents commun entre 13 mois), on peut obtenir de bonnes performances, ce qui laisse à penser qu'un petit nombre de documents suffit ou, du moins, concentre suffisamment d'informations pour adapter le modèle logique à une tâche spécifique.

7.5 Conclusion

Dans ce chapitre, nous avons examiné l'utilisation des sources web pour l'adaptation des modèles de langage pour un SRAP. Un ensemble d'expériences diachroniques a été réalisé. L'objectif principal est d'évaluer l'impact des données recueillies sur le web sur différentes périodes temporelles pour l'adaptation des modèles de langage. Sur une période de collecte d'un an (de Octobre 2018 jusqu'à Octobre 2019), nous avons pu montrer que, même si les données sur le web changent en partie d'un mois à l'autre, la performance des systèmes de transcription adaptés est restée constante (c'est-à-dire sans changement significatif de performance), peu importe la période considérée.

Chapitre 8

Structuration automatique de transcription

Contents

8.1 Introduction

Après avoir obtenu une transcription de bonne qualité, il est utile que les informations pertinentes au sein d'une transcription soient faciles à retrouver. La navigation au sein d'une transcription peut s'avérer longue et fastidieuse, surtout si l'information ciblée se trouve dans un document qui ne contient ni limites de paragraphes, ni ponctuations. La structuration de la transcription comme, par exemple la segmentation thématique, peut résoudre ce problème. Nous proposons dans ce chapitre deux types de structurations. La première consiste en une

segmentation thématique. La seconde concerne l'alignement des segments de transcription avec les diapositives.

8.2 Segmentation automatique de la transcription

Comme présenté dans le chapitre 3, il existe plusieurs techniques de segmentation. Ces techniques ont été appliquées aussi bien aux documents écrits qu'aux transcriptions automatiques.

8.2.1 Pourquoi TextTiling ?

Plusieurs raisons nous ont poussés à choisir l'algorithme TextTiling (section 3.2.1.2) pour implémenter nos expériences de segmentation de cours magistraux.

Le corpus PASTEL est un corpus peu volumineux qui contient 389 segments de type "Granularité 1" et 46 segments de type "Granularité 2" (section 5.3.2). Ce nombre de segments ne semble pas avantageux pour utiliser une technique de segmentation supervisée. TextTiling est un algorithme non supervisé qui ne nécessite aucune donnée d'apprentissage. Il repose sur des calculs de similarité entre blocs. Cela constitue aussi un avantage pour les cours magistraux, du fait que dans un cours les phrases ne sont pas riches lexicalement : il est donc plus intéressant de se baser sur des unités plus larges qu'une phrase (qui correspond aux blocs, dans le cas de l'algorithme TextTiling).

8.2.2 Pré-traitements et représentation vectorielle des blocs

Nous avons appliqué les pré-traitements classiques textuels sur les transcriptions : les transcriptions ont été lemmatisés en utilisant l'outil MACAON [1] (NASR et al., 2011) et les mots vides ont été supprimés.

L'algorithme TextTiling se base sur une représentation vectorielle des blocs adjacents pour calculer leur similarité. Le calcul des représentations vectorielles consiste à associer à chaque mot du corpus un score traduisant son importance au sein du segment. Pour représenter les blocs, nous avons choisi d'utiliser le critère TF-IDF, couramment utilisé en recherche d'information, pour traduire la capacité d'un mot à discriminer le document dans lequel il se trouve par rapport à une collection de textes. TF signifie "term frequency" et IDF "inverted document frequency". Grâce au TF, on détermine la fréquence relative d'un mot dans un document. Grâce à l'IDF, on mesure si le terme se trouve dans la plupart des documents ou non. Le poids du

1. https ://gitlab.lif.univ-mrs.fr/benoit.favre/macaon

terme t dans le document d est donné par :

$$w_{TF-IDF}(t,d) = TF(t,d) \times IDF(t) \tag{8.1}$$

avec $TF(t,d)$ la fréquence du terme t dans le document d et

$$IDF(t) = log(\frac{N}{n_t}) \tag{8.2}$$

où N est le nombre total de documents et n_t est le nombre de documents dans lequel le terme t apparaît. Dans notre cas, un document correspond à un groupe de souffle.

8.2.3 Résultats et discussion

Dans cette section, nous présentons les résultats obtenus pour la segmentation thématique. Étant données les limites des métriques Beeferman p_k et WindowDiff présentées dans la section 3.3, nous évaluons la performance de la segmentation en utilisant la métrique F-mesure. De façon similaire à (GUINAUDEAU, GRAVIER et SÉBILLOT, 2010) et (BOUCHEKIF, 2016), nous utilisons une tolérance de 10 secondes. Les résultats sont présentés dans le tableau 8.1. Comme notre algorithme est basé sur une fenêtre glissante paramétrable, nous avons appliqué la validation croisée (*leave-one-out cross-validation*) afin de fixer cette valeur pour chaque cours. Pour n cours, la valeur de la fenêtre est optimisé sur la moyenne de la F-mesure des $n-1$ autres cours du corpus. Puis, pour le énième cours, on valide en utilisant la valeur obtenu par optimisation. On répète ce processus n fois (pour tous les cours dans notre corpus).

TABLE 8.1 – Résultats (%) de segmentation en utilisant la transcription manuelle

	Précision	**Rappel**	**F-mesure**
TextTiling	27,53	65,66	38,79
Changement de diapositives	48,60	83,39	61,4
SliTextTiling	60,05	85,66	70,6

Les résultats présentés dans le tableau 8.1 montrent une faible performance en utilisant l'algorithme de segmentation TextTiling.
Comme les annotations manuelles de segmentation se sont appuyées sur les changements de diapositives, nous avons calculé la performance de segmentation en considérant que chaque changement de diapositive correspond à une frontière thématique. Les résultats sont présentés dans la deuxième ligne du tableau 8.1. Comme prévu, les résultats montrent que la structure en diapositives d'un cours donne une information importante pour la tâche de segmentation.
La prise en compte de cette information dans l'algorithme de segmentation peut peut-être

mieux guider le choix des frontières thématiques. Nous avons donc modifié l'algorithme TextTiling de sorte que, dans le calcul de la similarité, nous soyons capable de donner un poids relatif à la distance du bloc avec un changement de diapositive. La dernière ligne du tableau 8.1 présente la performance de ce nouvel algorithme, appelé SliTextTiling. Nous constatons que l'intégration de cette information de distance permet d'améliorer les résultats (de 38,79% à 70,6% en F-mesure).

8.3 Alignement automatique de la transcription avec les diapositives

Comme nous l'avons vu dans la section 4.5.4, l'alignement automatique de la transcription avec les diapositives a été étudié dans la littérature dans quelques travaux. L'alignement des transcriptions avec des diapositives est cependant une tâche complexe en raison des dépendances de ces systèmes :

— la précision des transcriptions de la parole ;
— le contenu des diapositives : les informations de surface dans les diapositives sont généralement très limitées (avec la présence de diapositives contenant uniquement des figures, uniquement des formules mathématiques, uniquement des tableaux) ;
— le discours de l'orateur : les orateurs suivent rarement exactement le contenu de leurs diapositives de présentation.

Nous présentons, dans cette section, la méthode d'alignement adoptée. Nous proposons d'étendre l'approche d'alignement de (YAMAMOTO, OGATA et ARIKI, 2003 ; JUNG, SHIN et KIM, 2018) en intégrant un mécanisme d'optimisation globale avec une contrainte de séquence pour respecter l'ordre séquentiel des diapositives et des transcriptions. En plus de la mesure de précision traditionnelle, nous considérons l'erreur quadratique moyenne qui offre un score d'évaluation plus fin en distinguant différemment les dés-alignements avec une distance courte et ceux avec une distance longue entre la référence et l'hypothèse. Enfin, nous considérons une mesure de confiance pour discuter de la fiabilité de l'approche proposée.

8.3.1 Pré-traitements et représentation vectorielle

La méthode que nous proposons exploite le contenu textuel des diapositives et la transcription de la parole pour effectuer l'alignement. Nous procédons de la même façon que pour la tâche de segmentation thématique : le texte des diapositives et les transcriptions automatiques et manuelles sont lemmatisés à l'aide de l'outil MACAON (NASR et al., 2011) et les mots vides sont supprimés. Nous construisons une représentation TF-IDF des segments de transcription et des diapositives qui calcule un score pour les mots en fonction de la collection

de documents à laquelle ils appartiennent (en considérant les diapositives et les segments de transcription comme deux collections distinctes).

8.3.2 Méthode proposée

La méthode proposée a été conçue de manière à prendre en compte :

— la similarité textuelle entre les diapositives et les segments de transcription de la parole.

— l'ordre linéaire des diapositives et des segments de transcription.

Nous construisons un module séparé pour chacune de ces deux analyses et les fusionnons pour obtenir la décision finale.

— **Similarité entre les diapositives et les segments de transcription**
Soit $S = \{s_1, s_2, ..., s_n\}$ l'ensemble des diapositives et soit $T = \{t_1, t_2, ..., t_m\}$ l'ensemble des segments de transcription, avec n le nombre de diapositives dans un cours et m le nombre de segments de transcription dans un cours. On définit $Sim(t_i, s_j)$ comme étant la similarité entre la représentation vectorielle du segment de transcription t_i et la représentation vectorielle de la diapositive s_j.
Soit $\Pi = \{\pi_1, \pi_2, ..., \pi_k\}$ l'ensemble de séquences possibles $\pi_x = [(t_1, s_i), (t_2, s_j), ..., (t_m, s_k)]$ de couples (segment de transcription, diapositive). Pour chaque séquence π_x, nous calculons le score $L(\pi_x)$, suivant la formule suivante :

$$L(\pi_x) = \prod_{(t_i, s_j) \in \pi_x} Sim(t_i, s_j) \tag{8.3}$$

— **Ordre linéaire des diapositives et des segments de transcription**
La contrainte d'ordre des segments de parole et des diapositives est définie pour imposer un ordre linéaire aux diapositives et aux segments de parole.
Soit $\alpha = [p_1, p_2, ..., p_m]$ une séquence de paires de diapositives et de segments de transcriptions qui respecte la contrainte d'ordre définie comme suit :

- Le segment de transcription de p_{i+1} est le segment qui suit le segment de transcription de p_i, du point de vue temporel.
- La diapositive associée à p_{i+1} peut être soit la même diapositive que celle de p_i ou soit la diapositive qui suit la diapositive de p_i.

$\beta = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ correspond ainsi à l'ensemble des séquences possibles α_i qui respectent les contraintes mentionnées ci-dessus.

— **Alignement de la transcription avec les diapositives**

Notre objectif est de trouver la meilleure séquence parmi Π, qui respecte l'ordre des diapositives et des segments de parole. Le processus décisionnel global consiste à choisir la séquence $\bar{H}$ qui maximise le score global obtenu par la fusion (intersection) de la *similarité entre diapositive et transcription de la parole* et de la *contrainte d'ordre des segments de diapositive et de parole*. La séquence $\bar{H}$ est calculée en utilisant l'équation suivante :

$$\bar{H} = \underset{\pi_x}{argmax}\ (\Pi \cap \beta) \tag{8.4}$$

8.3.3 Mesures d'évaluation

La précision est la métrique standard utilisée pour évaluer la performance de la tâche d'alignement de la transcription vers la diapositive (YAMAMOTO, OGATA et ARIKI, 2003 ; LU et al., 2014). La précision se calcule comme suit :

$$Précision = \frac{Nombre\ de\ segments\ de\ transcription\ correctement\ affectés\ \text{à}\ une\ diapositive}{Nombre\ total\ de\ segments\ de\ transcription} \tag{8.5}$$

Une précision de 100% signifie que l'ensemble des alignements proposés par le système est correct. L'alignement d'une hypothèse avec une diapositive proposé par le système n'est considéré comme correct que s'il coïncide exactement avec la diapositive réelle.

Étant donné que l'un des objectifs du projet PASTEL est de faciliter la navigation dans la transcription et les diapositives, la mesure de précision n'est pas adaptée à notre tâche car un petit ou un grand décalage dans l'alignement sont considérés comme faux avec la même pénalité. Par conséquent, pour une évaluation tenant mieux compte de la gravité d'une erreur, nous proposons d'utiliser l'erreur quadratique moyenne (MSE - Mean Square Error). L'erreur quadratique moyenne correspond à la perte quadratique moyenne pour chaque exemple (la distance en termes de nombre de diapositives entre la référence et l'hypothèse). Pour calculer l'erreur MSE, il faut additionner toutes les pertes quadratiques de chaque exemple, puis diviser cette somme par le nombre d'exemples (voir Équation 8.6).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\bar{S}_i - S_i)^2 \tag{8.6}$$

où m est le nombre total de segments de transcription, $\bar{S}_i$ correspond au numéro de diapositive affectée par notre système à un segment de transcription i et S_i est le numéro de diapositive réellement affectée au segment de transcription i.

8.3.4 Résultats et discussion

Nous présentons dans cette section les résultats expérimentaux de notre méthode.

8.3.4.1 Résultats du système de base

Le système de base consiste en un système de classification simple tel que celui présenté dans (JUNG, SHIN et KIM, 2018). La classification consiste à sélectionner, pour chaque segment de transcription, la diapositive la plus similaire en terme de distance cosinus entre les vecteurs TF-IDF. Les résultats en précision et en MSE du système de base sont présentés dans le tableau 8.2.

TABLE 8.2 – Résultats de l'alignement de la transcription avec les diapositives du système de base (JUNG, SHIN et KIM, 2018) à l'aide du MSE et de la précision (résultats pour la transcription manuelle, la transcription sans adaptation du ML, la transcription avec adaptation du ML)

	Sans adaptation du ML	**Avec adaptation du ML**	**Transcription manuelle**
Précision	18,96%	21,49%	24,98%
MSE	681,31	638,024	657,980

Les résultats dans le tableau 8.2 montrent une faible performance en précision et en MSE. Les valeurs élevées en MSE montrent que les hypothèses produites par le système d'alignement sont très distantes par rapport aux références. Ce phénomène met en évidence la nécessité d'une contrainte sur l'ordre des diapositives. Les résultats expérimentaux montrent également l'utilité de l'adaptation du ML pour la tâche d'alignement des diapositives.

8.3.4.2 Résultats de la méthode proposée

Le tableau 8.3 présente la performance d'alignement de la méthode proposée (avec contrainte sur l'ordre séquentiel des diapositives) en utilisant la transcription manuelle, la transcription automatique avec adaptation du ML et la transcription automatique sans adaptation du ML. Les lignes 1 et 2 illustrent respectivement les performances de notre système en termes de précision et de MSE.

TABLE 8.3 – Résultats de l'alignement de la transcription avec les diapositives de la méthode proposée à l'aide du MSE et de la précision (résultats pour la transcription manuelle, la transcription sans adaptation du ML, la transcription avec adaptation du ML)

	Sans adaptation du ML	**Avec adaptation du ML**	**Transcription manuelle**
Précision	44,32%	58,46%	63,28%
MSE	2,481	1,424	1,313

Les résultats montrent que l'approche proposée apporte des améliorations significatives par rapport au système de base. La méthode proposée améliore la précision de 24,98% à 63,28% et le MSE de 657,980 à 1,313 en utilisant la transcription manuelle.

La figure 8.1 illustre un exemple d'alignement pour le cours *"introduction à l'informatique"*. La ligne verte correspond à la référence. La ligne rouge correspond à la sortie du système d'alignement. Les carreaux en bleu correspondent à la similarité entre le vecteur TF-IDF de la diapositive i et le vecteur TF-IDF du segment de transcription j.

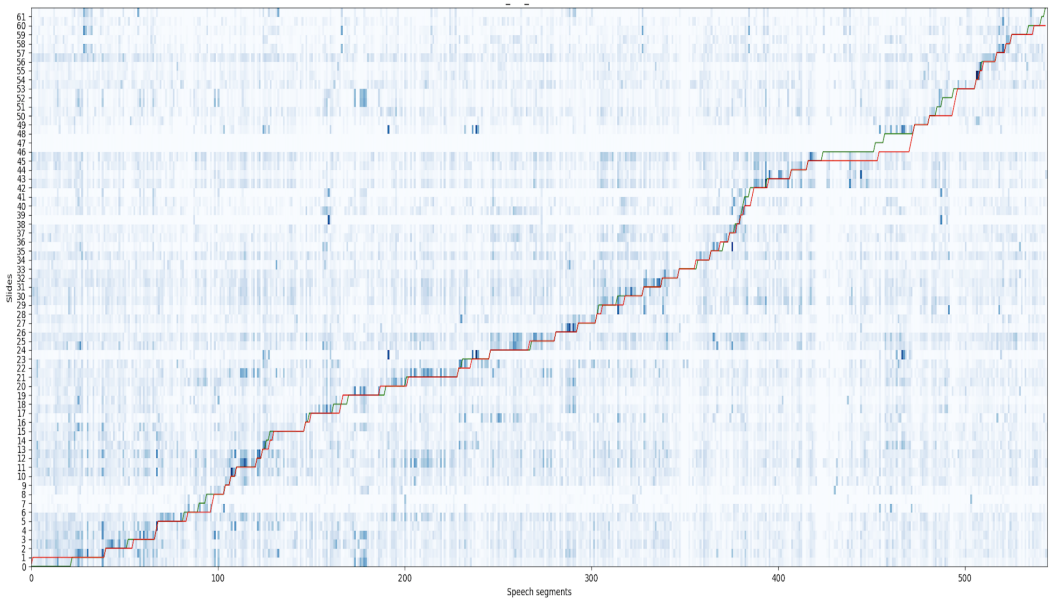


FIGURE 8.1 – Exemple d'alignement pour le cours *"introduction à l'informatique"*

8.3.4.3 Résultats en comparant le calcul des représentations TF-IDF

Le tableau 8.4 montre la performance d'alignement en utilisant la transcription manuelle et en considérant les diapositives et les segments de parole comme une seule collection pour construire la représentation TF-IDF. Ces résultats montrent l'utilité de la méthode que nous proposons en considérant les diapositives et les segments de transcription comme des collections distinctes pour le calcul des représentations TF-IDF. On observe une perte de 1,424 à 1,708 en MSE et de 58,46% à 56,11% sur la précision en utilisant la transcription automatique avec adaptation du ML.

TABLE 8.4 – Résultats de l'alignement de la transcription avec les diapositives de la méthode proposée à l'aide du MSE et de la précision (résultats pour la transcription manuelle, la transcription sans adaptation du ML, la transcription avec adaptation du ML)

	Sans adaptation du ML	**Avec adaptation du ML**	**Transcription manuelle**
Précision	41,19%	56,11%	62,64%
MSE	3,268	1,708	0,911

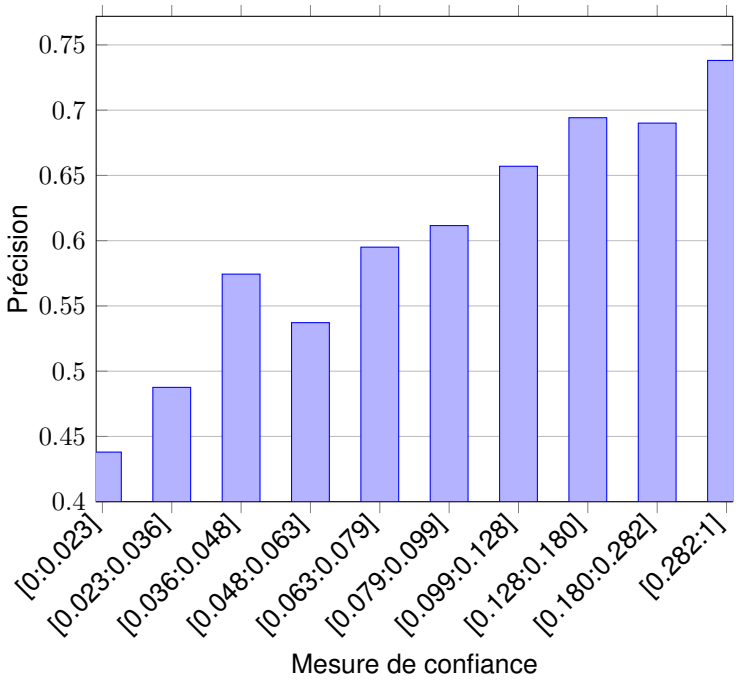


FIGURE 8.2 – Précision pour différentes valeurs de confiance

8.3.5 Mesure de confiance

Les mesures de confiance sont utilisées pour évaluer la fiabilité des résultats obtenus. Nous nous sommes intéressés à évaluer la fiabilité de la méthode proposée, en visualisant la performance sur la précision et sur le MSE, pour différents intervalles de mesure de confiance. Nous définissons le score de confiance comme suit :

$$Score_{ij} = \frac{Sim_{ij}}{\sum_{k=1}^{n} Sim_{kj}} \tag{8.7}$$

où i et j représentent respectivement les diapositives et les segments de transcription, Sim_{ij} représente la similarité cosinus entre le vecteur de représentation de i et le vecteur de représentation de j et n correspond au nombre total de diapositives.

Les figures 8.2 et 8.3 présentent à la fois les résultats en précision et MSE liés aux scores

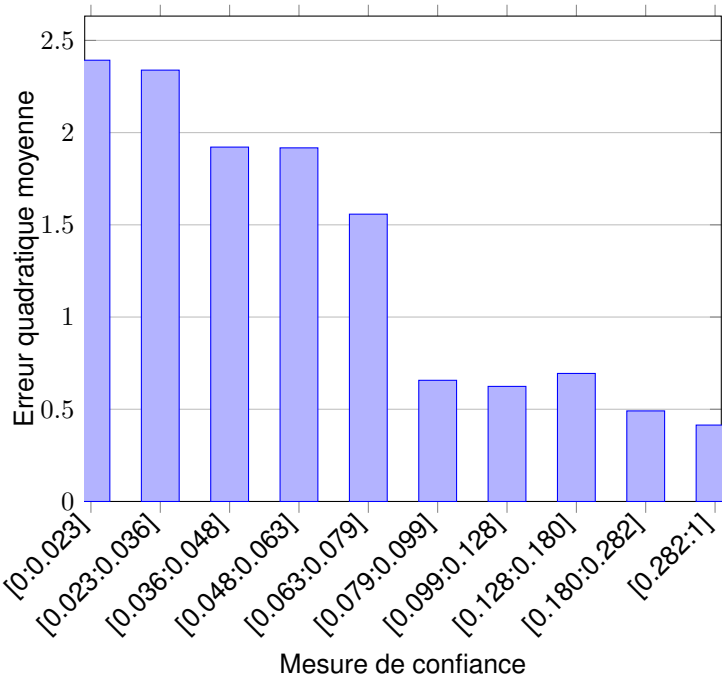


FIGURE 8.3 – Erreur quadratique moyenne (MSE) pour différentes valeurs de confiance

de confiance. Les intervalles sont générés en divisant les données en 10 parties égales. Les meilleurs résultats en termes de précision et de MSE sont atteints pour un score de confiance plus élevé. On observe que la mesure de confiance proposée a une forte corrélation avec la similitude entre les diapositives et les transcriptions. Cette corrélation a été statistiquement calculée en utilisant le coefficient de corrélation de Pearson. Entre les valeurs de précision et les valeurs de mesures de confiance, la valeur de corrélation est 0,807 (corrélation positive). Entre les valeurs de MSE et les valeurs de mesures de confiance, la valeur de corrélation est -0,743 (corrélation négative : ce qui signifie que les scores de confiance élevés tendent à augmenter lorsque celles de MSE diminuent). Cela illustre la fiabilité de la méthode proposée.

8.4 Conclusion

Dans ce chapitre, nous avons proposé deux techniques pour la structuration de la transcription. La première technique consiste à segmenter thématiquement la transcription. Cette segmentation repose sur l'algorithme TextTiling que nous avons adapté à la transcription automatique des cours magistraux. Nous avons notamment apporté des modifications en intégrant l'information de changement de diapositive dans le calcul de la similarité entre deux blocs de transcription. La deuxième technique de structuration concerne l'alignement de la transcription sur les diapositives. Afin d'accomplir cette tâche, nous avons calculé la similarité cosinus entre

la représentation TF-IDF des segments de transcription et la représentation TF-IDF des diapositives de texte et nous avons imposé une contrainte pour respecter l'ordre séquentiel des diapositives et des segments de transcription. Les résultats ont été comparés à une classification naïve et montrent une amélioration significative. Nous avons montré que l'adaptation des modèles de langage apporte également des améliorations significatives pour la tâche d'alignement de la transcription et des diapositives.

Conclusion et perspectives

Conclusion

Les systèmes de reconnaissance automatique de la parole (SRAP) sont sensible à la variation des sujets. On attend pourtant des systèmes de reconnaissance automatique de la parole qu'ils soient capables de transcrire précisément une très large palette de sujets.

Ce travail de thèse a porté sur plusieurs axes. Dans un premier temps, nous avons proposé la création et l'annotation d'un corpus lié à l'éducation dans le cadre du projet PASTEL. Ce corpus est composé de transcriptions manuelles de discours d'enseignants en situation de cours magistraux. En plus des discours, le corpus contient les supports de présentation du cours (diapositives) et la vidéo. Il s'accompagne d'informations annotées manuellement par des experts humains, à savoir une segmentation thématique des cours, une annotation en expressions clés, et enfin un alignement des diapositives avec la vidéo. Les données et les annotations seront distribuées sous licence libre à la communauté scientifique.

Les transcriptions de cours magistraux portant sur divers sujets, nous nous sommes intéressés, dans un deuxième temps, à la problématique de l'adaptation du modèle de langage de systèmes de reconnaissance automatique de la parole (SRAP). Pour adapter le modèle de langage, nous émettons l'hypothèse que l'enseignant collabore, a minima, en fournissant les diapositives de son cours. L'idée est d'utiliser les titres de diapositives comme des requêtes. Ces requêtes sont soumises à un moteur de recherche, le contenu des pages web étant alors utilisé pour l'adaptation.

Le WER est la métrique classiquement utilisée pour évaluer la qualité des sorties des systèmes de reconnaissance de la parole. Toutefois, cette mesure considère toutes les erreurs sur les mots comme ayant chacune la même importance. Dans un contexte applicatif, il apparaît clairement que les erreurs sur les mots transcrits automatiquement, par exemple utilisés comme source d'information d'autres systèmes (traduction automatique, indexation, recherche documentaire, etc.), n'ont pas le même impact. De nombreuses études ont démontré des in-

consistances entre les mesures données par le WER et les performances obtenues au niveau de l'application globale. C'est pour cette raison que nous avons exploré deux types d'évaluation : une évaluation extrinsèque et une évaluation intrinsèque. Pour l'évaluation intrinsèque, nous avons proposé une évolution de la métrique $IWER$, au travers de l'$IWER_{Average}$, qui permet d'évaluer la performance de la reconnaissance pour les mots d'un domaine dans un corpus multi-domaines. Pour l'évaluation extrinsèque, nous avons proposé deux métriques pour calculer la performance de la recherche d'information (taux de couverture) et de l'indexabilité (rang de corrélation de Spearman). En comparant le WER avec ces trois métriques, nous avons pu montrer que cette mesure ne permet pas de bien évaluer les performances pour de tâches applicatives, masquant alors les gains réels apportés par l'adaptation des modèles de langage. Cela illustre la nécessité d'utiliser de nouvelles mesures, telles que celles présentées dans ce manuscrit, pour évaluer notamment l'impact réel de l'adaptation des modèles de langage.

Étant donné que notre adaptation s'appuie sur des données provenant du web et, que celui-ci est connu pour son caractère évolutif, nous avons étudié, dans un troisième temps, la question de la reproductibilité des résultats d'adaptation sur une période temporelle. Pour ce faire, un ensemble d'expériences diachroniques ont été réalisées à partir de données collectées mensuellement. Ainsi, sur une période de collecte d'un an (octobre 2018 à octobre 2019), nous avons pu montrer que, même si les données sur le web évoluent en partie d'un mois à l'autre, le changement de résultats n'est pas significatif.

Finalement, notre dernière contribution concerne la structuration automatique de la transcription, comprenant alors une segmentation thématique ainsi qu'un alignement des diapositives avec la transcription. En intégrant l'information de changement de diapositives dans une méthode de segmentation thématique de l'état de l'art, nous avons pu améliorer les performances de l'algorithme en donnant plus de poids, lors du calcul de la similarité, aux blocs proches d'un changement de diapositive. Dans un second temps, à partir des segments de parole obtenus lors de la transcription et des diapositives, nous avons développé une technique d'alignement de la transcription avec les diapositives. Pour en faire, nous avons calculé la similarité cosinus entre la représentation TF-IDF des segments de transcription et la représentation TF-IDF des diapositives de texte, et nous avons imposé une contrainte pour respecter l'ordre séquentiel des diapositives et des segments de transcription. En plus de la mesure de précision traditionnelle, nous considérons l'erreur quadratique moyenne qui offre un score d'évaluation plus fin en distinguant différemment les dés-alignements selon la distance entre la référence et l'hypothèse. Enfin, nous avons considéré une mesure de confiance pour appréhender la fiabilité de l'approche proposée.

Perspectives

Un certain nombre de perspectives peuvent être envisagées afin de poursuivre les travaux présentés dans cette thèse. Nous donnons ici quelques pistes pour aller plus loin dans la continuité de ces travaux.

1. **Adaptation des modèles de langage**
   Bien que les contributions majeures que nous avons menées durant cette thèse s'appuient sur des modèles de langage n-grammes, nous avons vu, avec la compagne MGB (Annexe B), que les modèles de langage neuronaux permettent d'améliorer les performances. Dans notre contexte précis d'étude, à savoir la mise en place d'un système de transcription temps-réel dans le cadre du projet PASTEL, l'utilisation des MLs neuronaux n'avait pas été envisagée de part la latence introduite par cette opération de rescoring. Néanmoins, s'il l'on considère que la transcription de cours peut être réalisée a posteriori par les étudiants (par exemple lors des révisions), la contrainte de latence ne s'appliquerait plus. Il serait alors intéressant d'explorer l'utilisation de réseaux neuronaux, notamment en travaillant sur les modèles de langage neuronaux et en exploitant leurs architectures. Nous pourrions par exemple intégrer des informations provenant des diapositives pour l'adaptation des modèles de langage. La figure 8.4 illustre le principe d'adaptation du modèle de langage en utilisant des informations issues de diapositives. De manière plus précise, l'idée consiste à extraire un vecteur TF-IDF à partir de chaque diapositive du cours. Vu que la dimension des vecteurs TF-IDF peut être énorme, nous envisageons l'utilisation d'un auto-encodeur pour diminuer la dimension des vecteurs TF-IDF. Enfin, l'adaptation consiste à intégrer le vecteur fourni par l'auto-encodeur dans l'apprentissage du modèle neuronal : Feature based adaptation (section 2.3.2).

2. **Évaluation des SRAP**
   Pour l'évaluation extrinsèque, nous avons exploré deux types de tâches : la recherche d'informations et l'indexabilité. L'exploration d'autres contextes applicatifs similaires reste aussi possible dans le cadre du projet PASTEL. Nous pensons particulièrement à tout ce qui est traduction automatique et les systèmes questions/réponses.

3. **L'étude diachronique de l'adaptation du modèle de langage**
   L'étude diachronique a été réalisée pour une période temporelle d'un an, d'octobre 2018 à octobre 2019. Il est important de continuer l'analyse sur une période plus longue. Dans un article récent (publié en décembre 2019) du journal *Le Monde*[2] (figure 8.5), les auteurs ont annoncé la modification du moteur de recherche Google pour le français.

2. https ://www.lemonde.fr/pixels/article/2019/12/09/google-modifie-l-algorithme-de-son-moteur-de-recherche-en-francais_6022176_4408996.html

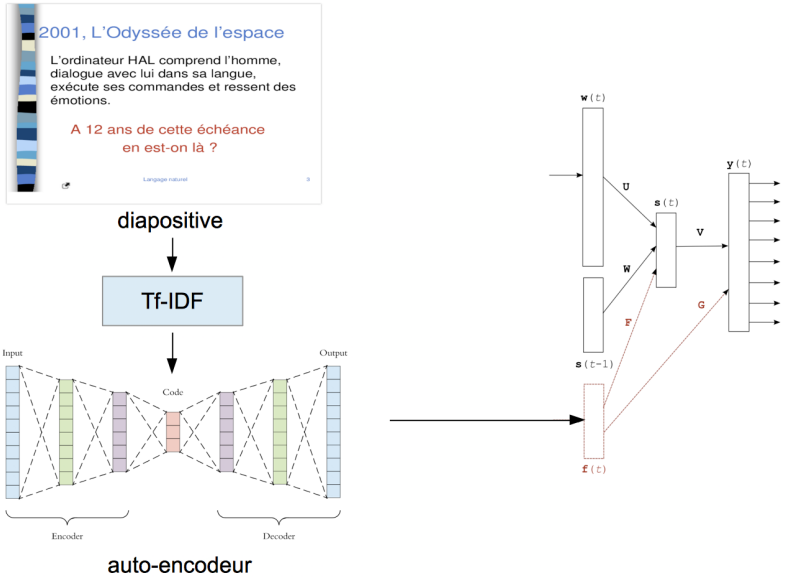


FIGURE 8.4 – Principe de l'adaptation du modèle de langage en utilisant des informations de diapositives

Ceci nous encourage à continuer l'étude afin de comparer les résultats avant et après cette modification.

Il pourrait également être intéressant d'évaluer l'impact de l'adaptation du ML sur un contexte de nature différente (par exemple les actualités) et aussi en utilisant d'autres moteurs de recherche tels que Yahoo, Bing et Qwant.

4. **Segmentation thématique**
   Appliquer nos contributions sur d'autres algorithmes de segmentation thématique. La méthode proposée repose sur l'intégration de l'information de changement de diapositives dans le calcul de la similarité. Le calcul de la similarité est une étape indispensable dans la plupart des algorithmes de segmentation s'appuyant sur le calcul de la cohésion lexicale. Alors, il est intéressant d'expérimenter cette méthodologie avec d'autres algorithmes.

Initiés dans le cadre du projet PASTEL, les travaux réalisés dans cette thèse ont principalement visé l'utilisation de la transcription automatique dans un contexte de cours magistraux. Cependant, ils ont un caractère plus général et sont très facilement portables vers d'autres types d'applications (conférences, séminaires, tutoriaux...). Il est également très probable que ces travaux puissent trouver un écho dans l'industrie, pour comparer des systèmes de reconnaissance de la parole, aligner automatiquement ou semi-automatiquement des diapositives sur des enregistrements audio ou vidéo, ou d'adapter le modèle de langage des systèmes de

FIGURE 8.5 – Capture d'écran de l'article du journal *"Le Monde"* sous le titre *"Google modifie l'algorithme de son moteur de recherche en français"*

reconnaissance de la parole à partir du web même sur différentes périodes temporelles.

Annexe A

Expériences du projet PASTEL

A.1 Étude de Cas : Instrumentation par la Technologie de Transcription de la Parole

Deux situations pédagogiques sont au coeur du projet PASTEL. La première est centrée sur les cours magistraux, la seconde sur le travail collaboratif mené lors de travaux dirigés ou de travaux pratiques. L'équipe EIAH ont développé les environnements d'instrumentation.

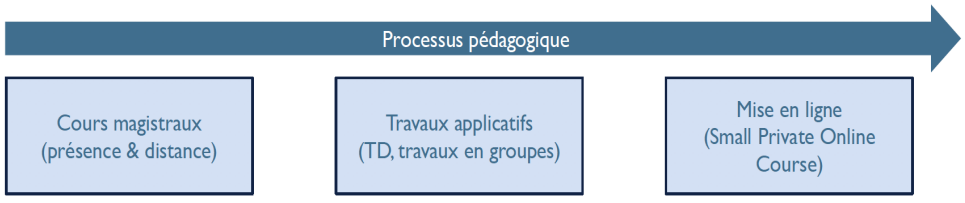


FIGURE A.1 – Processus pédagogique cible du projet PASTEL

A.1.1 Interface utilisée lors des cours magistraux

Les étudiants ont accès à la diffusion d'un flux audiovisuel montrant l'enseignant (Figure A.2.a) ainsi que des supports de cours, synchronisés avec ceux projetés en classe (Fig A.2.b). Lors du cours, des liens vers des ressources pertinentes sont recommandés aux étudiants (Figure A.2.c). La transcription est également affichée en temps réel dans une fenêtre restreinte similaire à des sous-titres (Figure A.2.d) qui peut être agrandie pour être lue plus confortablement. Les étudiants peuvent également prendre des notes dans un éditeur de texte (Figure A.2.e) et les exporter à l'issue du cours. La conception et l'architecture sont détaillées dans (BETTENFELD, CHOQUET et PIAU-TOFFOLON, 2018).

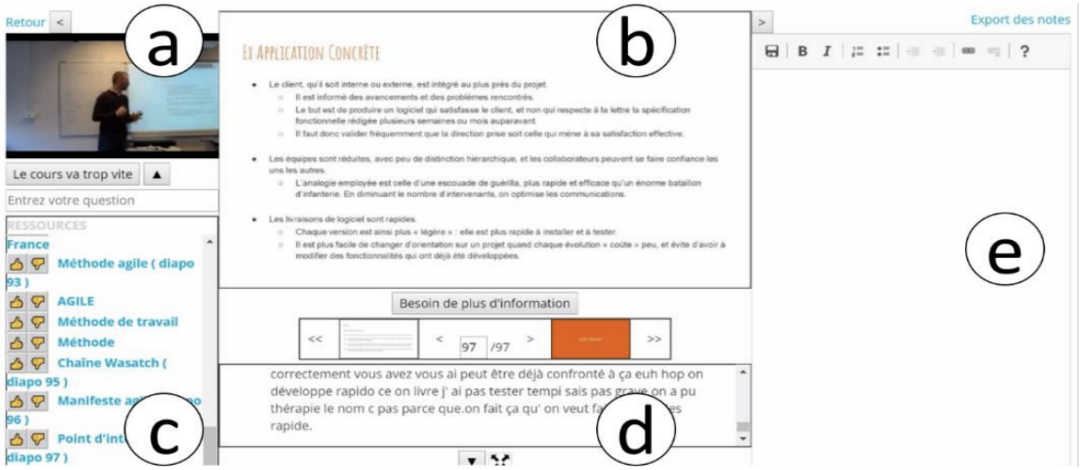


FIGURE A.2 – Interface proposée aux étudiants, comprenant le flux vidéo (a), les diapositives (b), la liste des ressources (c), la transcription automatique (d) ainsi que l'éditeur de texte (e) (BETTENFELD et al., 2019a)

A.1.2 Interface utilisée lors des travaux pratiques

Les étudiants disposent des vidéos des cours magistraux (Figure A.3.b), synchronisées avec leurs transcriptions respectives (Figure A.3.c). Afin de pouvoir naviguer plus facilement dans ces contenus vidéos et textuels, ils sont divisés en chapitres générés automatiquement (Figure A.3.d).

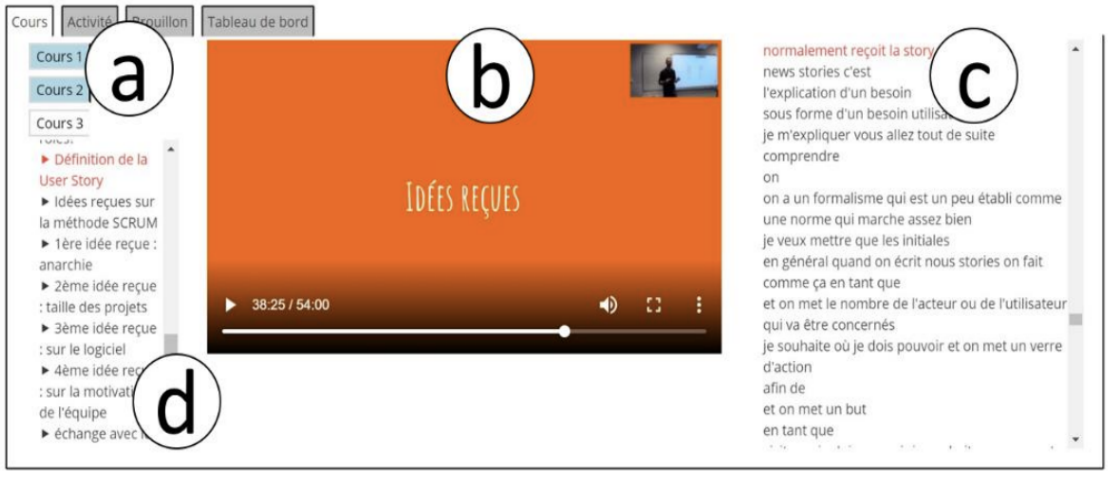


FIGURE A.3 – Interface proposée lors des travaux pratiques. L'un des onglets (a) permet d'afficher la vidéo correspondant à un cours (b), la transcription correspondant (c), ainsi que la liste des chapitres (d) (BETTENFELD et al., 2019a)

A.2 Expériences du projet PASTEL

Dans le cadre du projet PASTEL, trois expériences ont été réalisées dans un cadre d'expérimentation réel (en direct dans un cours avec un vrai enseignement et de vrais étudiants). Notre contribution à ces deux expériences repose sur l'adaptation d'un modèle de langage générique au domaine du cours. Cette adaptation repose sur le même principe présenté dans la section 5.6.3.

A.2.1 Expérience 1 (BETTENFELD et al., 2018)

— Date : 12 décembre 2017
— Place : Campus du Mans et de Nantes
— Nombre d'étudiants : 3 (Campus du Mans), 7 à distance (Campus de Nantes)
— Titre du cours : Traduction automatique

A.2.2 Expérience 2 (BETTENFELD, CRETIN-PIROLLI et CHOQUET, 2018)

— Date : 22 Mars 2018
— Place : Campus du Mans et de Nantes
— Nombre d'étudiants : 13, 16 à distance
— Titre du cours : Information et communication

A.2.3 Expérience 3

— Date : Janvier 2019
— Place : Campus de Laval
— Nombre d'étudiants : 23 étudiants
— Titre du cours : Gestion de projet

Annexe B

Étude des modèles neuronaux pour la modélisation linguistique (MGB)

Le Multi-Genre Broadcast Challenge [1] est une campagne d'évaluation pour la reconnaissance de la parole, la segmentation et le regroupement en locuteurs, et l'alignement semi-supervisé. Le LIUM a participé à la tâche de transcription de parole en anglais. Nous décrivons dans cette annexe notre contribution pour la modélisation de langage dans le cadre de MGB.

B.1 Description du système de reconnaissance de parole

Le SRAP utilisé pour les expériences présentées dans cette section est un système préliminaire développé dans le cadre de la campagne d'évaluation MGB 2017.

Un premier système consiste en un modèle triphone HMM-GMM avec des transformations indépendantes du locuteur appliquées aux caractéristiques MFCC-LDA-MLLT. Une technique de perturbation de la vitesse (KO et al., 2015) a été utilisée pour multiplier par trois la quantité des données d'apprentissage.

Ensuite, un modèle de type chain-TDNN (Lattice-free MMI TDNN (POVEY et al., 2016)) avec un apprentissage discriminant visant à minimiser le risque bayésien sur les états (sMBR (KINGSBURY, SAINATH et SOLTAU, 2012)) a été réalisé. Des i-vecteurs ont également été utilisés pour l'adaptation des modèles acoustiques neuronaux (SAON et al., 2013).

Les phonétisations des mots sont obtenues à l'aide d'un lexique construit manuellement, dérivé de Combilex, fourni par les organisateurs de la campagne d'évaluation. Combilex [2] est un lexique pour l'anglais, développé spécifiquement pour les technologies de la parole comme la synthèse de la parole et la reconnaissance automatique.

1. http ://www.mgb-challenge.org/
2. http ://homepages.inf.ed.ac.uk/korin/sitenew/Research/Combilex/index.html

B.2 Description des données MGB

Cette partie décrit les données utilisées pour implémenter le système de reconnaissance de parole dans le cadre de la compagne d'évaluation MGB.

B.2.1 Données pour le modèle acoustique

Les données fournies par la compagne d'évaluation MGB 2017 comprennent environ 328 heures d'audio enregistrées sur sept semaines à partir de toutes les chaînes de télévision de la BBC (BBC1, BBC2, BBC3, BBC4, etc.). Les données couvrent une grande variété de genres (documentaires, actualités, drames ,etc.).

Le corpus de développement contient 4 heures de parole qui ne pas incluses dans les données d'apprentissage. La campagne MGB n'a jamais fini, les données de test ne sont pas fournies. Par conséquent, nous avons éliminé les données de développement du corpus d'apprentissage et nous les avons utilisées pour évaluer notre système.

Les statistiques des données d'apprentissage et des données de développement sont présentées dans le tableau B.1.

TABLE B.1 – Statistiques du corpus d'apprentissage et corpus de développement MGB3

Corpus	**Corpus d'apprentissage**	**Corpus de développement**
# segments	237K	5856
# locuteurs	2719	302
Durée	324h	4h37

B.2.2 Données pour le modèle de langage

Pour la modélisation linguistique, nous avons utilisé les données fournies par la campagne d'évaluation. Le nombre de mots et le vocabulaire de ces données sont présentés dans le tableau B.2.

TABLE B.2 – Statistiques du corpus d'apprentissage pour la modélisation linguistique

	Corpus d'apprentissage
# mots total	645758K
Vocabulaire	757K
Vocabulaire utilisé pour les ML n-grammes	164K

B.3 Implémentation des modèles de langage

Afin de comparer la performance du SRAP, plusieurs ML ont été construits. Nous décrivons dans cette section les différentes implémentions réalisées (MDHAFFAR, LAURENT et ESTÈVE, 2018a).

B.3.1 Implémentation des modèles de langage neuronaux

Trois modèles neuronaux ont été implémentés : LSTM, GRU, GRU-Highway. L'outil CUED-RNNLM [3] (CHEN et al., 2016) a été utilisé. CUED-RNNLM est un outil "open source" destiné à la modélisation du langage avec les réseaux de neurones. Il comprend plusieurs types de réseaux de neurones tels que le modèle de langage neuronal "feed forward" et plusieurs variétés de RNN. La boîte à outils CUED-RNNLM fournit aussi des recettes pour diverses fonctions, notamment l'évaluation de la perplexité, la ré-évaluation de N-best, etc.

Afin de garantir une évaluation équitable entre les différents modèles, nous avons utilisé les mêmes réglages et paramètres pour tous les modèles RNN. Le nombre de couches cachées est 2. La taille de la couche cachée est de 200. La taille du vocabulaire de sortie (shortlist) est de 30K mots. Un modèle avec 60K mots a également été implémenté pour évaluer l'impact de l'augmentation de la taille du vocabulaire.

B.3.2 Implémentation d'un RNN arrière

Généralement, un modèle de langage est appris dans le sens de l'écrit de la langue : de gauche à droite dans le cas d'anglais. Cependant, dans le cadre de la modélisation linguistique, les informations du futur peuvent peut-être aider un modèle à estimer la probabilité d'apparition d'un mot. Avec cette hypothèse, des modèles de langage récurrents dans lesquels l'ordre des mots a été inversé "RNN arrière" ont été implémentés.

Un RNN arrière est similaire à un RNN classique à l'exception que dans l'apprentissage du modèle la phrase est donnée à l'envers : le premier mot devient le dernier et le dernier mot devient le premier. Par conséquent, un modèle de langage à l'arrière estime la probabilité d'un mot sachant le contexte futur $P(w_\alpha|w_{\alpha+1}, ..., w_{\alpha+n})$ où α est l'indice du mot courant et n est le nombre de mots dans le contexte.

B.3.3 Implémentation des modèles de langage N-grammes

Quatre modèles de langage n-grammes ont été implémentés : un modèle de langage 3-grammes, un modèle de langage 4-grammes, un modèle de langage 3-grammes arrière et

3. http ://mi.eng.cam.ac.uk/projects/cued-rnnlm/

un modèle de langage 4-grammes arrière. Le modèle de langage 3-grammes va servir au décodage du SRAP durant la première passe pour générer la liste de N-best. Il va être utilisé aussi comme base pour évaluer la performance sans et avec utilisation des réseaux neuronaux.

Les deux modèles 3-grammes et les deux 4-grammes vont servir par la suite pour effectuer une interpolation linéaire avec les modèles de langage neuronaux.

B.3.4 Processus de transcription de la parole

Le processus de transcription est composé de deux passes :

— La première passe utilise le modèle de langage 3-grammes afin de générer la liste de N-best.

— La liste de N-best obtenue par la première passe est ré-évaluée dans une deuxième passe avec les modèles de langage neuronaux. Le ré-évaluation a été effectué grâce à l'outil CUED-RNNLM. Nous avons modifié l'outil de rescoring afin qu'il soit capable d'utiliser des modèles de langage avant et arrière. Les modifications apportées impliquent une interpolation des résultats entre les modèles avant et les modèles arrière. La figure B.1 illustre dans la partie droite l'architecture du rescoring avant et arrière.

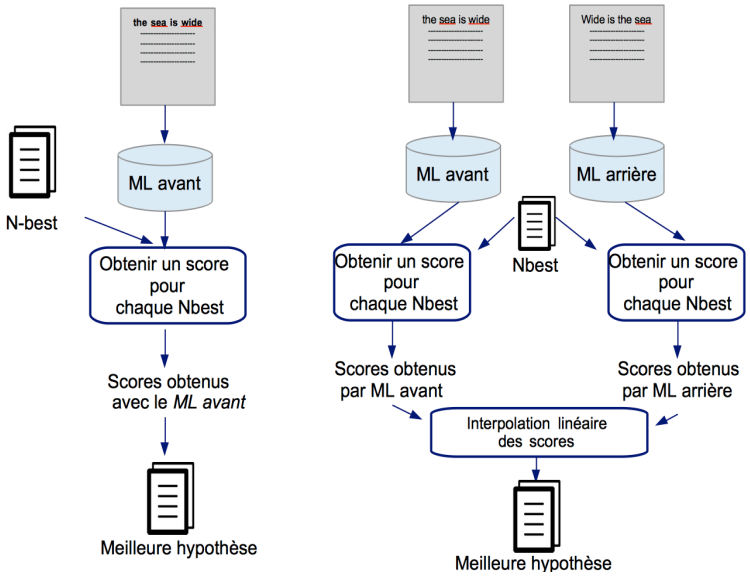


FIGURE B.1 – Rescoring avec des modèles de langage neuronaux en utilisant un modèle classique avant (gauche) et une interpolation des résultats entre les modèles avant et les modèles arrière

B.4 Résultats et discussion

Le tableau B.3 présente les différents résultats expérimentaux. La troisième colonne présente les résultats en terme de WER. La colonne 4 (δ) présente le gain absolu par rapport au système de base sans effectuer la deuxième passe (1-best). Les poids d'interpolation utilisés sont 0,5 dans le cas d'interpolation de deux modèles et 0,25 dans le cas d'interpolation de quatre modèles.

TABLE B.3 – Résultats obtenus en utilisant différents modèles de langage neuronaux

		WER %	δ
1	3-grammes (système de base)	24,7	-
2	LSTM	22,6	2,1
3	GRU	22,5	2,2
4	GRU-Highway	22,3	2,4
5	LSTM + 3-grammes	22,1	2,6
6	GRU + 3-grammes	22,0	2,7
7	GRU + 4-grammes	21,7	3
8	GRU-Highway + 3-grammes	21,6	3,1
9	GRU + 3-grammes (|shortlist|= 60K)	21,8	2,9
10	GRU + 4-grammes (|shortlist|= 60K)	21,6	3,1
11	GRU arrière	22,5	2,2
12	GRU arrière + 3-grammes arrière + GRU avant + 3-grammes avant	21,6	3,1
13	GRU arrière + 4-grammes arrière + GRU avant + 4-grammes avant	21,4	3,3

Les résultats obtenus par les différents modèles neuronaux ((2) LSTM, (3) GRU et (4) GRU-Highway) donnent de meilleures performances comparées à la performance obtenue par le système de base (1). Les résultats expérimentaux de ces trois modèles montrent que le GRU-Highway a donné un résultat plus performant que le LSTM et le GRU en terme de WER (22,3% WER pour le GRU-Highway, 22,6% WER pour le LSTM, 22,5% WER pour le GRU).

L'interpolation de ces modèles neuronaux avec des modèles n-grammes a permis d'améliorer les performances du système (réduction de 0,5% absolu en WER dans le cas de LSTM et GRU et réduction de 0,6% absolu en WER dans le cas GRU-Highway). De plus, l'interpolation avec un modèle 4-grammes (7) donne de meilleurs résultats que l'interpolation avec un modèle 3-grammes (réduction de 0,5% absolu en WER dans le cas de LSTM).

En ré-évaluant les n-best avec le GRU arrière (11) sans interpolation, on obtient les mêmes performances qu'avec le GRU classique (3).

Les résultats dans les lignes (12) et (13) prouvent tout l'intérêt d'utiliser à la fois un contexte gauche et un contexte droit au niveau de la modélisation linguistique. L'interpolation des modèles avant et arrière a amélioré significativement les résultats en termes de WER (de 22% (6) à 21,6% (12) en terme de WER dans le cas de 3-grammes et de 21,7% (7) à 21,4% (13) en

terme de WER dans le cas de 4-grammes).

Nous avons également montré dans cette étude que l'utilisation de 60000 mots dans la shortlist (9,10) a donné de meilleurs résultats par rapport à l'utilisation de 30000 mots (6,7).

Acronymes

ADSL Asymmetric Digital Subscriber Line
ANR Agence Nationale de la Recherche
BIC Bayesian Information Criterion
BN Bottleneck
CHIL Computers In the Human Interaction Loop
CNN Convolutional Neural Network
CREN Centre de Recherche en Education de Nantes
CRF Conditional Random Fields
COCo Comin Open Courseware
CSAIL Computer Science and Artificial Intelligence Laboratory
DER Diarization Error Rate
DNN Deep Neural Networks
EIAH Environnements Informatiques pour l'Apprentissage Humain
ELAN EUDICO Linguistic Annotator
EMMA European Multiple MOOC Aggregator
ENT Environnements Numériques de Travail
GMM Gaussian Mixture Models
GRU Gated Recurrent Unit
HMM Hidden Markov Model
IWER Individual Word Error Rate
IDF Inverted Document Frequency
LDA Linear Discriminant Analysis
LIUM Laboratoire d'Informatique de l'Université du Maine
LHN Linear Hidden Layer
LPC Linear Predictive Codes
LS2N Laboratoire des Sciences du Numérique de Nantes

LSA Latent Semantic Analysis
LST Language and Speech Technology
LSTM Long Short-Term Memory
MAP Maximum a posteriori
MDI minimum d'information discriminante
MFCC Mel Frequency Cepstral Coefficients
MGB Multi-Genre Broadcast Challenge
ML Modèle de Langage
MMC Modèle de Markov Caché
MOOC Massive Open Online Course
MSE Mean Square Error
Net4voice New technologies for Voice-converting in barrier-free learning environments
OOV Out Of Vocabulary
PASTEL Performing Automatic Speech Transcription for Enhanced Learning
PLP Perceptual Linear Prediction
LPCC Linear Predictive Cepstral Coefficients
PPL PerPLexité
pLSA Probabilistic Latent Semantic Analysis
RASTA-PLP Relative Spectral PLP
SRAP Système de Reconnaissance Automatique de la Parole
SVM Support Vector Machines
TALN Traitement Automatique du Langage Naturel
TF Term Frequency
TraMOOC Translation for Massive Open Online Courses
TransLectures Transcription and Translation of Video Lectures
WD WindowDiff
WER Word Error Rate
WLL Weighted Lexical Links

Références personnelles

Références personnelles liées à la thèse

BETTENFELD, Vincent, Salima MDHAFFAR, Christophe CHOQUET et Claudine PIAU-TOFFOLON (2018). « Instrumentation of Classrooms Using Synchronous Speech Transcription ». In : *European Conference on Technology Enhanced Learning*. Springer, p. 648–651.

BETTENFELD, Vincent, Salima MDHAFFAR, Claudine PIAU-TOFFOLON et Christophe CHOQUET (2019b). « Instrumentation of learning situation using automated speech transcription : A prototyping approach ». In : *11th International Conf on Computer Supported Education (CSEDU 2019)*.

MDHAFFAR, Salima, Antoine LAURENT et Yannick ESTÈVE (2018a). « Etude de performance des réseaux neuronaux récurrents dans le cadre de la campagne d'évaluation Multi-Genre Broadcast challenge 3 (MGB3) ». In : *XXXIIe Journées d'Etudes sur la Parole (JEP 2018)*.

MDHAFFAR, Salima, Antoine LAURENT et Yannick ESTÈVE (2018b). « Le corpus PASTEL pour le traitement automatique de cours magistraux ». In : *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018)*.

MDHAFFAR, Salima, Yannick ESTÈVE, Nicolas HERNANDEZ, Antoine LAURENT et Solen QUINIOU (2019a). « Apport de l'adaptation automatique des modèles de langage pour la reconnaissance de la parole : évaluation qualitative extrinsèque dans un contexte de traitement de cours magistraux ». In : *26e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2019)*.

MDHAFFAR, Salima, Yannick ESTÈVE, Nicolas HERNANDEZ, Antoine LAURENT, Richard DUFOUR et Solen QUINIOU (2019b). « Qualitative evaluation of ASR adaptation in a lecture context : Application to the PASTEL corpus ». In : *Proc. Interspeech 2019*, p. 569–573.

MDHAFFAR, Salima, Yannick ESTÈVE, Antoine LAURENT, Nicolas HERNANDEZ, Richard DUFOUR, Charlet DELPHINE, Damnati GÉRALDINE, Solen QUNIOU et Camelin NATHALIE (2020).

« A Multimodal Educational Corpus of Oral Courses : Annotation, Analysis and Case Study ». In : *LREC*.

Références personnelles non liées à la thèse

MASMOUDI, Abir, Salima MDHAFFAR, Rahma SELLAMI et Lamia Hadrich BELGUITH (2019). « Automatic diacritics restoration for Tunisian dialect ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18.3, p. 28.

MDHAFFAR, Salima, Fethi BOUGARES, Yannick ESTÈVE et Lamia HADRICH-BELGUITH (2017). « Sentiment analysis of Tunisian dialects : Linguistic ressources and experiments ». In : *Proceedings of the third Arabic natural language processing workshop*, p. 55–61.

Références

ABDELALI, Ahmed, Francisco GUZMAN, Hassan SAJJAD et Stephan VOGEL (2014). « The AMARA Corpus : Building Parallel Language Resources for the Educational Domain ». In : *Language Resources and Evaluation Conference (LREC)*. T. 14, p. 1044–1054.

ABDULLAH, Badr, Irina ILLINA et Dominique FOHR (2018). « Dynamic Extension of ASR Lexicon Using Wikipedia Data ». In : *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, p. 36–42.

ABE, M (1992). « A Study on Speaker Individuality Control ». Thèse de doct. NTT Human Interface Laboratories.

AKITA, Yuya, Yizheng TONG et Tatsuya KAWAHARA (2015). « Language model adaptation for academic lectures using character recognition result of presentation slides ». In : *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 5431–5435.

ALHARBI, Ghada et Thomas HAIN (2015). « Using Topic Segmentation Models for the Automatic Organisation of MOOCs resources. » In : *International Conference on Educational Data Mining (EDM)*, p. 524–527.

ALLAUZEN, Alexandre et Jean-Luc GAUVAIN (2003). « Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés : Modélisation probabiliste du langage naturel ». In : *TAL. Traitement automatique des langues* 44.1, p. 11–31.

ALLAUZEN, Alexandre et Jean-Luc GAUVAIN (2005). « Diachronic vocabulary adaptation for broadcast news transcription ». In : *Ninth European Conference on Speech Communication and Technology*.

ALUMÄE, Tanel et Mikko KURIMO (2010). « Domain adaptation of maximum entropy language models ». In : *Association for Computational Linguistics (ACL)*, p. 301–306.

ARANSA, Walid, Holger SCHWENK et Loic BARRAULT (2015). « Improving continuous space language models using auxiliary features ». In : *Proceedings of the 12th International Workshop on Spoken Language Translation*, p. 151–158.

ARNOLD, Sebastian, Rudolf SCHNEIDER, Philippe CUDRÉ-MAUROUX, Felix A GERS et Alexander LÖSER (2019). « SECTOR : A Neural Model for Coherent Topic Segmentation and Classification ». In : *Transactions of the Association for Computational Linguistics* 7, p. 169–184.

AUBERT, Olivier et Joscha JAEGER (2014). « Annotating Video with Open Educational Resources in a Flipped Classroom Scenario ». In : *OCWC Global Conference*.

AUBERT, Olivier, Yannick PRIÉ et Camila CANELLAS (2014). « Leveraging Video Annotations in Video-based e-Learning ». In : *Proceedings of the 6th International Conference on Computer Supported Education*, p. 479–485.

AUER, Eric, Albert RUSSEL, Han SLOETJES, Peter WITTENBURG, Oliver SCHREER, Stefano MASNIERI, Daniel SCHNEIDER et Sebastian TSCHÖPEL (2010). « ELAN as flexible annotation framework for sound and image processing detectors ». In : *Seventh conference on International Language Resources and Evaluation [LREC 2010]*. European Language Resources Association (ELRA), p. 890–893.

AUZANNE, Cédric, John S GAROFOLO, Jonathan G FISCUS et William M FISHER (2000). « Automatic language model adaptation for spoken document retrieval ». In : *Content-Based Multimedia Information Access-Volume 1*, p. 132–141.

BACCHIANI, Michiel, Michael RILEY, Brian ROARK et Richard SPROAT (2006). « MAP adaptation of stochastic grammars ». In : *Computer speech & language* 20.1, p. 41–68.

BADJATIYA, Pinkesh, Litton J KURISINKEL, Manish GUPTA et Vasudeva VARMA (2018). « Attention-based neural text segmentation ». In : *European Conference on Information Retrieval*. Springer, p. 180–193.

BAIN, Keith, Sara H BASSON et Mike WALD (2002). « Speech recognition in university classrooms : liberated learning project ». In : *Proceedings of the fifth international ACM conference on Assistive technologies*. ACM, p. 192–196.

BALAGOPALAN, Arun, Lalitha Lakshmi BALASUBRAMANIAN, Vidhya BALASUBRAMANIAN, Nithin CHANDRASEKHARAN et Aswin DAMODAR (2012). « Automatic keyphrase extraction and segmentation of video lectures ». In : *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*. IEEE, p. 1–10.

BALASUBRAMANIAN, Vidhya, Sooryanarayan Gobu DORAISAMY et Navaneeth Kumar KANAKARAJAN (2016). « A multimodal approach for extracting content descriptive metadata from lecture videos ». In : *Journal of Intelligent Information Systems* 46.1, p. 121–145.

BARRAS, Claude, Edouard GEOFFROIS, Zhibiao WU et Mark LIBERMAN (1998). « Transcriber : a free tool for segmenting, labeling and transcribing speech ». In : *First international conference on language resources and evaluation (LREC)*, p. 1373–1376.

BARRAS, Claude, Edouard GEOFFROIS, Zhibiao WU et Mark LIBERMAN (2001). « Transcriber : development and use of a tool for assisting speech corpora production ». In : *Speech Communication* 33.1-2, p. 5–22.

BAUM, Leonard (1972). « An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process ». In : *Inequalities* 3, p. 1–8.

BÉCHET, Frédéric (2001). « LIA—PHON : Un système complet de phonétisation de textes ». In : *TAL. Traitement automatique des langues* 42.1, p. 47–67.

BEEFERMAN, Doug, Adam BERGER et John LAFFERTY (1997). « Text Segmentation Using Exponential Models ». In : *Second Conference on Empirical Methods in Natural Language Processing*.

BEHNKE, Maximiliana, Antonio Valerio MICELI BARONE, Rico SENNRICH, Vilelmini SOSONI, Thanasis NASKOS, Eirini TAKOULIDOU, Maria STASIMIOTI, van Zaanen MENNO, Sheila CASTILHO, Federico GASPARI et al. (2018). « Improving machine translation of educational content via crowdsourcing ». In : *International Language Resources and Evaluation (LREC)*. European Language Resource Association.

BELL, Peter, Hitoshi YAMAMOTO, Pawel SWIETOJANSKI, Youzheng WU, Fergus MCINNES, Chiori HORI et Steve RENALS (2013). « A lecture transcription system combining neural network acoustic and language models. » In : *INTERSPEECH*, p. 3087–3091.

BELLEGARDA, Jerome R (2000). « Exploiting latent semantic information in statistical language modeling ». In : *Proceedings of the IEEE* 88.8, p. 1279–1296.

BELLEGARDA, Jerome R (2004). « Statistical language model adaptation : review and perspectives ». In : *Speech communication* 42.1, p. 93–108.

BEN JANNET, Mohamed Ameur (2015). « Évaluation adaptative des systèmes de transcription en contextes applicatifs ». Thèse de doct. PhD Thesis, University Paris Sud.

BEN JANNET, Mohamed Ameur, Olivier GALIBERT, Martine ADDA-DECKER et Sophie ROSSET (2015). « How to evaluate ASR output for named entity recognition ? » In : *Sixteenth Annual Conference of the International Speech Communication Association*.

BEN JANNET, Mohamed Ameur Ben, Martine ADDA-DECKERδ, Olivier GALIBERTγ, Juliette KAHNγ et Sophie ROSSETα (2014). « ETER : a new metric for the evaluation of hierarchical named entity recognition ». In : *Ninth International Conference on Language Resources and Evaluation (LREC 14)*.

BENGIO, Yoshua, Réjean DUCHARME, Pascal VINCENT et Christian JAUVIN (2003). « A neural probabilistic language model ». In : *Journal of machine learning research* 3.Feb, p. 1137–1155.

BETTENFELD, Vincent, Christophe CHOQUET et Claudine PIAU-TOFFOLON (2018). « Lecture instrumentation based on synchronous speech transcription ». In : *International Conference on Advanced Learning Technologies (ICALT)*. IEEE, p. 11–15.

BETTENFELD, Vincent, Raphaëlle CRETIN-PIROLLI et Christophe CHOQUET (2018). « PASTEL : un environnement outillé exploitant la reconnaissance de la parole dans les situations de cours ». In : *IHM*.

BETTENFELD, Vincent, Salima MDHAFFAR, Christophe CHOQUET et Claudine PIAU-TOFFOLON (2018). « Instrumentation of Classrooms Using Synchronous Speech Transcription ». In : *European Conference on Technology Enhanced Learning*. Springer, p. 648–651.

BETTENFELD, Vincent, Raphaëlle CRETIN-PIROLLI, Claudine PIAU-TOFFOLON et Christophe CHOQUET (2019a). « Elaboration d'une méthodologie d'instrumentation pédagogique en contexte universitaire ». In : *9 ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 19)*.

BETTENFELD, Vincent, Salima MDHAFFAR, Claudine PIAU-TOFFOLON et Christophe CHOQUET (2019b). « Instrumentation of learning situation using automated speech transcription : A prototyping approach ». In : *11th International Conf on Computer Supported Education (CSEDU 2019)*.

BIGI, Brigitte, Renato DE MORI et Spriet THIERRY (2000). « Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langages thématiques ». In : *Journées d'Etudes sur la Parole (JEP)*.

BLEI, David M, Andrew Y NG et Michael I JORDAN (2003). « Latent dirichlet allocation ». In : *Journal of machine Learning research* 3.Jan, p. 993–1022.

BOUCHEKIF, Abdessalam, Géraldine DAMNATI, Yannick ESTEVE, Delphine CHARLET et Nathalie CAMELIN (2015). « Diachronic semantic cohesion for topic segmentation of tv broadcast news ». In : *Sixteenth Annual Conference of the International Speech Communication Association*.

BOUCHEKIF, Abdesselam (2016). « Structuration automatique de documents audio ». Thèse de doct. Le Mans.

BOUGARES, Fethi (2012). « Attelage de systèmes de transcription automatique de la parole ». Thèse de doct. Le Mans.

BOURLARD, Herve et Christian J WELLEKENS (1989). « Links between Markov models and multilayer perceptrons ». In : *Advances in neural information processing systems*, p. 502–510.

BOURLARD, Hervé, Marc FERRAS, Nikolaos PAPPAS, Andrei POPESCU-BELIS, Steve RENALS, Fergus MCINNES, Peter J BELL, Sandy INGRAM et Maël GUILLEMOT (2013). « Processing and linking audio events in large multimedia archives : The eu inevent project ». In : *First Workshop on Speech, Language and Audio in Multimedia*.

BROWN, Gillian, Gillian D BROWN, Gillian R BROWN, Brown GILLIAN et George YULE (1983). *Discourse analysis*. Cambridge university press.

BROWN, Peter F, Peter V DESOUZA, Robert L MERCER, Vincent J Della PIETRA et Jenifer C LAI (1992). « Class-based n-gram models of natural language ». In : *Computational linguistics* 18.4, p. 467–479.

CANELLAS, Camila, Olivier AUBERT et Yannick PRIÉ (2015). « Prise de note collaborative en vue d'une tâche : une étude exploratoire avec COCoNotes Live ». In :

CERDÀ, Silvestre, Joan ALBERT, Miguel Angel DEL AGUA TEBA, Gonzalo Vicente GARCÉS DÍAZ-MUNÍO, GUILLEM GASCÓ MORA, Adrián GIMÉNEZ PASTOR, Adrià Agustí MARTÍNEZ-VILLARONGA, Alejandro Manuel Pérez González de MARTOS, Isaías SÁNCHEZ-CORTINA, Nicolás SERRANO MARTÍNEZ-SANTOS et al. (2012). « TransLectures ». In : *IberSPEECH 2012-VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*. IberSPEECH 2012, p. 345–351.

CERVA, Petr, Jan SILOVSKY, Jindrich ZDANSKY, Jan NOUZA et Jiri MALEK (2012). « Real-time lecture transcription using ASR for czech hearing impaired or deaf students ». In : *Thirteenth Annual Conference of the International Speech Communication Association*.

CHELBA, Ciprian, Timothy J HAZEN et Murat SARACLAR (2008). « Retrieval and browsing of spoken content ». In : *IEEE Signal Processing Magazine* 25.3, p. 39–49.

CHELBA, Ciprian et Frederick JELINEK (2000). « Structured language modeling ». In : *Computer Speech & Language* 14.4, p. 283–332.

CHELBA, Ciprian, David ENGLE, Harry PRINTZ, Frederick JELINEK, Eric RISTAD, Victor JIMENEZ, Ronald ROSENFELD, Sanjeev KHUDANPUR, Andreas STOLCKE, Lidia MANGUE et al. (1997). *Structure and performance of a dependency language model*. Rapp. tech. sri international menlo park ca speech technology et research lab.

CHEN, Langzhou et Taiyi HUANG (1999). « An improved MAP method for language model adaptation ». In : *Sixth European Conference on Speech Communication and Technology*.

CHEN, Stanley F (2009). « Shrinking exponential language models ». In : *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, p. 468–476.

CHEN, Stanley F et Joshua GOODMAN (1999). « An empirical study of smoothing techniques for language modeling ». In : *Computer Speech & Language* 13.4, p. 359–394.

CHEN, Xie, Tian TAN, Xunying LIU, Pierre LANCHANTIN, Moquan WAN, Mark JF GALES et Philip C WOODLAND (2015). « Recurrent neural network language model adaptation for multi-genre broadcast speech recognition ». In : *Sixteenth Annual Conference of the International Speech Communication Association*.

CHEN, Xie, Xunying LIU, Yanmin QIAN, MJF GALES et Philip C WOODLAND (2016). « CUED-RNNLM—An open-source toolkit for efficient training and evaluation of recurrent neural network language models ». In : *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, p. 6000–6004.

CHEN, Yun-Nung, Yu HUANG, Sheng-Yi KONG et Lin-Shan LEE (2010). « Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features ». In : *2010 IEEE Spoken Language Technology Workshop*. IEEE, p. 265–270.

CHO, Eunah, Christian FÜGEN, Teresa HERRMANN, Kevin KILGOUR, Mohammed MEDIANI, Christian MOHR, Jan NIEHUES, Kay ROTTMANN, Christian SAAM, Sebastian STÜKER et al. (2013). « A real-world system for simultaneous translation of German lectures. » In : *INTERSPEECH*, p. 3473–3477.

CHO, Kyunghyun, Bart VAN MERRIËNBOER, Caglar GULCEHRE, Dzmitry BAHDANAU, Fethi BOUGARES, Holger SCHWENK et Yoshua BENGIO (2014). « Learning phrase representations using RNN encoder-decoder for statistical machine translation ». In : *Empirical Methods in Natural Language Processing (EMNLP)*.

CHOI, Freddy YY (2000). « Advances in domain independent linear text segmentation ». In : *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, p. 26–33.

CHUNG, Junyoung, Caglar GULCEHRE, Kyunghyun CHO et Yoshua BENGIO (2014). « Empirical evaluation of gated recurrent neural networks on sequence modeling ». In : *NIPS 2014 Workshop on Deep Learning, December 2014*.

DAHL, George E, Dong YU, Li DENG et Alex ACERO (2011a). « Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition ». In : *IEEE Transactions on audio, speech, and language processing* 20.1, p. 30–42.

DAHL, George E, Dong YU, Li DENG et Alex ACERO (2011b). « Large vocabulary continuous speech recognition with context-dependent DBN-HMMs ». In : *ICASSP*, p. 4688–4691.

DAVIS, Steven et Paul MERMELSTEIN (1980). « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences ». In : *IEEE transactions on acoustics, speech, and signal processing* 28.4, p. 357–366.

DEENA, Salil, Madina HASAN, Mortaza DOULATY, Oscar SAZ et Thomas HAIN (2016). « Combining feature and model-based adaptation of RNNLMs for multi-genre broadcast speech recognition ». In : *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Sheffield, p. 2343–2347.

DEENA, Salil, Madina HASAN, Mortaza DOULATY, Oscar SAZ et Thomas HAIN (2019). « Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition and Alignment ». In : *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 27.3, p. 572–582.

DORAI, Chitra, Vincent ORIA et Viswanath NEELAVALLI (2003). « Structuralizing educational videos based on presentation content ». In : *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. T. 2. IEEE, p. II–1029.

ELLOUMI, Zied (2019). « Prédiction de performances des systèmes de Reconnaissance Automatique de la Parole ». Thèse de doct.

ESTÈVE, Yannick (2002). « Intégration de sources de connaissances pour la modélisation stochastique du langage appliquée à la parole continue dans un contexte de dialogue oral homme-machine ». Thèse de doct.

FAUCONNIER, Jean-Philippe, Laurent SORIN, Mouna KAMEL, Mustapha MOJAHID et Nathalie AUSSENAC-GILLES (2014). « Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux ». In : *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*.

FEDERICO, Marcello (1999). « Efficient language model adaptation through MDI estimation ». In : *Sixth European Conference on Speech Communication and Technology*.

FEDERICO, Marcello et Nicola BERTOLDI (2004). « Broadcast news LM adaptation over time ». In : *Computer Speech & Language* 18.4, p. 417–435.

FETTERLY, Dennis, Mark MANASSE, Marc NAJORK et Janet L WIENER (2004). « A large-scale study of the evolution of Web pages ». In : *Software : Practice and Experience* 34.2, p. 213–237.

FORNEY, G David (1973). « The viterbi algorithm ». In : *Proceedings of the IEEE* 61.3, p. 268–278.

FÜGEN, Christian (2008). « A system for simultaneous translation of lectures and speeches ». Thèse de doct. Verlag nicht ermittelbar.

FÜGEN, Christian, Matthias WÖLFEL, John W MCDONOUGH, Shajith IKBAL, Florian KRAFT, Kornel LASKOWSKI, Mari OSTENDORF, Sebastian STÜKER et Kenichi KUMATANI (2006a). « Advances in lecture recognition : The isl rt-06s evaluation system ». In : *Ninth International Conference on Spoken Language Processing*.

FÜGEN, Christian, Shajith IKBAL, Florian KRAFT, Kenichi KUMATANI, Kornel LASKOWSKI, John W MCDONOUGH, Mari OSTENDORF, Sebastian STÜKER et Matthias WÖLFEL (2006b). « The ISL RT-06S speech-to-text system ». In : *International Workshop on Machine Learning for Multimodal Interaction*. Springer, p. 407–418.

FUJII, Atsushi, Katunobu ITOU et Tetsuya ISHIKAWA (2006). « Lodem : A system for on-demand video lectures ». In : *Speech Communication* 48.5, p. 516–531.

FUJII, Yasuhisa, Kazumasa YAMAMOTO, Norihide KITAOKA et Seiichi NAKAGAWA (2008). « Class lecture summarization taking into account consecutiveness of important sentences ». In : *Ninth Annual Conference of the International Speech Communication Association*.

FURUI, Sadaoki, Kikuo MAEKAWA et Hitoshi ISAHARA (2000). « A Japanese national project on spontaneous speech corpus and processing technology ». In : *ASR2000-Automatic Speech Recognition : Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.

FURUI, Sadaoki, Koji IWANO, Chiori HORI, Takahiro SHINOZAKI, Yohei SAITO et Satoshi TAMURA (2001). « Ubiquitous speech processing ». In : *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*. T. 1. IEEE, p. 13–16.

GALIBERT, Olivier et Juliette KAHN (2013). « The first official repere evaluation ». In : *First Workshop on Speech, Language and Audio in Multimedia*.

GALLEY, Michel, Kathleen MCKEOWN, Eric FOSLER-LUSSIER et Hongyan JING (2003). « Discourse segmentation of multi-party conversation ». In : *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, p. 562–569.

GALLIANO, Sylvain, Guillaume GRAVIER et Laura CHAUBARD (2009). « The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts ». In : *Tenth Annual Conference of the International Speech Communication Association*.

GALLIANO, Sylvain, Edouard GEOFFROIS, Djamel MOSTEFA, Khalid CHOUKRI, Jean-François BONASTRE et Guillaume GRAVIER (2005). « The ESTER phase II evaluation campaign for the rich transcription of French broadcast news ». In : *Ninth European Conference on Speech Communication and Technology*.

GELAN, Anouk (2010). « Language and Text-to-Speech technologies for highly accessible language & culture learning ». In : International Association of Online Engineering.

GEORGESCUL, Maria, Alexander CLARK et Susan ARMSTRONG (2006). « Word distributions for thematic segmentation in a support vector machine approach ». In : *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, p. 101–108.

GHANNAY, Sahar (2017). « Etude sur les representations continues de mots appliquees a la detection automatique des erreurs de reconnaissance de la parole ». Thèse de doct. Le Mans.

GILDEA, Daniel et Thomas HOFMANN (1999). « Topic-based language models using EM ». In : *Sixth European Conference on Speech Communication and Technology*.

GISH, Herbert, M-H SIU et Robin ROHLICEK (1991). « Segregation of speakers for speech recognition and speaker identification ». In : *[Proceedings] ICASSP 91 : 1991 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, p. 873–876.

GLASS, James, Timothy J HAZEN, Lee HETHERINGTON et Chao WANG (2004). « Analysis and processing of lecture audio data : Preliminary investigations ». In : *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*. Association for Computational Linguistics, p. 9–12.

GLASS, James, Timothy J HAZEN, Scott CYPHERS, Igor MALIOUTOV, David HUYNH et Regina BARZILAY (2007). « Recent progress in the MIT spoken lecture processing project ». In : *Eighth Annual Conference of the International Speech Communication Association*.

GLASS, James R, Timothy J HAZEN, D Scott CYPHERS, Ken SCHUTTE et Alex PARK (2005). « The MIT spoken lecture processing project ». In : *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, p. 28–29.

GOLDWATER, Sharon, Dan JURAFSKY et Christopher D MANNING (2008). « Which words are hard to recognize ? Prosodic, lexical, and disfluency factors that increase ASR error rates ». In : *Proceedings of ACL-08 : HLT*, p. 380–388.

GOLDWATER, Sharon, Dan JURAFSKY et Christopher D MANNING (2010). « Which words are hard to recognize ? Prosodic, lexical, and disfluency factors that increase speech recognition error rates ». In : *Speech Communication* 52.3, p. 181–200.

GOOD, Irving J (1953). « The population frequencies of species and the estimation of population parameters ». In : *Biometrika* 40.3-4, p. 237–264.

GOODMAN, Joshua T (2001). « A bit of progress in language modeling ». In : *Computer Speech & Language* 15.4, p. 403–434.

GRAVIER, G, JF BONASTRE, E GEOFFROIS, S GALLIANO, K MCTAIT et K CHOUKRI (2004). « ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français ». In : *Proc. Journées d'Etude sur la Parole (JEP)*.

GRÉZL, Frantisek, Martin KARAFIÁT, Stanislav KONTÁR et Jan CERNOCKY (2007). « Probabilistic and bottle-neck features for LVCSR of meetings ». In : *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. T. 4. IEEE, p. IV–757.

GUINAUDEAU, Camille, Guillaume GRAVIER et Pascale SÉBILLOT (2010). « Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations ». In : *Eleventh Annual Conference of the International Speech Communication Association*.

GUINAUDEAU, Camille et Julia HIRSCHBERG (2011). « Accounting for Prosodic Information to Improve ASR-Based Topic Tracking for TV Broadcast News ». In : *Twelfth Annual Conference of the International Speech Communication Association*.

HE, Xiaodong, Li DENG et Alex ACERO (2011). « Why word error rate is not a good metric for speech recognizer training for the speech translation task ? » In : *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 5632–5635.

HEARST, Marti A (1997). « TextTiling : Segmenting text into multi-paragraph subtopic passages ». In : *Computational linguistics* 23.1, p. 33–64.

HENTSCHEL, Michael, Marc DELCROIX, Atsunori OGAWA, Tomoharu IWATA et Tomohiro NAKATANI (2019a). « A Unified Framework for Feature-based Domain Adaptation of Neural Network Language Models ». In : *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 7250–7254.

HENTSCHEL, Michael, Marc DELCROIX, Atsunori OGAWA, Tomoharu IWATA et Tomohiro NAKATANI (2019b). « Feature Based Domain Adaptation for Neural Network Language Models

with Factorised Hidden Layers ». In : *IEICE transactions on Information and Systems* 102.3, p. 598–608.

HERMANSKY, Hynek et Louis Anthony COX JR (1991). « Perceptual linear predictive (PLP) analysis-resynthesis technique ». In : *Second European Conference on Speech Communication and Technology*.

HERMANSKY, Hynek et Sangita SHARMA (1999). « Temporal patterns (TRAPS) in ASR of noisy speech ». In : *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. T. 1. IEEE, p. 289–292.

HERMANSKY, Hynek, Nelson MORGAN, Aruna BAYYA et Phil KOHN (1992). « RASTA-PLP speech analysis technique ». In : *Proceedings ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 1. IEEE, p. 121–124.

HERNANDEZ, Nicolas (2004). « Description et détection automatique de structures de texte ». Thèse de doct.

HERNANDEZ, Nicolas et Brigitte GRAU (2005). « Détection automatique de structures fines de texte ». In : *Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*.

HINTON, Geoffrey, Li DENG, Dong YU, George DAHL, Abdel-rahman MOHAMED, Navdeep JAITLY, Andrew SENIOR, Vincent VANHOUCKE, Patrick NGUYEN, Brian KINGSBURY et al. (2012). « Deep neural networks for acoustic modeling in speech recognition ». In : *IEEE Signal processing magazine* 29.

HO, Ivan, Hajime KIYOHARA, Akira SUGIMOTO et Kazuo YANA (2005). « Enhancing global and synchronous distance learning and teaching by using instant transcript and translation ». In : *Cyberworlds, 2005. International Conference on*. IEEE, 5–pp.

HOCHREITER, Sepp et Jürgen SCHMIDHUBER (1997). « Long short-term memory ». In : *Neural computation* 9.8, p. 1735–1780.

HOFMANN, Thomas (1999). « Probabilistic latent semantic analysis ». In : *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 289–296.

HOFMANN, Thomas (2017). « Probabilistic latent semantic indexing ». In : *ACM SIGIR Forum*. T. 51. 2. ACM, p. 211–218.

HSU, Bo-June (2007). « Generalized linear interpolation of language models ». In : *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, p. 136–140.

HSU, Bo-June Paul et James GLASS (2006). « Style & topic language model adaptation using HMM-LDA ». In : *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, p. 373–381.

HUANG, Jing, Martin WESTPHAL, Stanley CHEN, Olivier SIOHAN, Daniel POVEY, Vit LIBAL, Alvaro SONEIRO, Henrik SCHULZ, Thomas ROSS et Gerasimos POTAMIANOS (2006). « The

IBM Rich Transcription Spring 2006 speech-to-text system for lecture meetings ». In : *International Workshop on Machine Learning for Multimodal Interaction*. Springer, p. 432–443.

HWANG, Wu-Yuin, Rustam SHADIEV, Tony CT KUO et Nian-Shing CHEN (2012). « Effects of speech-to-text recognition application on learning performance in synchronous cyber classrooms ». In : *Journal of Educational Technology & Society* 15.1, p. 367.

IGLESIAS, Ana, Lourdes MORENO et Javier JIMÉNEZ (2010). « Supporting Teachers to Automatically Build Accessible Pedagogical Resources : The APEINTA Project ». In : *Technology Enhanced Learning. Quality of Teaching and Educational Reform*. Springer Berlin Heidelberg, p. 620–624.

IGLESIAS, Ana, Javier JIMÉNEZ, Pablo REVUELTA et Lourdes MORENO (2016). « Avoiding communication barriers in the classroom : the APEINTA project ». In : *Interactive Learning Environments* 24.4, p. 829–843.

JANSEN, Dean, Aleli ALCALA et Francisco GUZMAN (2014). « Amara : A sustainable, global solution for accessibility, powered by communities of volunteers ». In : *International Conference on Universal Access in Human-Computer Interaction*. Springer, p. 401–411.

JELINEK, Fred, Robert L MERCER, Lalit R BAHL et James K BAKER (1977). « Perplexity—a measure of the difficulty of speech recognition tasks ». In : *The Journal of the Acoustical Society of America* 62.S1, S63–S63.

JELINEK, Frederick (1976). « Continuous speech recognition by statistical methods ». In : *Proceedings of the IEEE* 64.4, p. 532–556.

JOTY, Shafiq, Giuseppe CARENINI, Gabriel MURRAY et Raymond T NG (2011). « Supervised topic segmentation of email conversations ». In : *Fifth International AAAI Conference on Weblogs and Social Media*.

JUAN, Sarah Flora Samson (2015). « Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from Malaysia ». Thèse de doct.

JUNG, Hyeungshik, Hijung Valentina SHIN et Juho KIM (2018). « DynamicSlide : Exploring the Design Space of Reference-based Interaction Techniques for Slide-based Lecture Videos ». In : *Proceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface*. ACM, p. 33–41.

KAHN, Juliette, Olivier GALIBERT, Ludovic QUINTARD, Matthieu CARRÉ, Aude GIRAUDEL et Philippe JOLY (2012). « A presentation of the REPERE challenge ». In : *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, p. 1–6.

KAN, M-Y (1998). « Linear Segmentation and Segment Significance ». In : *Proc. of WVLC-6, 1998*.

KAWAHARA, Tatsuya, Yusuke NEMOTO et Yuya AKITA (2008). « Automatic lecture transcription by exploiting presentation slide information for language model adaptation ». In : *2008 IEEE*

International Conference on Acoustics, Speech and Signal Processing. IEEE, p. 4929–4932.

KELLNER, Andreas (1998). « Initial language models for spoken dialogue systems ». In : *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. T. 1. IEEE, p. 185–188.

KIM, Hyun Hee et Yong Ho KIM (2016). « Generic speech summarization of transcribed lecture videos : Using tags and their semantic relations ». In : *Journal of the Association for Information Science and Technology* 67.2, p. 366–379.

KIM, Yoon, Yacine JERNITE, David SONTAG et Alexander M RUSH (2016). « Character-aware neural language models ». In : *Thirtieth AAAI Conference on Artificial Intelligence*.

KINGSBURY, Brian, Tara N SAINATH et Hagen SOLTAU (2012). « Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization ». In : *Interspeech*, p. 10–13.

KLAKOW, Dietrich (1998). « Log-linear interpolation of language models ». In : *Fifth International Conference on Spoken Language Processing*.

KNESER, Reinhard et Hermann NEY (1995). « Improved backing-off for m-gram language modeling ». In : *1995 International Conference on Acoustics, Speech, and Signal Processing*. T. 1. IEEE, p. 181–184.

KO, Tom, Vijayaditya PEDDINTI, Daniel POVEY et Sanjeev KHUDANPUR (2015). « Audio augmentation for speech recognition ». In : *Sixteenth Annual Conference of the International Speech Communication Association*.

KORDONI, Valia, APJ van den BOSCH, Katia Lida KERMANIDIS, Vilelmini SOSONI, Kostadin CHOLAKOV, IHE HENDRICKX et Matthias HUCK (2016). « Enhancing access to online education : Quality machine translation of MOOC content ». In : *Language Resources and Evaluation Conference (LREC)*. European Language Resources Association.

KOSHOREK, Omri, Adir COHEN, Noam MOR, Michael ROTMAN et Jonathan BERANT (2018). « Text Segmentation as a Supervised Learning Task ». In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 469–473.

KUHN, Roland et Renato DE MORI (1990). « A cache-based natural language model for speech recognition ». In : *IEEE transactions on pattern analysis and machine intelligence* 12.6, p. 570–583.

KULLBACK, Solomon (1997). *Information theory and statistics*. Courier Corporation.

KURATA, Gakuto, Bhuvana RAMABHADRAN, George SAON et Abhinav SETHY (2017). « Language modeling with highway LSTM ». In : *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, p. 244–251.

LAMEL, Lori, Gilles ADDA, Eric BILINSKI et Jean-Luc GAUVAIN (2005). « Transcribing lectures and seminars ». In : *Ninth European Conference on Speech Communication and Technology*.

LAMPRIER, Sylvain, Tassadit AMGHAR, Bernard LEVRAT et Frederic SAUBION (2007). « On evaluation methodologies for text segmentation algorithms ». In : *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*. T. 2. IEEE, p. 19–26.

LAU, Raymond, Ronald ROSENFELD et Salim ROUKOS (1993). « Trigger-based language models : A maximum entropy approach ». In : *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 2. IEEE, p. 45–48.

LECORVÉ, Gwénolé (2010). « Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole ». Thèse de doct.

LECORVÉ, Gwénolé, Guillaume GRAVIER et Pascale SÉBILLOT (2008). « An unsupervised web-based topic language model adaptation method ». In : *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, p. 5081–5084.

LECOUTEUX, Benjamin (2008). « Reconnaissance automatique de la parole guidée par des transcriptions a priori ». Thèse de doct.

LECUN, Yann, Bernhard E BOSER, John S DENKER, Donnie HENDERSON, Richard E HOWARD, Wayne E HUBBARD et Lawrence D JACKEL (1990). « Handwritten digit recognition with a back-propagation network ». In : *Advances in neural information processing systems*, p. 396–404.

LI, Jinyu, Li DENG, Reinhold HAEB-UMBACH et Yifan GONG (2015). *Robust automatic speech recognition : a bridge to practical applications*. Academic Press.

LIN, Ming (2006). « Automated Lecture Video Segmentation : Facilitate Content Browsing and Retrieval ». Thèse de doct. The University of Arizona Tucson.

LIN, Ming, Jay F NUNAMAKER, Michael CHAU et Hsinchun CHEN (2004). « Segmentation of lecture videos based on text : a method combining multiple linguistic features ». In : *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE, 9–pp.

LIN, Ming, Michael CHAU, Jinwei CAO et Jay F NUNAMAKER JR (2005). « Automated video segmentation for lecture videos : A linguistics-based approach ». In : *International Journal of Technology and Human Interaction (IJTHI)* 1.2, p. 27–45.

LING, Zhang (2019). « An Acoustic Model for English Speech Recognition Based on Deep Learning ». In : *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, p. 610–614.

LIU, Xunying, Mark JF GALES et Philip C WOODLAND (2008). « Context dependent language model adaptation ». In : *Ninth Annual Conference of the International Speech Communication Association*.

LIU, Yu-Chi, Xi-Dao LUAN, Yu-Xiang XIE, Duan-Hui DAI et Ling-Da WU (2006). « Narrative structure analysis of lecture video with hierarchical hidden markov model for e-learning ». In : *International Conference on Technologies for E-Learning and Digital Entertainment*. Springer, p. 429–437.

LU, Han, Sheng-syun SHEN, Sz-Rung SHIANG, Hung-yi LEE et Lin-shan LEE (2014). « Alignment of spoken utterances with slide content for easier learning with recorded lectures using structured support vector machine (svm) ». In : *Fifteenth Annual Conference of the International Speech Communication Association*.

LUPPI, Elena, Raffaella PRIMIANI, Carla RAFFAELLI, Daniela TIBALDI, Ivan TRAINA et Anna VIOLI (2009). « Net4voice : new technologies for voice-converting in barrier-free learning environments ». In : *eLearning papers* 13, p. 4.

MAERGNER, Paul, Alex WAIBEL et Ian LANE (2012). « Unsupervised vocabulary selection for real-time speech recognition of lectures ». In : *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 4417–4420.

MAKHOUL, John, Francis KUBALA, Richard SCHWARTZ, Ralph WEISCHEDEL et al. (1999). « Performance measures for information extraction ». In : *Proceedings of DARPA broadcast news workshop*. Herndon, VA, p. 249–252.

MALIOUTOV, Igor Igor Mikhailovich (2006). « Minimum cut model for spoken lecture segmentation ». Thèse de doct. Massachusetts Institute of Technology.

MANGU, Lidia, Eric BRILL et Andreas STOLCKE (2000). « Finding consensus in speech recognition : word error minimization and other applications of confusion networks ». In : *Computer Speech & Language* 14.4, p. 373–400.

MAREÜIL, Philippe Boula de, Christophe D'ALESSANDRO, François YVON, Véronique AUBERGÉ, Jacqueline VAISSIÈRE et Angélique AMELOT (2000). « A French Phonetic Lexicon with Variants for Speech and Language Processing. » In : *LREC*.

MARKEL, John E et AH GRAY (1982). *Linear Prediction of Speech*. Springer-Verlag.

MARQUARD, Stephen (2012). « Improving searchability of automatically transcribed lectures through dynamic language modelling ». Thèse de doct. University of Cape Town.

MARTÍNEZ-VILLARONGA, Adrià, A MIGUEL, Jesús ANDRÉS-FERRER et Alfons JUAN (2013). « Language model adaptation for video lectures transcription ». In : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, p. 8450–8454.

MARTINS, Ciro, António TEXEIRA et Joao NETO (2006). « Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system ». In : *2006 IEEE Spoken Language Technology Workshop*. IEEE, p. 146–149.

MASMOUDI, Abir, Salima MDHAFFAR, Rahma SELLAMI et Lamia Hadrich BELGUITH (2019). « Automatic diacritics restoration for Tunisian dialect ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18.3, p. 28.

MASUMURA, Ryo, Seongjun HAHM et Akinori ITO (2011). « Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition ». In : *Twelfth Annual Conference of the International Speech Communication Association*.

MDHAFFAR, Salima, Antoine LAURENT et Yannick ESTÈVE (2018a). « Etude de performance des réseaux neuronaux récurrents dans le cadre de la campagne d'évaluation Multi-Genre Broadcast challenge 3 (MGB3) ». In : *XXXIIe Journées d'Etudes sur la Parole (JEP 2018)*.

MDHAFFAR, Salima, Antoine LAURENT et Yannick ESTÈVE (2018b). « Le corpus PASTEL pour le traitement automatique de cours magistraux ». In : *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018)*.

MDHAFFAR, Salima, Fethi BOUGARES, Yannick ESTÈVE et Lamia HADRICH-BELGUITH (2017). « Sentiment analysis of Tunisian dialects : Linguistic ressources and experiments ». In : *Proceedings of the third Arabic natural language processing workshop*, p. 55–61.

MDHAFFAR, Salima, Yannick ESTÈVE, Nicolas HERNANDEZ, Antoine LAURENT et Solen QUINIOU (2019a). « Apport de l'adaptation automatique des modèles de langage pour la reconnaissance de la parole : évaluation qualitative extrinsèque dans un contexte de traitement de cours magistraux ». In : *26e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2019)*.

MDHAFFAR, Salima, Yannick ESTÈVE, Nicolas HERNANDEZ, Antoine LAURENT, Richard DUFOUR et Solen QUINIOU (2019b). « Qualitative evaluation of ASR adaptation in a lecture context : Application to the PASTEL corpus ». In : *Proc. Interspeech 2019*, p. 569–573.

MDHAFFAR, Salima, Yannick ESTÈVE, Antoine LAURENT, Nicolas HERNANDEZ, Richard DUFOUR, Charlet DELPHINE, Damnati GÉRALDINE, Solen QUNIOU et Camelin NATHALIE (2020). « A Multimodal Educational Corpus of Oral Courses : Annotation, Analysis and Case Study ». In : *LREC*.

MEIGNIER, Sylvain et Teva MERLIN (2010). « LIUM SpkDiarization : an open source toolkit for diarization ». In : *CMU SPUD Workshop*.

MIKOLOV, Tomáš (2012). « Statistical language models based on neural networks ». Thèse de doct. Brno University of Technology.

MIKOLOV, Tomas et Geoffrey ZWEIG (2012). « Context dependent recurrent neural network language model ». In : *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, p. 234–239.

MIKOLOV, Tomáš, Martin KARAFIÁT, Lukáš BURGET, Jan ČERNOCKỲ et Sanjeev KHUDANPUR (2010). « Recurrent neural network based language model ». In : *Eleventh annual conference of the international speech communication association*.

MIKOLOV, Tomáš, Stefan KOMBRINK, Lukáš BURGET, Jan ČERNOCKỲ et Sanjeev KHUDANPUR (2011). « Extensions of recurrent neural network language model ». In : *2011 IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, p. 5528–5531.

MILLER, George A (1998). *WordNet : An electronic lexical database*. MIT press.

MIRANDA, Joao, Joao Paulo NETO et Alan W BLACK (2013). « Improving ASR by integrating lecture audio and slides ». In : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, p. 8131–8135.

MIRÓ, Juan Daniel Valor (2017). « Evaluation of innovative computer-assisted transcription and translation strategies for video lecture repositories ». Thèse de doct. Universitat Politècnica de València.

MISRA, Hemant, François YVON, Joemon M JOSE et Olivier CAPPE (2009). « Text segmentation via topic modeling : an analytical study ». In : *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, p. 1553–1556.

MOORE, Robert C et William LEWIS (2010). « Intelligent selection of language model training data ». In : *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, p. 220–224.

MOREAU, Nicolas, Djamel MOSTEFA, Rainer STIEFELHAGEN, Susanne BURGER et Khalid CHOUKRI (2008). « Data Collection for the CHIL CLEAR 2007 Evaluation Campaign. » In : *LREC*. T. 8, p. 28–30.

MORGAN, Nelson et Herve BOURLARD (1990). « Continuous speech recognition using multilayer perceptrons with hidden Markov models ». In : *International conference on acoustics, speech, and signal processing*. IEEE, p. 413–416.

MOUGARD, Hugo, Matthieu RIOU, Colin DE LA HIGUERA, Solen QUINIOU et Olivier AUBERT (2015). « The Paper or the Video : Why Choose ? » In : *International World Wide Web Conference (WWW'2015)*. ACM Press.

MÜLLER, Markus, Thai Son NGUYEN, Jan NIEHUES, Eunah CHO, Bastian KRÜGER, Thanh-Le HA, Kevin KILGOUR, Matthias SPERBER, Mohammed MEDIANI, Sebastian STÜKER et al. (2016). « Lecture Translator-Speech translation framework for simultaneous lecture translation ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, p. 82–86.

MUNTEANU, Cosmin, Ronald BAECKER, Gerald PENN, Elaine TOMS et David JAMES (2006). « The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives ». In : *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, p. 493–502.

MURAMATSU, Brandon, Andrew MCKINNEY, Phillip D LONG et John ZORNIG (2009). « SpokenMedia project : Media-linked transcripts and rich media notebooks for learning and teaching ». In : *2009 International Workshop on Technology for Education*. IEEE, p. 6–9.

NASR, Alexis, Frédéric BÉCHET, Jean-François REY, Benoît FAVRE et Joseph LE ROUX (2011). « Macaon : An nlp tool suite for processing word lattices ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Systems Demonstrations*. Association for Computational Linguistics, p. 86–91.

NICKEL, Kai, Tobias GEHRIG, Hazim K EKENEL, John MCDONOUGH et Rainer STIEFELHAGEN (2006). « An audio-visual particle filter for speaker tracking on the clear'06 evaluation dataset ». In : *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, p. 69–80.

NTOULAS, Alexandros (2006). *Crawling and searching the Hidden Web*. T. 68. 02. Citeseer.

OGER, Stanislas (2011). « Modèles de langage ad hoc pour la reconnaissance automatique de la parole ». Thèse de doct.

OGER, Stanislas, Vladimir POPESCU et Georges LINARES (2009). « Using the world wide web for learning new words in continuous speech recognition tasks : Two case studies ». In : *Proc. Speech and Computer*, p. 76–81.

OGER, Stanislas, Georges LINARES, Frédéric BÉCHET et Pascal NOCERA (2008). « On-demand new word learning using world wide web ». In : *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, p. 4305–4308.

PALLETT, David S (2003). « A look at NIST's benchmark ASR tests : past, present, and future ». In : *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, p. 483–488.

PALMER, David D et Mari OSTENDORF (2005). « Improving out-of-vocabulary name resolution ». In : *Computer Speech & Language* 19.1, p. 107–128.

PARK, Alex, Timothy J HAZEN et James R GLASS (2005). « Automatic processing of audio lectures for information retrieval : Vocabulary selection and language modeling ». In : *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* T. 1. IEEE, p. I–497.

PARK, Junho, Xunying LIU, Mark JF GALES et Phil C WOODLAND (2010). « Improved neural network based language modelling and adaptation ». In : *Eleventh Annual Conference of the International Speech Communication Association*.

PEDDINTI, Vijayaditya, Daniel POVEY et Sanjeev KHUDANPUR (2015). « A time delay neural network architecture for efficient modeling of long temporal contexts ». In : *Sixteenth Annual Conference of the International Speech Communication Association*.

PEVZNER, Lev et Marti A HEARST (2002). « A critique and improvement of an evaluation metric for text segmentation ». In : *Computational Linguistics* 28.1, p. 19–36.

POVEY, Daniel, Vijayaditya PEDDINTI, Daniel GALVEZ, Pegah GHAHREMANI, Vimal MANOHAR, Xingyu NA, Yiming WANG et Sanjeev KHUDANPUR (2016). « Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. » In : *Interspeech*, p. 2751–2755.

PRZYBOCKI, Mark A., Jonathan G. FISCUS, John S. GAROFOLO et David S. PALLETT (1999). « 1998 HUB-4 INFORMATION EXTRACTION EVALUATION ». In : *Proceedings of DARPA broadcast news workshop*.

RAUX, Antoine, Brian LANGNER, Alan W BLACK et Maxine ESKENAZI (2003). « Let's go : Improving spoken dialog systems for the elderly and non-natives ». In : *Eighth European Conference on Speech Communication and Technology*.

REYNAR, Jeffrey C (1994). « An automatic method of finding topic boundaries ». In : *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 331–333.

REYNAR, Jeffrey C (1998). « Topic segmentation : Algorithms and applications ». Thèse de doct. University of Pennsylvania.

RIEDHAMMER, Korbinian, Martin GROPP et Elmar NÖTH (2012). « The FAU video lecture browser system ». In : *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, p. 392–397.

RIEDL, Martin et Chris BIEMANN (2012a). « Text segmentation with topic models ». In : *Journal for Language Technology and Computational Linguistics* 27.1, p. 47–69.

RIEDL, Martin et Chris BIEMANN (2012b). « TopicTiling : a text segmentation algorithm based on LDA ». In : *Proceedings of ACL 2012 Student Research Workshop*. Association for Computational Linguistics, p. 37–42.

ROSENFELD, Roni (1996). « A maximum entropy approach to adaptive statistical language modeling ». In : *Computer Speech and Language*. T. 10, p. 187–228.

ROUSSEAU, Anthony (2013). « XenC : An open-source tool for data selection in natural language processing ». In : *The Prague Bulletin of Mathematical Linguistics* 100, p. 73–82.

ROUSSEAU, Anthony, Gilles BOULIANNE, Paul DELÉGLISE, Yannick ESTÈVE, Vishwa GUPTA et Sylvain MEIGNIER (2014). « LIUM and CRIM ASR system combination for the REPERE evaluation campaign ». In : *International Conference on Text, Speech, and Dialogue*. Springer, p. 441–448.

RUMELHART, David E, Geoffrey E HINTON, Ronald J WILLIAMS et al. (1988). « Learning representations by back-propagating errors ». In : *Cognitive modeling* 5.3, p. 1.

SADAMITSU, Kugatsu, Takuya MISHINA et Mikio YAMAMOTO (2007). « Topic-based language models using Dirichlet Mixtures ». In : *Systems and Computers in Japan* 38.12, p. 76–85.

SANDERS, Gregory A, Audrey N LE et John S GAROFOLO (2002). « Effects of word error rate in the DARPA Communicator data during 2000 and 2001 ». In : *Seventh International Conference on Spoken Language Processing*.

SAON, George, Hagen SOLTAU, David NAHAMOO et Michael PICHENY (2013). « Speaker adaptation of neural network acoustic models using i-vectors. » In : *ASRU*, p. 55–59.

SARIKAYA, Ruhi, Agustin GRAVANO et Yuqing GAO (2005). « Rapid language model development using external resources for new spoken dialog domains ». In : *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* T. 1. IEEE, p. I–573.

SCHULTE, Olaf A, Tobias WUNDEN et Armin BRUNNER (2008). « Replay : an integrated and open solution to produce, handle, and distributeaudio-visual (lecture) recordings ». In : *Proceedings of the 36th annual ACM SIGUCCS fall conference : moving mountains, blazing trails*. ACM, p. 195–198.

SCHWENK, Holger (2007). « Continuous space language models ». In : *Computer Speech & Language* 21.3, p. 492–518.

SCHWENK, Holger et Jean-Luc GAUVAIN (2004). « Neural network language models for conversational speech recognition ». In : *Eighth International Conference on Spoken Language Processing*.

SENAY, Grégory (2011). « Approches semi-automatiques pour la recherche d'information dans les documents audio ». Thèse de doct. Thèse de doctorat.

SENAY, Grégory, Benjamin LECOUTEUX et Georges LINARÈS (2012). « Prédiction de l'indexabilité d'une transcription (Prediction of transcription indexability)[in French] ». In : *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1 : JEP*, p. 697–705.

SHADIEV, Rustam, Wu-Yuin HWANG, Nian-Shing CHEN et Huang YUEH-MIN (2014). « Review of speech-to-text recognition technology for enhancing learning ». In : *Journal of Educational Technology & Society* 17.4, p. 65.

SHEIKH, Imran, Irina ILLINA et Dominique FOHR (2016). « How diachronic text corpora affect context based retrieval of OOV proper names for audio news ». In : *International Conference on Language Resources and Evaluation (LREC)*.

SINGH-MILLER, Natasha et Michael COLLINS (2007). « Trigger-based language modeling using a loss-sensitive perceptron algorithm ». In : *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. T. 4. IEEE, p. IV–25.

SITBON, Laurianne et Patrice BELLOT (2007). « Topic segmentation using weighted lexical links (wll) ». In : *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 737–738.

SIU, M-H, George YU et Herbert GISH (1992). « An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers ». In : *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. T. 2. IEEE, p. 189–192.

SOUTNER, Daniel et Luděk MÜLLER (2013). « Application of LSTM neural networks in language modelling ». In : *International Conference on Text, Speech and Dialogue*. Springer, p. 105–112.

SOUVIGNIER, Bernd, Andreas KELLNER, Bernhard RUEBER, Hauke SCHRAMM et Frank SEIDE (2000). « The thoughtful elephant : Strategies for spoken dialog systems ». In : *IEEE Transactions on Speech and Audio Processing* 8.1, p. 51–62.

SRIVASTAVA, Rupesh Kumar, Klaus GREFF et Jürgen SCHMIDHUBER (2015). « Highway networks ». In : *arXiv preprint arXiv :1505.00387*.

STIEFELHAGEN, Rainer, Keni BERNARDIN, Hazim K EKENEL et Michael VOIT (2008). « Tracking identities and attention in smart environments-contributions and progress in the CHIL project ». In : *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, p. 1–8.

STOKES, Nicola, Joe CARTHY et Alan F SMEATON (2004). « SeLeCT : a lexical cohesion based news story segmentation system ». In : *AI communications* 17.1, p. 3–12.

SUNDERMEYER, Martin, Ralf SCHLÜTER et Hermann NEY (2012). « LSTM neural networks for language modeling ». In : *Thirteenth annual conference of the international speech communication association*.

SZÖKE, Igor, Jan CERNOCKỲ, M FAPŠO et J ZIŽKA (2010). « Speech@ FIT lecture browser ». In : *2010 IEEE Spoken Language Technology Workshop*. IEEE, p. 169–170.

TRANCOSO, Isabel, Ricardo NUNES et Luís NEVES (2006). « Classroom lecture recognition ». In : *International Workshop on Computational Processing of the Portuguese Language*. Springer, p. 190–199.

TRANCOSO, Isabel, Rui MARTINS, Helena MONIZ, Ana Isabel MATA et M Céu VIANA (2008). « The Lectra corpus classroom lecture transcriptions in European Portuguese ». In : *Economic Theory* 1.17, p. 15–1.

TRENTIN, Edmondo et Marco GORI (2001). « A survey of hybrid ANN/HMM models for automatic speech recognition ». In : *Neurocomputing* 37.1-4, p. 91–126.

TÜR, Gökhan, Dilek HAKKANI-TÜR, Andreas STOLCKE et Elizabeth SHRIBERG (2001). « Integrating prosodic and lexical cues for automatic topic segmentation ». In : *Computational linguistics* 27.1, p. 31–57.

UTIYAMA, Masao et Hitoshi ISAHARA (2001). « A statistical model for domain-independent text segmentation ». In : *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 499–506.

VALOR MIRÓ, Juan Daniel, Rachel Nadine SPENCER, A Pérez González de MARTOS, G GARCÉS DÍAZ-MUNÍO, CIVERA TURRÓ, J CIVERA et A JUAN (2014). « Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures ». In : *Open Learning : The Journal of Open, Distance and e-Learning* 29.1, p. 72–85.

VESELỲ, Karel, Shinji WATANABE, Katerina ŽMOLÍKOVÁ, Martin KARAFIÁT, Lukáš BURGET et Jan Honza ČERNOCKỲ (2016). « Sequence summarizing neural network for speaker adap-

tation ». In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 5315–5319.

VITERBI, Andrew (1967). « Error bounds for convolutional codes and an asymptotically optimum decoding algorithm ». In : *IEEE transactions on Information Theory* 13.2, p. 260–269.

WANG, Liang, Sujian LI, Yajuan LV et WANG HOUFENG (2017). « Learning to rank semantic coherence for topic segmentation ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1340–1344.

WANG, Xiaoxuan, Lei XIE, Mimi LU, Bin MA, Eng Siong CHNG et Haizhou LI (2012). « Broadcast news story segmentation using conditional random fields and multimodal features ». In : *IEICE TRANSACTIONS on Information and Systems* 95.5, p. 1206–1215.

WANG, Ye-Yi, Alex ACERO et Ciprian CHELBA (2003). « Is word error rate a good indicator for spoken language understanding accuracy ». In : *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, p. 577–582.

WITTEN, Ian H et Timothy C BELL (1991). « The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression ». In : *Ieee transactions on information theory* 37.4, p. 1085–1094.

WITTENBURG, Peter, Hennie BRUGMAN, Albert RUSSEL, Alex KLASSMANN et Han SLOETJES (2006). « ELAN : a professional framework for multimodality research ». In : *5th International Conference on Language Resources and Evaluation (LREC 2006)*, p. 1556–1559.

XIONG, Wayne, Jasha DROPPO, Xuedong HUANG, Frank SEIDE, Mike SELTZER, Andreas STOLCKE, Dong YU et Geoffrey ZWEIG (2016). « Achieving human parity in conversational speech recognition ». In : *arXiv preprint arXiv :1610.05256*.

YAMAMOTO, Natsuo, Jun OGATA et Yasuo ARIKI (2003). « Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition ». In : *Eighth European Conference on Speech Communication and Technology*.

YAMAZAKI, Hiroki, Koji IWANO, Koichi SHINODA, Sadaoki FURUI et Haruo YOKOTA (2007). « Dynamic language model adaptation using presentation slides for lecture speech recognition ». In : *Eighth Annual Conference of the International Speech Communication Association*.

YU, Dong, Li DENG et George DAHL (2010). « Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition ». In : *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

YU, Dong et Michael L SELTZER (2011). « Improved bottleneck features using pretrained deep neural networks ». In : *Twelfth annual conference of the international speech communication association*.

ZHAI, Chengxiang et John LAFFERTY (2004). « A study of smoothing methods for language models applied to information retrieval ». In : *ACM Transactions on Information Systems (TOIS)* 22.2, p. 179–214.

ZHAI, Chengxiang et John LAFFERTY (2017). « A study of smoothing methods for language models applied to ad hoc information retrieval ». In : *ACM SIGIR Forum*. T. 51. 2. ACM, p. 268–276.

ZHANG, Justin Jian et Pascale FUNG (2009). « Active learning of extractive reference summaries for lecture speech summarization ». In : *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*. Association for Computational Linguistics, p. 23–26.

ZITOUNI, Imed (2000). « Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires : application à MAUD ». Thèse de doct.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

Titre : Reconnaissance de la parole dans un contexte de cours magistraux : évaluation, avancées et enrichissement

Résumé : Cette thèse s'inscrit dans le cadre d'une étude sur le potentiel de la transcription automatique pour l'instrumentation de situations pédagogiques. Notre contribution porte sur plusieurs axes. Dans un premier temps, nous décrivons l'enrichissement et l'annotation du corpus COCo que nous avons réalisés dans le cadre du projet ANR PASTEL. Ce corpus est composé de vidéos de différents cours magistraux, chacun étant spécialisé dans un domaine particulier (langage naturel, graphes, fonctions...). Dans ce cadre multi-thématiques, nous nous sommes ensuite intéressés à la problématique de l'adaptation linguistique des systèmes de reconnaissance automatique de la parole (SRAP). La proposition d'adaptation des modèles s'appuie à la fois sur les supports de présentation de cours fournis par les enseignants et sur des données spécialisées récoltées automatiquement à partir du web. Puis, nous nous sommes focalisés sur la problématique de l'évaluation des SRAP, les métriques existantes ne permettant pas une évaluation précise de la qualité des transcriptions dans un cadre applicatif déterminé. Ainsi, nous avons proposé deux protocoles d'évaluation. Le premier porte sur une évaluation intrinsèque, permettant d'estimer la performance seulement pour des mots spécialisés de chacun des cours ($IWER_{Average}$). D'autre part, nous proposons une évaluation extrinsèque, qui estime la performance pour deux tâches exploitant la transcription : la recherche d'informations et l'indexabilité. Nos résultats expérimentaux montrent que le taux d'erreurs-mots global (WER) masque les apports effectifs de l'adaptation des modèles de langage et prouve la nécessité d'utiliser de nouvelles mesures, telles que celles présentées dans ce manuscrit, pour évaluer l'apport réel de l'adaptation des modèles de langage. L'adaptation reposant sur une collecte de données issues du web, nous avons cherché à rendre compte de la reproductibilité des résultats sur l'adaptation de modèles de langage en comparant les performances obtenues sur une longue période temporelle. Nos résultats expérimentaux montrent que même si les données sur le web changent en partie d'une période à l'autre, la variabilité de la performance des systèmes de transcription adaptés est restée non significative à partir d'un nombre minimum de documents collectés. Enfin, nous avons proposé une approche permettant de structurer la sortie de la transcription automatique en segmentant thématiquement la transcription et en alignant la transcription avec les diapositives des supports de cours. Pour la segmentation, l'intégration de l'information de changement de diapositives dans l'algorithme TextTiling apporte un gain significatif en termes de F-mesure. Pour l'alignement, nous avons développé une technique basé sur des représentations TF-IDF en imposant une contrainte pour respecter l'ordre séquentiel des diapositives et des segments de transcription et nous avons vérifié la fiabilité de l'approche utilisée à l'aide d'une mesure de confiance.

Title: Speech recognition in the context of lectures: evaluation, progress and enrichment

Keywords: Language model, transcription, evaluation, adaptation, automatic structuration

Abstract: This thesis is part of a study that explores automatic transcription potential for the instrumentation of educational situations. Our contribution covers several axes. First, we describe the enrichment and the annotation of COCo dataset that we produced as part of the ANR PASTEL project. This corpus is composed of different lectures' videos. Each lecture is related to a particular field (natural language, graphs, functions ...). In this multi-thematic framework, we are interested in the problem of the linguistic adaptation of automatic speech recognition systems (ASR). The proposed language model adaptation is based both on the lecture presentation supports provided by the teacher and in-domain data collected automatically from the web. Then, we focused on the ASR evaluation problem. The existing metrics don't allow a precise evaluation of the transcriptions' quality. Thus, we proposed two evaluation protocols. The first one deals with an intrinsic evaluation, making it possible to estimate performance only for domain words of each lecture ($IWER_{Average}$). The second protocol offers an extrinsic evaluation, which estimates the performance for two tasks exploiting transcription: information retrieval and indexability. Our experimental results show that the global word error rate (WER) masks the gain provided by language model adaptation. So, to better evaluate this gain, it seems particularly relevant to use specific measures, like those presented in this thesis. As LM adaptation is based on a collection of data from the web, we study the reproducibility of language model adaptation results by comparing the performances obtained over a long period of time. Over a collection period of one year, we were able to show that, although the data on the Web changed in part from one month to the next, the performance of the adapted transcription systems remained constant (i.e. no significant performance changes), no matter the period considered. Finally, we are interested on thematic segmentation of ASR output and alignment of slides with oral lectures. For thematic segmentation, the integration of slide's change information into the TextTiling algorithm provides a significant gain in terms of F-measure. For alignment of slides with oral lectures, we have calculated a cosine similarity between the TF-IDF representation of the transcription segments and the TF-IDF representation of text slides and we have imposed a constraint to respect the sequential order of the slides and transcription segments. Also, we have considered a confidence measure to discuss the reliability of the proposed approach.