De l'utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole

Sylvain Raybaud

UNIVERSITÉ DE LORRAINE

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**

De l'utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole.

THÈSE

présentée et soutenue publiquement le 5 décembre 2012

pour l'obtention du

Doctorat de l'université de Lorraine

(spécialité informatique)

par

Sylvain Raybaud

Composition du jury

Président :	François Charpillet
Rapporteurs :	François Yvon Laurent Besacier
Examinateurs :	Guillaume Gravier François Charpillet Kamel Smaïli David Langlois

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

Table des matières

Table des figures

Première partie

Introduction et état de l'art

L'Humain se caractérise par deux traits essentiels : c'est un être social, et il aime se compliquer la vie. S'ensuivent deux tendances opposées : se regrouper et partager d'une part, ériger des barrières de l'autre. Et parfois les deux se marient : les accords transnationaux fleurissent, la culture se diffuse au-delà des frontières, l'Europe se construit. Aujourd'hui, des centaines de millions d'internautes peuvent visionner des millions de vidéos dans des centaines de langues différentes ; des milliers d'interprètes ont la lourde tâche de diffuser dans vingt trois langues officielles les documents produits par les instances européennes, et d'assurer la capacité des citoyens à communiquer avec des instances dans n'importe laquelle de ces langues [Eise 11]. La Commission Européenne est le premier employeur mondial de traducteurs et d'interprètes, et l'Union Européenne y consacre 1% de son budget, soit plus d'un milliard d'euros par an [Fuge 08]. Les institutions canadiennes, l'ONU et l'Union Africaine sont également d'importants consommateurs de traduction. Et pourtant, on ne le constate que trop, le dialogue international, voire intra-communautaire, se heurte à la plus ancienne et la plus infranchissable des barrières : celle de la langue. Cette barrière est tellement profonde, tellement symbolique, qu'elle est passée dans le langage courant : « Je n'y comprends, rien, *c'est du chinois !* ». On retrouve ces expressions dans toutes les langues ; les anglophones diront « It's all Greek to me ! » (figure 1) [1].

Rendre la traduction accessible au plus grand nombre est donc un intérêt sociétal majeur. Les technologies de traduction automatiques sont déjà utilisées par certains services (Youtube propose la traduction en temps réel de certains programmes), dans certaines grandes entreprises (l'entreprise suisse Autodesk utilise la traduction automatique pour sa documentation) ou institutions (au Parlement Européen par exemple). Afin d'être utilisables, des contraintes fortes pèsent sur ces systèmes :

- rapidité, surtout dans le cas de la traduction simultanée de la parole,
- robustesse aux erreurs de langue,
- robustesse aux disfluences et aux mauvaises conditions de prise de son dans le cas de la traduction de la parole,
- fiabilité du résultat.

Différentes technologies sont donc sollicitées : analyse du signal, reconnaissance de la parole, traduction de parole ou de texte, estimation de qualité. Prenons l'exemple du sous-titrage automatique de programmes télédiffusés. Dans certaines scènes d'un film, ou lors d'une interview en extérieur par exemple, les conditions de prise de son peuvent être assez mauvaises (bruit de fond, locuteur éloigné du micro, etc.). Le locuteur peut avoir un registre de langue familier, se répéter, commettre des disfluences, parler rapidement. Ces conditions difficiles rendent plus que probables l'apparition d'erreurs dans la traduction automatique. De plus, la majorité des usagers d'un système de sous-titrage automatique ne maîtrisent pas la langue originale, et sont donc dans l'incapacité de repérer ces erreurs. À défaut d'être suffisamment fiable, la traduction automatique doit donc au moins être vérifiable : on peut pour cela utiliser des techniques automatiques d'évaluation de la qualité d'une traduction, les mesures de confiance.

Le premier objectif de cette thèse sera donc le développement et l'évaluation de mesures de confiance pour la traduction automatique. Je commencerai par proposer une formalisation rigoureuse du problème (section 1). Je m'attacherai à détailler tous les concepts qui rendent ce problème accessible au calcul numérique, à fournir tous les détails permettant de reproduire les expériences, et à développer un cadre facilitant l'implémentation et la comparaison de nouvelles mesures de confiance. Dans la section 3, je présenterai plusieurs mesures de confiance représentatives de l'état de l'art (sans prétendre à l'exhaustivité), ainsi que des mesures utilisées en

1. Merci à Mark Liberman pour ce graphe délectable (`http://languagelog.ldc.upenn.edu/nll/?p=1024` — `http://www.ling.upenn.edu/~myl/`)

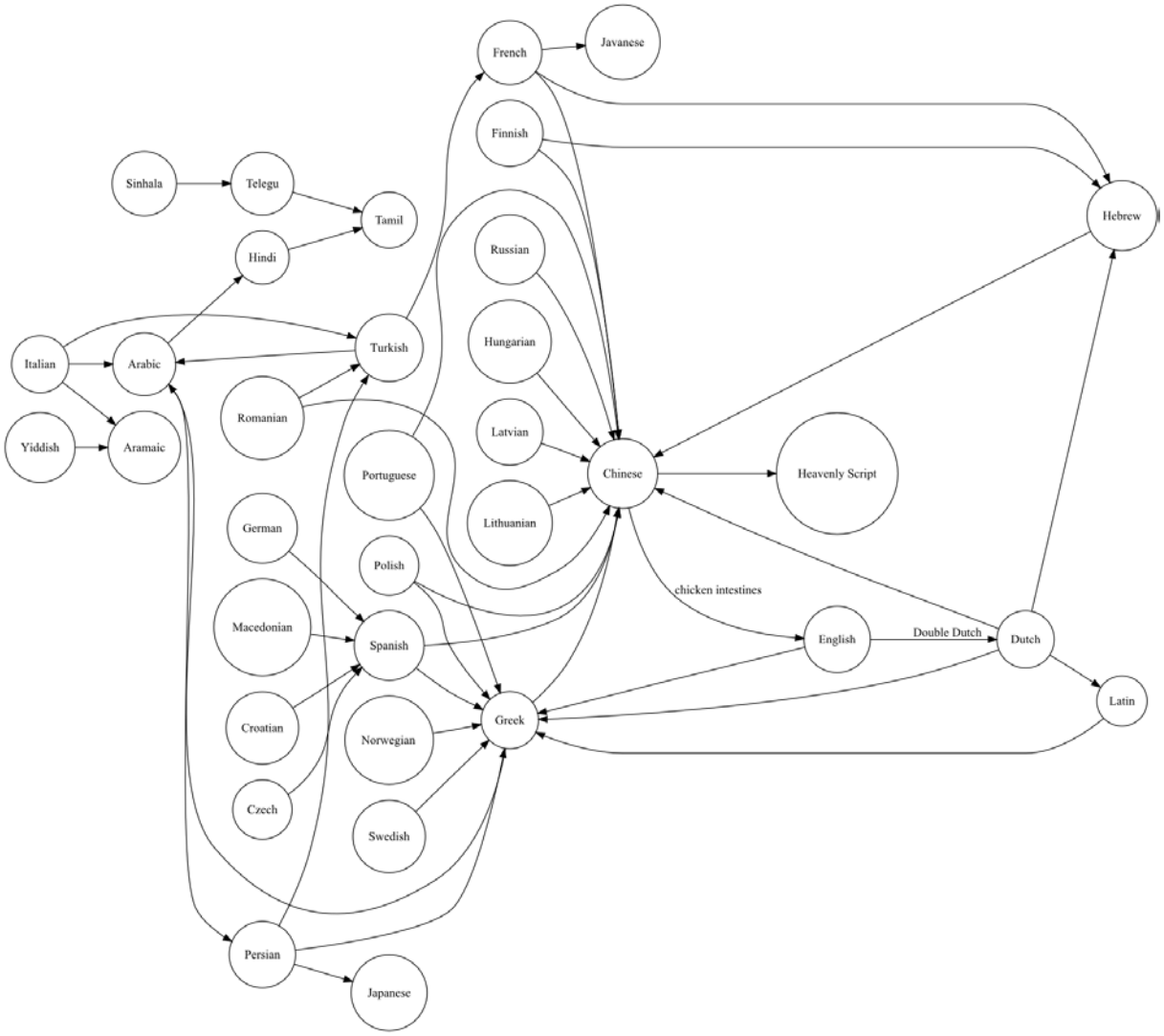


FIGURE 1 – Graphe dirigé des incompréhensions stéréotypiques.

reconnaissance vocale et adaptées pour être utilisées en traduction automatique, et des mesures originales (sections 3.1.4, 3.1.5 et 3.2.2). Outre une évaluation directe par différentes métriques, une expérience de post-édition de traductions automatiques a été menée, afin de mesurer l'aide que peuvent apporter des mesures de confiance sur cette tâche (section 4).

Enfin, je propose un système de traduction de la parole intégrant les mesures de confiance, tant comme indication de fiabilité pour l'utilisateur que comme source d'information pour le décodage (partie III). J'ai voulu rendre ce système aussi pérenne que possible, afin qu'il puisse être utilisé par qui le souhaite, notamment au sein de mon équipe de recherche et de la communauté. Une base logicielle commune ne peut qu'être bénéfique pour la reproductibilité des expériences et la comparabilité des résultats. Le programme et les bibliothèques sont donc libres (la dernière version est disponible sur `http://perso.crans.org/raybaud/maparole.tar.bz2`), et un soin particulier a été apporté à la clarté du code et à la documentation. J'ai également voulu ce système polyvalent, et utilisable hors du milieu académique. Il doit donc être capable :

– d'avoir un temps de calcul raisonnable,

– de reconnaître la parole d'un ou plusieurs locuteurs dans un contexte de discours spontané,
– de fournir une estimation de la qualité de la traduction aux utilisateurs qui ne maîtrisent pas la langue source, ou à un traducteur qui édite les propositions du système.

Au terme de cette thèse, nous disposons d'un prototype capable de traduire un dialogue spontané entre plusieurs locuteurs enregistrés par un seul micro. Il intègre pour cela des algorithmes de segmentation et de robustesse aux erreurs. Il est également capable d'informer l'utilisateur sur la fiabilité de la traduction grâce à l'utilisation de mesures de confiance. S'il a été testé sur une tâche de traduction en anglais de programmes d'informations radiodiffusés en français, ce logiciel est en théorie utilisable dans n'importe quelle situation de parole spontanée et pour d'autres paires de langues, à condition de construire les modèles de traduction et de langage pour ces langues. La qualité du résultat, bien sûr, dépendra du style de parole, de la qualité de la prise de son et de la paire de langues. Tous les objectifs n'ont cependant pas été atteints, car la vitesse de traduction reste nettement insuffisante pour permettre une application à l'interprétariat (qui nécessite une traduction en temps réel), et si les mesures de confiance fournissent une information utile pour déterminer la fiabilité générale d'une traduction, elles restent insuffisamment précises pour être utilisables sur une tâche de post-édition.

1

La traduction automatique

Je vais commencer par donner une présentation générale de la Traduction Automatique (TA) et de son histoire. Je m'attacherai à donner un aperçu de toutes les approches existant ou ayant existé, mais j'insisterai plus spécialement sur les points qui seront nécessaires à la compréhension de mes travaux, à savoir les mesures de confiance (partie II) et la traduction automatique de la parole par des méthodes statistiques (partie III). Je passerai plus rapidement sur les aspects purement culturels ou, plus éloignés de mon sujet de recherche. Le lecteur intéressé pourra se référer à l'état de l'art très complet proposé dans [Lave 10].

1.1 Un bref historique

Avant le besoin de traduire, il y a le besoin de communiquer. Les premières idées pour développer une méthode rationnelle et systématique de communication consistèrent à proposer une langue universelle, capable, comme l'écrivait René Descartes dans sa *Lettre au P. Mersenne* du 20 novembre 1629, d'« établi[r] un ordre entre toutes les pensées qui peuvent entrer en l'Esprit humain » [2]. On pense aussi bien sûr à l'*espéranto*, la langue construite (c'est-à-dire inventée, par opposition aux langues naturelles dont l'apparition et l'évolution ne sont — *a priori* — pas le résultat d'une démarche pensée, planifiée et organisée d'un individu ou groupe d'individus) du docteur Zamenhof. Cette idée d'une langue unique se rapproche du concept d'*interlangue* qui sera developpé dans la deuxième moitié du XXe siècle par certains chercheurs en traduction automatique. En attendant, les premiers développements théoriques, dans les années 1930, ont suivi une approche beaucoup moins sophistiquée : Georges Artsrouni proposait de développer une machine automatique traduisant « mot-à-mot » à l'aide d'un simple dictionnaire bilingue ; Petr Trojanskij a proposé un système utilisant un dictionnaire, des règles de grammaire, et des règles de transfert inspirées de l'*espéranto*. Dans son mémorandum publié en 1947, Warren Weaver (reproduit en 1955 dans [Weav 55]) propose de s'attaquer aux problèmes de polysémie, et suggère de s'appuyer sur les méthodes de cryptographie, sur l'interprétation des langues naturelles en termes de logique, et sur une supposée « linguistique universelle » pour réaliser la traduction automatique. Aucun système concret n'a cependant été présenté avant 1954, quand des chercheurs d'IBM ont utilisé un ordinateur pour traduire, devant les journalistes, quelques dizaines de phrases russes en anglais. L'enthousiasme soulevé fut énorme, et induisit un effort de recherche intense.

Cette catégorie de méthodes est appelée *approche experte* (section 1.3), car elle repose sur des informations fournies et encodées par des humains pour effectuer la traduction (le lexique,

2. `http://interlanguages.net/Hdesc.html`

les règles de grammaire et de transfert, les modèles logiques, etc.). Ces approches sont encore utilisées de nos jours, mais ont cédé beaucoup de terrain. En effet, devant la lenteur des progrès, l'enthousiasme et le financement retombèrent en 1966, avec la douche froide du rapport de l'ALPAC (*Automatic Language Processing Advisory Committee*), qui montra que la traduction automatique était moins précise, plus lente et plus coûteuse que la traduction réalisée par des experts. Les travaux ne reprirent sérieusement qu'à la fin des années soixante-dix lorsque les progrès de l'informatique ouvrirent des horizons jusque là inédits, permettant le développement de méthodes dites *empiriques* (sections 1.4 et 1.5). Avec l'avènement de ces méthodes, on renonce à « inculquer » à la machine les règles des langages et de la traduction sous forme de grammaire et de lexique. La tâche est en effet titanesque : toutes ces règles ne sont pas formalisées, et on n'est pas sûr de pouvoir les écrire aussi vite qu'elles évoluent. Au contraire, on développe des algorithmes qui, d'une grande quantité de textes déjà traduits (les corpora bilingues — section 1.5.6) vont extraire, pour les réutiliser, des exemples de traductions (approche par *analogie*, section 1.4) ou déterminer les lois statistiques qui caractérisent la correspondance entre un élément de texte et sa traduction (approche *statistique*, section 1.5).

1.2 Généralités sur la traduction

C'est presque un poncif mais il est bon de le rappeler : la traduction est un problème particulièrement complexe même pour un expert. Traduire un document, c'est bien plus que remplacer les mots d'une langue par les mots d'une autre langue, même quand les deux langues ont des notions similaires de « mot », ce qui n'est pas toujours le cas. Les mots peuvent avoir plusieurs sens. L'ordre des mots peut être modifié, la structure de la phrase totalement transformée. Le contexte, la culture de l'auteur et des destinataires interviennent également : quelles sont les connotations d'un terme ? y a-t-il des sous-entendus ? des références ? Les expressions sont une difficulté supplémentaire : « L'habit ne fait pas le moine » sera traduit en anglais par « Don't juge a book by its cover » (« Ne jugez pas un livre à sa couverture »), et non par « The dress doesn't make the monk ». Les exemples de telles difficultés sont innombrables. Dans les grandes lignes, pour un humain, la traduction d'un texte se déroule en trois phases :

1. Lire et comprendre le texte source.
2. Transformation de la structure du texte pour l'adapter à la langue cible.
3. Génération (écriture) d'un texte dans la langue cible.

Ces étapes sont illustrées par le fameux triangle de Vauquois (figure 1.1), nommé ainsi d'après Bernard Vauquois, un des pionniers de la TA dans les années 1970 : le traducteur commence par comprendre le texte (c'est *l'analyse* dans le triangle). Il arrive ainsi à une représentation mentale du sens du texte, qui correspond au sommet du triangle (certains systèmes génèrent une telle représentation, soit sous forme abstraite, soit en utilisant une *interlangue*, c'est-à-dire un énoncé rédigé dans un « langage » artificiel). Enfin, le traducteur écrit un texte qu'il estime être la meilleure traduction du texte source dans la langue désirée.

Malheureusement, la compréhension d'un texte en langue naturelle est encore largement inaccessible aux méthodes actuelles. La plupart des systèmes actuels de traduction automatique se contente donc de niveaux d'analyse intermédiaires, c'est-à-dire de règles de transfert entre des représentations de plus ou moins haut niveau de la langue source, ou empruntent un raccourci dans ce triangle : la « TA directe », c'est-à-dire les méthodes traduisant le texte comme on transformerait un signal, sans analyse. Dans les sections qui suivent, je vais m'attacher à donner un aperçu des méthodes à base de règles de transfert, avant de m'attarder plus longuement sur

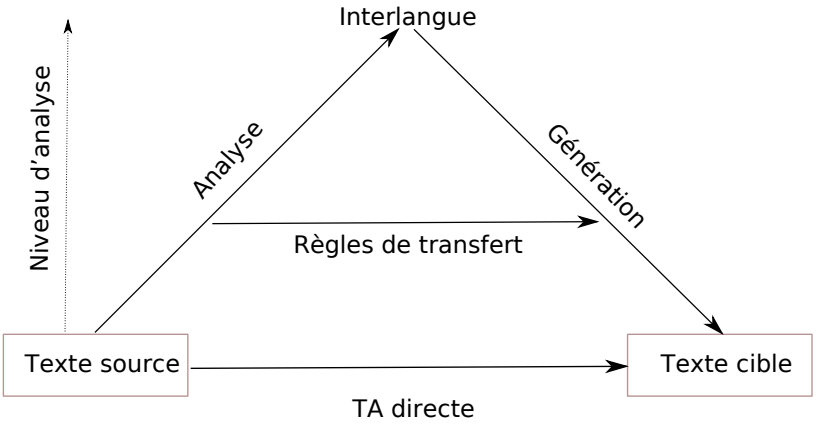


FIGURE 1.1 – Le triangle de Vauquois.

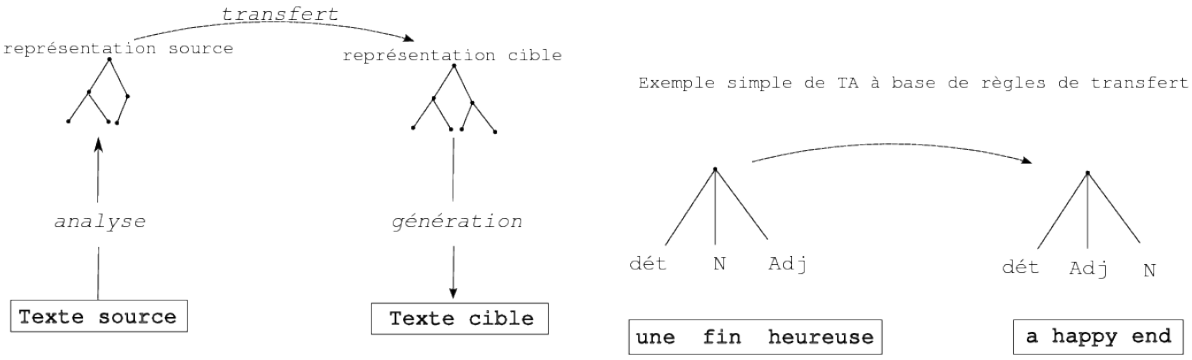
la TA directe statistique, sur laquelle sont fondés les systèmes que j'ai utilisés au cours de mon doctorat. Évidemment, la distinction entre les différentes approches est parfois floue, et il existe des systèmes hybrides.

1.3 Les approches expertes

Cette première famille regroupe les méthodes utilisant des règles établies par des experts. La traduction selon ces méthodes se déroule en trois phases similaires à celles exposées dans la section précédente :

1. *Analyse de la phrase source.* Selon la méthode employée, des unités plus ou moins fines sont distinguées dans la phrase source : morpho-syntaxiques, grammaticales ou sémantiques. Cette étape peut s'appuyer, selon la méthode, sur des lexiques de formes fléchies, des règles de grammaire et de syntaxe, un étiquetage en Parties-du-Discours, ou des modèles sémantiques.
2. *Transfert.* Les unités de la phrase source sont transformées en unités équivalentes de la langue cible à l'aide de *règles de transfert*. Ces règles peuvent être similaires à des lexiques si l'analyse de la première étape était très superficielle, ou plus complexes, impliquant des transformations d'arbre comme illustré dans la figure 1.2.
3. *Génération.* La représentation abstraite de la phrase traduite est ensuite transformée en texte. Cette étape implique la génération de termes fléchis à partir d'éléments abstraits, et souvent un réordonnancement des termes.

Les niveaux d'analyse peuvent aller de la simple analyse morpho-syntaxique (c'est l'analyse la plus superficielle, presque à la base du triangle de Vauquois, juste au dessus de la TA directe) à l'interlangue, une analyse tellement sophistiquée que la représentation de la phrase source peut être vue comme une traduction dans une langue intermédiaire, universelle. Le transfert



(a) Principe général [Jura 00]

(b) Illustration sur une expression simple

FIGURE 1.2 – Traduction automatique à base de règles

n'a alors plus lieu d'être, on passe directement de l'analyse à la génération. En pratique, les systèmes effectivement implémentés, à ma connaissance, ne dépassent pas un niveau intermédiaire de représentation, basé sur la syntaxe, la grammaire, et éventuellement certaines notions de sémantique.

Plusieurs systèmes commerciaux utilisent aujourd'hui une approche à base de règles, comme Babelfish [3] ou Systran [Sene 01]. Ces systèmes reposent sur des lexiques et des règles déterminées et encodées par des experts. Si cela leur permet de générer des traductions d'excellente qualité, c'est au prix d'un effort (et donc d'un temps et d'un coût de développement) considérable, ou d'une couverture réduite. De plus, ces approches sont difficilement robustes aux phrases mal construites comme on peut en trouver à l'oral ou dans des documents rédigés sans soin. C'est pour résoudre ces limitations qu'ont été développées les *approches empiriques* qui seront présentées dans les prochaines sections : le principe de ces approches est de tirer parti des gros volumes de données de plus en plus disponibles (des documents déjà traduits, par exemple les transcriptions des sessions du Parlement Européen) et des capacités de calcul toujours croissantes des ordinateurs pour déterminer automatiquement les règles de transfert entre les différentes langues.

1.4 La traduction par analogie

La traduction par analogie est née en 1984 de l'idée que les traducteurs traduisaient souvent en se référant à des exemples [Naga 84]. On retrouve cette idée dans certains outils d'aide à la traduction utilisés par les professionnels, comme par exemple les mémoires de traduction, qui étant donné une phrase ou des mots sources, suggèrent à l'utilisateur des traductions qu'il a déjà utilisées.

L'*approche par analogie* ou *traduction à base d'exemples* consiste à, dans un premier temps, constituer une *base d'exemples* la plus fournie possible, en tout cas couvrant le mieux possible le domaine ciblé. Cette base de données contient des couples composés d'un élément de la langue source (une phrase, une séquence de mots, une expression ou un mot) et sa traduction dans la langue cible. Une fois cette base constituée, une traduction se déroule en deux étapes :

1. *Correspondance.* Étant donnée une phrase à traduire **s**, le système recherche dans la base

3. `http://fr.babelfish.yahoo.com`

d'exemples l'ensemble d'éléments $\mathcal{C}(\mathbf{s})$ qui s'en rapprochent le plus selon une certaine *mesure de similarité* d (c'est-à-dire un voisinage de $\mathbf{s}$, si on interprète d comme une distance).

2. La deuxième phase est *l'adaptation.* Soit $\mathcal{T}(\mathcal{C}(\mathbf{s}))$ l'ensemble des traductions des éléments de $\mathcal{C}(\mathbf{s})$ dans la base d'exemples. Les éléments de $\mathcal{T}(\mathcal{C}(\mathbf{s}))$ sont adaptés et réarrangés pour générer une phrase dans la langue cible.

Le choix d'une mesure de similarité d et le choix des modèles d'adaptation sont donc cruciaux. Les auteurs de [Veal 97], par exemple, mesurent la similarité des phrases entières pour le système GAIJIN. Le système PANGLOSS [Brow 96], lui, segmente la phrase à traduire en séquences, et cherche dans la base d'exemples les entrées similaires à ces séquences. [Naga 84, Alp 08] génèrent les arbres représentant la syntaxe de la phrase et des exemples à l'aide d'un analyseur syntaxique, et recherchent les arbres les plus semblables. Le rôle de la phase d'adaptation est de générer une phrase intelligible dans la langue cible. Il faut donc s'appuyer sur un modèle de langue pour réarranger les traductions des éléments identifiés, comme décrit par exemple dans [Brow 96]. Yves Lepage propose dans [Lepa 05] une approche originale basée sur la notion *d'analogie proportionnelle*, qui repose sur une formalisation rigoureuse des analogies entre phrases, et propose une méthode pour transférer les analogies de la langue source en des analogies de la langue cible.

1.5 La traduction automatique statistique

Nous abordons maintenant un nouveau pan des méthodes empiriques, la *traduction automatique statistique*. C'est une approche largement répandue, utilisée notamment par le système de traduction de Google. Toutes nos expériences ont été réalisées en utilisant les logiciels Giza++ [Och 03] et Moses [Koeh 07a], qui implémentent cette approche. Je vais donc l'exposer plus en détail. Je commencerai par présenter les principes qui fondent la modélisation statistique du langage (section 1.5.1). Je présenterai ensuite les modèles de traduction, qui permettent de déterminer les éléments de la phrase traduite en fonction de la phrase source (section 1.5.2) et de langue, qui permettent d'assurer que la phrase dans la langue cible est correcte (section 1.5.4). Nous verrons comment, étant donnés ces modèles et une phrase source, les systèmes déterminent la traduction *la plus probable*. Enfin, je donnerai un aperçu des données nécessaires à l'estimation des paramètres de ces modèles.

1.5.1 Modélisation statistique de la traduction

L'approche statistique de la traduction automatique consiste à représenter les phrases dans la langue source (la langue à traduire) et celles de la langue cible (la langue dans laquelle on souhaite traduire) par deux variables aléatoires $\mathbf{S}$ et $\mathbf{T}$. Ces variables sont à valeurs dans $\mathcal{V}_S^*$ et $\mathcal{V}_T^*$ respectivement, c'est-à-dire qu'elles sont des séquences de mots dans les vocabulaires source ($\mathcal{V}_S$) et cible ($\mathcal{V}_T$, T pour *target*) respectivement. Un vocabulaire est l'ensemble des mots (y compris les formes fléchies) et des symboles de ponctuation qui existent dans une langue. On y ajoute des symboles de début et fin de phrase `<s>` et `</s>`, ainsi qu'un symbole spécial `UNK` (*unknown*) représentant tous les mots rencontrés qui ne feraient pas partie du vocabulaire. Pour la traduction de la parole (partie III) il pourra être enrichi de symboles particuliers tels que les *silences*, les *bruits*, etc. Le i-ème mot d'une phrase est désigné par la variable aléatoire S_i (respectivement T_i), et sa longueur par $Len(\mathbf{S})$ (respectivement $Len(\mathbf{T})$) [4]. La traduction

4. Les caractères majuscules désignent des variables aléatoires, et les minuscules correspondantes leurs réalisations. Les symboles en gras sont des vecteurs ou des séquences, les symboles en maigre des scalaires. Les lettres

automatique consiste, étant donnée une phrase source $\mathbf{s}$, à déterminer une phrase $\mathbf{t}^*$ dans la langue cible telle que :

$$\mathbf{t}^* = \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} P_\Theta(\mathbf{T} = \mathbf{t} | \mathbf{S} = \mathbf{s}) \tag{1.1}$$

où Θ désigne l'ensemble des paramètres du modèle, qui représentent les connaissances dont on dispose. En français, cela signifie qu'on cherche la traduction *la plus probable* de $\mathbf{s}$ étant données les connaissances dont on dispose. En l'état, le problème est trop compliqué pour être correctement modélisé. On va donc transformer cette équation en utilisant la règle de Bayes pour distinguer deux sous-problèmes :

$$\begin{aligned} \mathbf{t}^* &= \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} P_\Theta(\mathbf{T} = \mathbf{t} | \mathbf{S} = \mathbf{s}) \\ &= \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} \frac{P_\Theta(\mathbf{S} = \mathbf{s} | \mathbf{T} = \mathbf{t}) \times P_\Theta(\mathbf{T} = \mathbf{t})}{P_\Theta(\mathbf{S} = \mathbf{s})} \\ &= \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} P_\Theta(\mathbf{S} = \mathbf{s} | \mathbf{T} = \mathbf{t}) \times P_\Theta(\mathbf{T} = \mathbf{t}) \end{aligned} \tag{1.2}$$

La dernière simplification est possible car la recherche d'une traduction de $\mathbf{s}$ se fait bien entendu à $\mathbf{s}$ fixé. Le dénominateur $P_\Theta(\mathbf{S} = \mathbf{s})$ est donc constant et peut être ignoré lors du calcul de l'arg max. On distingue donc ici :

- $P_\Theta(\mathbf{S} = \mathbf{s} | \mathbf{T} = \mathbf{t})$ (qui sera abrégé en $P(\mathbf{s}|\mathbf{t})$ lorsqu'il n'y a pas ambiguïté) : le modèle de traduction (section 1.5.2),
- $P_\Theta(\mathbf{T} = \mathbf{t})$ (qui sera abrégé en $P(\mathbf{t})$) : le modèle de langue cible (section 1.5.4).

Étant données ces deux distributions, on peut donc déterminer la traduction la plus probable de $\mathbf{s}$, c'est-à-dire estimer le *argmax* de la formule 1.2. Ce calcul est effectué par le *décodeur*.

Historiquement, cette approche est celle dite du *canal bruité* : $\mathbf{s}$ est assimilé à un signal observé qui a subi des « erreurs de transmissions » (ici, elles représentent le transfert dans une autre langue). On veut retrouver le signal original $\mathbf{t}$, sachant que l'on connaît la distribution des perturbations $P(\mathbf{s}|\mathbf{t})$ et la distribution du signal original $P(\mathbf{t})$. Cette approche repose sur l'hypothèse que le signal original peut être modélisé comme une chaîne de Markov discrète à états finis (les états étant les mots). Cette hypothèse ne rend pas parfaitement compte de la complexité des langues naturelles, ne serait-ce que parce qu'il existe des dépendances longue distance (c'est-à-dire que la probabilité d'apparition d'un mot ne dépend pas uniquement du mot qui le précède) et parce que la probabilité d'une perturbation (la traduction d'un mot) ne dépend pas seulement du mot courant mais aussi de l'historique (le contexte). Ce formalisme est cependant assez puissant pour permettre une modélisation assez efficace de la traduction. Il domine également en reconnaissance automatique de la parole, problème formellement très similaire : le signal original est la transcription de la phrase prononcée, et le signal perturbé est le son enregistré. Ce procédé est illustré dans la figure 1.3.

Les modèles de langue et de traduction seront donc estimés séparément à partir de grandes quantités de textes (voir la section 1.5.6 pour plus d'explications sur ces **corpora**). Pour tenir

cursives sont des ensembles. Ces notations seront reprises et complétées dans la partie II, section 1.

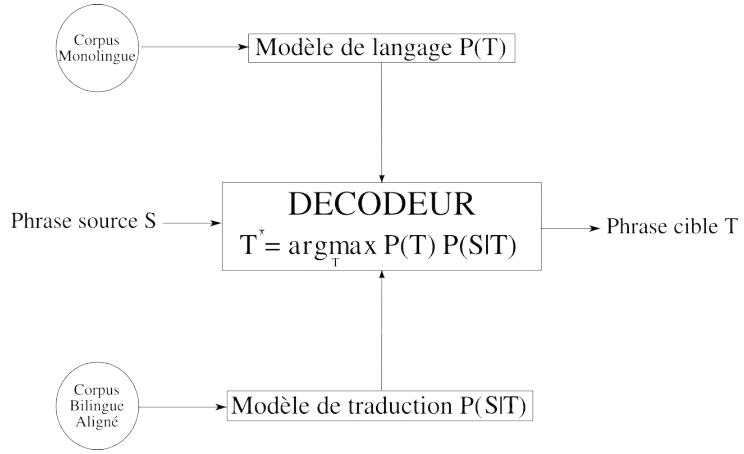


FIGURE 1.3 – Schéma de principe de la traduction automatique statistique selon le paradigme du canal bruité.

compte du fait que les modèles sont plus ou moins bien estimés, et que les estimations de probabilités pourraient être biaisées, on ajoute à l'équation 1.2 un facteur de pondération λ :

$$\mathbf{t}^* = \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} P_\Theta(\mathbf{S} = \mathbf{s} | \mathbf{T} = \mathbf{t}) \times P_\Theta(\mathbf{T} = \mathbf{t})^\lambda \tag{1.3}$$

Les auteurs de [Och 02] proposent de ne pas se restreindre à un modèle de traduction et un modèle de langue, mais d'utiliser une multitude de modèles, chacun pouvant être vu comme une source d'information supplémentaire sur le processus de traduction :

$$\mathbf{t}^* = \arg\max_{\mathbf{t} \in \mathcal{V}_T^*} \prod_i h_i(\mathbf{s}, \mathbf{t})^{\lambda_i}$$

On voit que la formulation du canal bruité (formule 1.2) est un cas particulier de celle-ci, en utilisant seulement deux modèles : $h_1(\mathbf{s}, \mathbf{t}) = P(\mathbf{s}|\mathbf{t})$ et $h_2(\mathbf{s}, \mathbf{t}) = P(\mathbf{t})$.

Les paramètres des modèles sont appris sur un corpus bilingue aligné (c'est-à-dire un ensemble de phrases dans la langue source et leurs traductions dans la langue cible, réalisées par des humains — section 1.5.6), et leurs poids sont estimés selon l'algorithme du maximum d'entropie. Je ne m'étendrai pas plus sur ce modèle, car j'ai pour mes travaux suivi l'approche du canal bruité classique.

1.5.2 Les modèles de traduction

Je vais dans cette section expliquer la forme du modèle de traduction, c'est-à-dire l'estimation de $P(\mathbf{s}|\mathbf{t})$ (formule 1.2). On peut imaginer une multitude de formes pour ces modèles. Je décrirai ceux proposés dans [Brow 93], qui sont les plus utilisés aujourd'hui. Les algorithmes pour l'entraînement de ces modèles (et de certaines variantes) sont implémentés dans le logiciel Giza++ [Och 03]. La traduction selon ces modèles est implémentée notamment dans le logiciel Moses [Koeh 07a].

Les modèles d'IBM

Les modèles d'IBM sont construits selon l'idée que chaque élément de la phrase à traduire (mot ou groupe de mots, signe de ponctuation) a un ensemble de traductions possibles (appelons les *alternatives*), chacune étant associée à une probabilité, et que la traduction d'une phrase consiste à choisir les bonnes alternatives et à les écrire dans le bon ordre. L'ordre est déterminé par la correspondance entre les termes de la traduction et les termes de la phrase source. Cette correspondance s'appelle *l'alignement*. Cet alignement n'est pas forcément monotone, c'est-à-dire que l'ordre des mots de la traduction n'est pas forcément l'ordre des mots de la phrase source avec lesquels ils sont alignés (figure 1.4). Formellement, l'alignement est une variable aléatoire cachée $\mathbf{A}$ du modèle de traduction :

$$\begin{aligned} P(\mathbf{S}=\mathbf{s}|\mathbf{T}=\mathbf{t}) &= \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} P(\mathbf{S}=\mathbf{s},\mathbf{A}=\mathbf{a}|\mathbf{T}=\mathbf{t}) \\ &= \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} P(\mathbf{A}=\mathbf{a}|\mathbf{T}=\mathbf{t})P(\mathbf{S}=\mathbf{s}|\mathbf{T}=\mathbf{t},\mathbf{A}=\mathbf{a}) \end{aligned} \tag{1.4}$$

où $\mathbf{a}$ représente l'alignement et $\mathcal{A}(\mathbf{s},\mathbf{t})$ l'ensemble des alignements possibles entre les phrases $\mathbf{s}$ et $\mathbf{t}$. Si on se restreint à la traduction basée sur les mots (la traduction basée sur les séquences sera abordée plus tard), un alignement est formellement le choix d'une fonction de $\{1,..,Len(\mathbf{s})\}$ dans $\{0,..,Len(\mathbf{t})\}$, c'est-à-dire que chaque mot de la phrase source $\mathbf{s}$ est aligné avec un et un seul mot de la phrase cible $\mathbf{t}$, mais il peut aussi n'être aligné avec aucun mot, ce qui est représenté par l'alignement avec le mot t_0 (noté $NULL$). On identifiera souvent la fonction $\mathbf{a}$ et la séquence $a_1,..,a_{Len(\mathbf{s})} = \mathbf{a}(1),..,\mathbf{a}(Len(\mathbf{s}))$. Chaque mot a une certaine probabilité d'être la traduction du mot avec lequel il est aligné (donnée par la table de traduction estimée sur un corpus d'apprentissage). Un exemple est donné dans la figure 1.4.

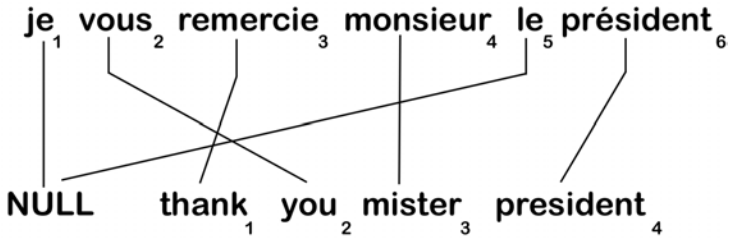


FIGURE 1.4 – Exemple d'alignement entre une phrase anglaise et une phrase française.

Or les variables aléatoires $\mathbf{S}$, $\mathbf{T}$ et $\mathbf{A}$ sont elles-mêmes des séquences de variables aléatoires S_i, T_j et A_k, et la longueur des séquences $(S_i)_i$ et $(A_k)_k$ est la variable aléatoire $Len(\mathbf{S})$. Selon le théorème de Bayes, la probabilité de $\mathbf{s}$ sachant $\mathbf{t}$ s'écrit donc dans ce modèle, en toute généralité :

$$\begin{aligned} P(\mathbf{s}|\mathbf{t},\mathbf{a}) &= \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} P(\mathbf{a}|\mathbf{t}) \times P(Len(\mathbf{s})|\mathbf{t},\mathbf{a}) \times P(\mathbf{s}|\mathbf{t},\mathbf{a},Len(\mathbf{s})) \\ &= P(Len(\mathbf{s})|\mathbf{t}) \times \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} P(\mathbf{a}|\mathbf{t}) \times P(\mathbf{s}|\mathbf{t},\mathbf{a},Len(\mathbf{s})) \\ &= P(Len(\mathbf{s})|\mathbf{t}) \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} \prod_{i=1}^{Len(\mathbf{s})} P(a_i|\mathbf{a}_1^{i-1},\mathbf{s}_1^{i-1},Len(\mathbf{s}),\mathbf{t})P(s_i|\mathbf{a}_1^{i},\mathbf{s}_1^{i-1},Len(\mathbf{s}),\mathbf{t}) \end{aligned} \tag{1.5}$$

Les auteurs de [Brow 93] proposent cinq modèles, connus sous le nom de modèles IBM-1 à 5, de complexité croissante. La différence principale entre ces modèles réside dans la façon d'appréhender l'alignement. En pratique, c'est le modèle IBM-5 ou une variante qui est en général utilisé pour le décodage. Les autres modèles sont utilisés lors de l'estimation des paramètres : chaque modèle ayant strictement moins de paramètres que le suivant, on commence par entraîner le modèle le plus simple, et les paramètres déterminés servent de valeur initiale pour l'entraînement du modèle suivant.

Le modèle IBM-1 : ce premier modèle est le plus simple, presque simpliste. Il repose sur trois facteurs :

- La distribution de probabilité des longueurs des phrases sources, $P(Len(\mathbf{s})|\mathbf{t})$.
- La distribution de probabilité des alignements, $P(\mathbf{a}|\mathbf{t})$.
- Les probabilités de traduction mot-à-mot $\{P(s|t)\}_{(s,t)\in\mathcal{V}_S\times\mathcal{V}_T}$ ou *table de traduction.*

Ce modèle fait deux hypothèses fortement simplificatrices :

1. La probabilité de la longueur est constante (formellement, $P(Len(\mathbf{s})|\mathbf{t}) = \frac{1}{L_{max}}$ pour $Len(\mathbf{s}) \in \{1, .., L_{max}\}$ avec L_{max} arbitrairement et suffisamment grand, et est nulle au-delà. Mais sa valeur n'importe pas pour le calcul d'une traduction, il suffit de savoir qu'elle est constante. Nous la noterons ϵ dans ce qui suit).
2. Tous les alignements sont équiprobables. La probabilité d'un alignement est donc uniforme étant données les longueurs des phrases source et cible :

$$\forall \mathbf{a} \in \mathcal{A}(\mathbf{s}, \mathbf{t}).\, P(\mathbf{a}|\mathbf{t}) = \frac{1}{(1 + Len(\mathbf{t}))^{Len(\mathbf{s})}}$$

La probabilité de $\mathbf{s}$ sachant $\mathbf{t}$ selon le modèle IBM-1 est donc :

$$\begin{aligned} P(\mathbf{s}|\mathbf{t}) &= P(Len(\mathbf{s})|\mathbf{t}) \times \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} P(\mathbf{a}|\mathbf{t}) P(\mathbf{s}|\mathbf{t}, \mathbf{a}) \\ &= \frac{\epsilon}{(1 + Len(\mathbf{t}))^{Len(\mathbf{s})}} \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{t})} \prod_{i=1}^{Len(\mathbf{s})} P(s_i|t_{\mathbf{a}(i)}) \end{aligned}$$

Nous passerons les détails (la preuve est donnée dans [Brow 93]), mais on montre que cette expression peut s'écrire :

$$P(\mathbf{s}|\mathbf{t}) = \frac{\epsilon}{(1 + Len(\mathbf{t}))^{Len(\mathbf{s})}} \prod_{i=1}^{Len(\mathbf{s})} \sum_{i=0}^{Len(\mathbf{t})} \theta(s_i, t_j) \tag{1.6}$$

L'ensemble des paramètres $\{\theta(s,t)\}_{(s,t)\in\mathcal{V}_S\times\mathcal{V}_T}$ constitue la table de traduction, et sera estimé par l'algorithme *Expectation-Maximization* [Demp 77] sur le corpus d'apprentissage.

Ce modèle sera également utilisé comme mesure de confiance (partie II, sections 3.1.6 et 3.2.3).

Le modèle IBM-2 : la nouveauté introduite par ce modèle est d'assigner une probabilité non uniforme aux différents alignements. On fait ici l'hypothèse que $P(a_i|\mathbf{a}_1^{i-1}, \mathbf{s}_1^{i-1}, Len(\mathbf{s}), \mathbf{t})$ (équation 1.6) ne dépend que de $Len(\mathbf{s})$, $Len(\mathbf{t})$, i et a_i. Ce qui nous donne un nouvel ensemble de paramètres, les *probabilités d'alignement* que nous noterons α :

$$\alpha(a_i, i, Len(\mathbf{s}), Len(\mathbf{t})) \stackrel{def}{=} P(a_i|\mathbf{a}_1^{i-1}, \mathbf{s}_1^{i-1}, Len(\mathbf{s}), \mathbf{t})$$

Les nouveaux paramètres $\alpha(...)$ sont donc des approximations des « vraies » probabilités d'alignement, au sens où on fait l'hypothèse que la position a_i du mot de la phrase cible avec lequel est aligné le mot de la phrase source s_i ne dépend que de la position i, et des longueurs des phrases source et cible. L'équation 1.6 (approchée par l'équation 1.6 dans le modèle IBM-1) devient donc :

$$P(\mathbf{s}|\mathbf{t}) = \epsilon \sum_{a_1=0}^{Len(\mathbf{t})} ... \sum_{a_{Len(\mathbf{s})}=0}^{Len(\mathbf{t})} \prod_{i=1}^{Len(\mathbf{s})} \theta(s_i, t_{a_i}) \alpha(a_i, i, Len(\mathbf{s}), Len(\mathbf{t}))$$

avec ϵ un facteur de normalisation.

Une variante du modèle IBM-2 est le modèle HMM proposé par [Voge 96]. Il repose sur l'observation que des mots voisins dans $\mathbf{s}$ ont tendance a être alignés avec des mots voisins dans $\mathbf{t}$. Cela signifie que a_i a des chances d'être « proche » de a_{i-1}. Pour prendre en compte cette dépendance, l'alignement est modélisé par une chaîne de Markov, c'est-à-dire qu'on peut écrire, dans l'équation 1.6 :

$$P(a_i|\mathbf{a}_1^{i-1}, \mathbf{s}_1^{i-1}, Len(\mathbf{s}), \mathbf{t}) = P(a_i|a_{i-1}, Len(\mathbf{t}))$$

Les modèles IBM-3, 4 et 5. Ces modèles introduisent progressivement des sophistications supplémentaires :

- *Fertilité :* un mot dans une langue n'est pas toujours traduit par un seul mot dans une autre. Par exemple, le terme français *bourse* peut être traduit en anglais par *purse*, *scholarship* ou *stock market*, selon le sens qu'il prend. Le paramètre $n(\phi|t = \text{bourse})$ est la probabilité que *bourse* soit aligné avec ϕ mots. Dans notre exemple on pourrait avoir $n(1|t = \text{bourse}) = 2/3$ et $n(2|t = \text{bourse}) = 1/3$.
- *Distorsion :* un autre ensemble de paramètres modélise les positions des mots de $\mathbf{s}$ en fonction des mots de $\mathbf{t}$ avec lesquels ils sont alignés, ainsi que les écarts de position entre les mots de $\mathbf{s}$ qui sont alignés avec un même mot t.

Nous ne présenterons pas ici les équations complexes qui décrivent ces modèles.

Les modèles de triggers

Je vais décrire ici un autre modèle de traduction, qui ne repose pas sur la notion d'alignement mais sur la notion de *cooccurrence* ou de *déclenchement* d'un mot par un autre (*trigger*). Il a été décrit notamment dans [Lave 07, Lave 08]. Les modèles de triggers peuvent être vus comme une généralisation des modèles de cache (section 1.5.4), qui estiment la probabilité d'apparition d'un mot en fonction de sa fréquence d'apparition dans l'historique. La spécificité des modèles de triggers est que la probabilité d'apparition d'un mot ne dépend pas seulement de l'apparition de ce mot dans l'historique, mais de celle d'une liste de mots qui lui sont « associés » (les

déclencheurs). Cette propriété permet d'utiliser les triggers pour construire des modèles de traduction, et pas seulement des modèles de langue.

Il existe plusieurs méthodes pour déterminer les paires de mots (déclencheur, déclenché). Les auteurs de [Lave 07] proposent d'utiliser *l'information mutuelle*. L'information mutuelle mesure la quantité d'information que la connaissance d'une variable aléatoire apporte au sujet d'une autre variable aléatoire. Formellement, si X et Y sont deux variables aléatoires à valeurs dans les ensembles discrets $\mathcal{X}$ et $\mathcal{Y}$ respectivement :

$$I(X,Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

Dans le cas de la modélisation de la traduction, S est une variable aléatoire représentant un mot de la phrase source, T est une variable aléatoire représentant un mot de la phrase traduite, $P(s)$ la probabilité que le mot s soit présent dans une phrase de la langue source, $P(t)$ la probabilité que le mot t soit présent dans une phrase de la langue cible, et $P(s,t)$ la probabilité que t soit présent dans la traduction d'une phrase contenant s (on parle de *cooccurrence*). Pour construire la table de traduction, on considère la contribution de chaque paire de mots à l'information mutuelle totale. La table de traduction est donc l'ensemble :

$$\left\{\left(s,t,\theta(s,t) \stackrel{def}{=} \lambda(s)P(s,t)\log\left(\frac{P(s,t)}{P(s)P(t)}\right)\right)\right\}_{s\in\mathcal{V}_S,t\in\mathcal{V}_T}$$

Avec $\lambda(s)$ un facteur de normalisation assurant que $\sum_t \theta(s,t) = 1$.

En pratique, pour éviter d'obtenir des tables trop grosses et contenant surtout des triplets avec des scores très bas, on ne gardera, pour chaque mot du vocabulaire source, que les meilleures traductions (voir les détails dans [Lave 07]). Notons que même si $P(s,t)\log\left(\frac{P(s,t)}{P(s)P(t)}\right)$ est élevé, t n'est pas forcément une traduction de s. Par exemple, la présence du mot « économie » dans la phrase source est un bon indicateur de la présence du mot « crisis » dans la traduction. « économie » est donc un déclencheur plausible de « crisis », bien qu'ils ne soient pas traduction l'un de l'autre. Ce modèle sera également utilisé comme mesure de confiance (partie II, sections 3.1.5 et 3.2.2).

D'autres approches des modèles de triggers existent. Les auteurs de [Maus 09], par exemple, utilisent non pas un mais deux mots déclencheurs. Leur modèle est basé sur la probabilité que le mot déclenché apparaisse dans la traduction, sachant que les déclencheurs sont présents dans la phrase source :

$$\{(s_1,s_2,t,P(t|s_1,s_2))\}_{s_1,s_2\in\mathcal{V}_S,t\in\mathcal{V}_T}$$

Les auteurs de [Kim 03] proposent d'utiliser l'information mutuelle pour déterminer des triggers basés sur les cooccurrences de mots à l'échelle du document : $P(s)$ est la probabilité que le mot s apparaisse dans le document (dans leur cas, un article) et $P(s,t)$ est la probabilité que le terme t apparaisse dans la traduction d'un article contenant s. Leur modèle tient également compte de l'information apportée par le fait qu'un terme *n'apparaît pas* dans un document.

1.5.3 Les modèles à base de segments

Une extension très intéressante des modèles exposés dans les sections précédentes est *la traduction par segments*. Dans les langues indo-européennes, telles que l'anglais et le français par exemple, une phrase peut être vue comme une succession de groupes de mots : expression lexicale,

groupe nominal, groupe verbal, proposition relative, etc. Ces groupes de mots sont en général de meilleures unités de traduction que les mots pris isolément : en effet, considérer un groupe de mots dans son ensemble peut permettre de lever des ambiguïtés, ou d'identifier une expression qui ne sera pas traduite mot-à-mot. Cela permet aussi de prendre en compte implicitement les réarrangements locaux, et de diminuer la complexité de l'alignement. Prenons l'exemple de la paire de phrases suivante :

français	`J'ai vendu mon bateau de pêche.`
anglais	`I sold my fishing boat.`

`J'ai vendu` (trois mots) est traduit par `I sold` (deux mots), et `mon bateau de pêche` (quatre mots) par `my fishing boat` (trois mots). Le meilleur alignement mot-à-mot de cette phrase pourtant simple implique donc des réarrangements, et est assez peu satisfaisant, notamment parce que `de` n'est pas traduit par le mot avec lequel il est aligné (figure 1.5) :

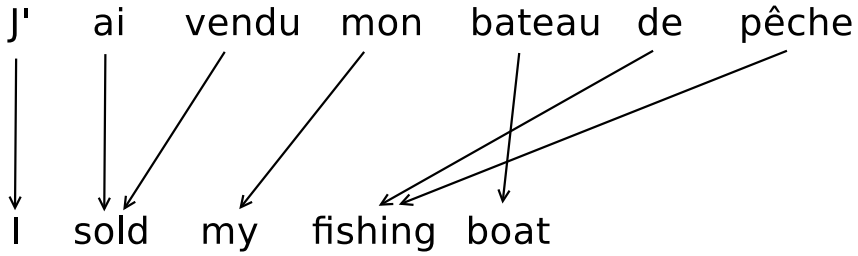


FIGURE 1.5 – Alignement mot-à-mots d'une phrase simple.

Si on considère les groupes de mots dans leur ensemble, en revanche, l'alignement devient trivial (figure 1.6) :

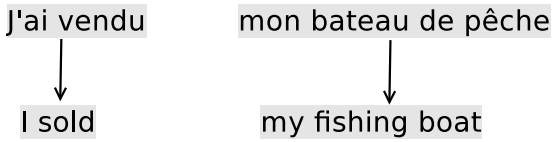


FIGURE 1.6 – Alignement par séquences d'une phrase simple avec réarrangements locaux.

La contrepartie de cette simplification est que la table de traduction devra contenir non plus les alternatives de traduction de chaque mot du vocabulaire, mais les traductions de chaque segment qui aura été retenu lors de l'apprentissage, ce qui représente une combinatoire bien plus élevée. De plus, les explications et les exemples donnés ci-dessus reposent sur des notions linguistiques (groupe verbal, groupe nominal, etc.). Or de telles notions sont absentes des modèles statistiques. La très grande majorité des segments retenus par les algorithmes d'apprentissage n'ont donc aucun sens pour un humain : ils auront été sélectionés parce qu'ils permettent une bonne représentation des données d'apprentissage au sens statistique du terme, c'est-à-dire une perplexité basse.

Une des contributions fondatrices dans ce domaine est [Och 99, Och 00]. Dans ces articles, Och *et al.* proposent, d'une part, l'usage de *patrons d'alignement* utilisant les classes grammaticales des mots, afin de modéliser les réarrangements locaux fréquemment observés dans le corpus d'apprentissage (par exemple, pour une traduction du français vers l'anglais, le segment

<NOM> <ADJECTIF> sera souvent aligné avec un segment <ADJECTIF> <NOM>) ; et d'autre part, d'appliquer à l'échelle de la phrase un réarrangement de ces segments.

Le modèle de triggers a également été adapté pour utiliser des segments : dans ce cas, l'algorithme d'apprentissage estime d'abord l'information mutuelle entre les segments exactement comme il le faisait avec les mots, puis détermine les segments pertinents à incorporer dans la table de traduction en utilisant un algorithme de *recuit simulé* (voir [Lave 10] pour les détails).

L'avènement des méthodes utilisant les segments a été un progrès majeur pour la traduction automatique statistique. Ces méthodes sont aujourd'hui utilisées par les systèmes les plus performants [Koeh 03].

1.5.4 Les modèles de langue

Le rôle du modèle de langue est de contraindre la recherche d'une traduction à des suites de mots $\mathbf{t}$ qui sont des phrases correctes dans la langue cible. Idéalement, toute phrase correcte, indépendamment de la phrase source dont elle doit être une traduction, aurait une probabilité $P(\mathbf{t}) = 1$ et toute phrase incorrecte (mal formée) aurait une probabilité nulle $P(\mathbf{t}) = 0$. Mais d'une part, il n'existe à l'heure actuelle aucun modèle permettant d'affirmer, parmi l'infinie variété des phrases possibles, lesquelles sont correctes ou non ; et d'autre part, la notion même de phrase correcte est ambigüe (voir la partie II sur les mesures de confiance). Le modèle de langue assignera donc une probabilité à chaque phrase dans l'intervalle $[0, 1]$.

Dans ce formalisme la probabilité d'une phrase est modélisée de la façon suivante :

$$P(\mathbf{t}) = \prod_{j=1}^{Len(\mathbf{t})} P(t_j|t_1, .., t_{j-1})$$

C'est-à-dire que la probabilité d'une phrase est le produit de la probabilité des mots qui la composent (on devrait parler plus précisément de la *vraisemblance de la séquence* $t_1, .., t_{Len(\mathbf{t})}$), et la probabilité d'un mot dépend uniquement de son historique (son contexte gauche). Le rôle des modèles de langue sera donc d'estimer les probabilités des mots étant donnés leurs historiques, et la vraisemblance de la phrase en découlera. Lorsque cela n'est pas ambigu, l'historique $\mathbf{h} = t_1, .., t_j$ sera noté $\mathbf{t}_1^j$.

Les modèles n-grammes

L'ensemble des historiques est $\mathcal{V}_T^*$, c'est-à-dire l'ensemble des séquences finies de mots du vocabulaire. Il est donc infini. On ne peut donc pas estimer directement l'ensemble des probabilités :

$$\{P(t|h)\}_{(t,h)\in\mathcal{V}_T\times\mathcal{V}_T^*}$$

Les modèles n-grammes reposent sur l'hypothèse que les séquences de mots peuvent être modélisées par un processus de Markov d'ordre $n-1$, c'est-à-dire que la probabilité d'un état (un mot) dépend uniquement des $n-1$ états précédents. La probabilité d'un mot dans un modèle n-gramme s'écrit donc :

$$P(t_j|\mathbf{t}_1^{j-1}) = P(t_j|\mathbf{t}_{j-n+1}^{j-1})$$

L'espace des historiques est donc réduit à $\cup_{i=0}^{n-1}\mathcal{V}_T^i$ (on considérant les historiques de longueur inférieure à $n-1$, notamment en début de phrase). Il est donc fini. Sa taille reste bien trop

importante pour estimer les distributions pour tous les historiques possibles, mais lors de la phase d'apprentissage, seuls les historiques effectivement observés seront retenus.

Pour $n = 1$, on parle de modèle *unigramme*. Pour $n = 2$, de *bigramme* ; *trigramme* et *quadrigramme* pour pour $n = 3$ et $n = 4$. On utilisera plus rarement *pentagramme* pour $n = 5$. Dans le cas général, on parle simplement de n-gramme. On confond systématiquement le nom du modèle et le nom de l'évènement $\mathbf{h}w$ (l'historique $\mathbf{h}$ suivi du mot w).

Entraînement des modèles n-grammes : Entraîner un modèle n-gramme revient simplement à mesurer la fréquence avec laquelle chaque mot apparaît après chaque historique observé dans le corpus d'apprentissage, pour chaque historique observé au moins une fois :

$$\forall w \in \mathcal{V}_T, \forall \mathbf{h} \in \bigcup_{i=0}^{n-1} \mathcal{V}_T^i, P(w|\mathbf{h}) = \frac{\#\mathbf{h}w}{\#\mathbf{h}}$$

$\#\mathbf{h}$ est le nombre d'occurrences de l'historique $\mathbf{h}$ dans le corpus d'entraînement (section 1.5.6), et $\#\mathbf{h}w$ le nombre d'occurrences de $\mathbf{h}$ suivi de w.

Lissage des modèles n-gramme : la précision des estimations de probabilité ainsi réalisées dépend de la taille du corpus d'apprentissage et de la qualité de l'échantillonnage dont il résulte (un corpus d'apprentissage peut être vu comme un échantillonage de l'ensemble des textes possibles). Mais ces estimations ne seront jamais parfaites. Que l'évènement $\mathbf{h}w$ n'ait pas été observé ne signifie pas qu'il ne peut exister. On imagine aisément, par exemple, qu'avant mai 2012, aucun corpus d'apprentissage extrait des archives du journal Le Monde ne contient le trigramme `le président Hollande`. Ce n'est pourtant manifestement pas un trigramme impossible. Des méthodes dites de *lissage* ont donc été développées pour assigner une probabilité non nulle aux évènements non observés. Le principe est en général de soustraire une certaine masse de probabilité γ aux évènements observés et de la redistribuer sur les évènements non observés :

$$P_{\text{lisse}}(w_i|\mathbf{w}_1^{i-1}) = \begin{cases} P(w_i|\mathbf{w}_{i-n+1}^{i-1}) - \gamma & \text{si l'évènement } \mathbf{w}_{i-n+1}^{i} \text{ a été observé} \\ & \text{lors de l'apprentissage} \\ b(\mathbf{w}_{i-n+1}^{i-1}) \times P_{\text{lisse}}(w_i|\mathbf{w}_{i-n+2}^{i-1}) & \text{sinon} \end{cases}$$

En français, cela signifie que si la séquence $\mathbf{w}_{i-n+1}^{i}$ n'a pas été observée dans le corpus d'apprentissage, alors on estime qu'on n'a pas d'information sur la probabilité $P(w_i|\mathbf{w}_{i-n+1}^{i-1})$. Dans ce cas, on se *replie* (*backoff*) sur un historique plus court, donc moins informatif mais qui a peut-être, lui, été observé. Ce repli a un coût $b(\mathbf{w}_{i-n+1}^{i-1})$ qui dépend de l'historique initial.

Dans notre exemple, la séquence `président Hollande` n'a sans doute pas été observée non plus. La probabilité de `Hollande` après l'historique `le président` est donc :

$$\begin{aligned} P_{\text{lisse}}(\text{Hollande}|\text{le président}) &= b(\text{le président}) \times P_{\text{lisse}}(\text{Hollande}|\text{président}) \\ &= b(\text{le président}) \times b(\text{président}) \times P_{\text{lisse}}(\text{Hollande}) \end{aligned}$$

Les différentes méthodes de lissage diffèrent essentiellement par le calcul de la masse de probabilité γ retirée aux évènements observés, et par le calcul du coût b du repli. Ce dernier peut ne dépendre que de la longueur de l'historique, du nombre d"occurence de la séquence, de l'identité des mots de l'historique... Pour plus de détails sur différentes techniques de lissage, le lecteur se réfèrera avec profit à [Jeli 80, Good 53, Ney 94, Knes 95].

Le repli du modèle de langue sera utilisé comme mesure de confiance (partie II, section 3.1.1).

Modèle cache

Les modèles n-grammes fournissent la distribution de probabilité d'occurrence des mots du vocabulaire étant donné un historique en se basant sur les fréquences d'apparition. Mais ils sont incapables de prendre en compte les variations locales, qui peuvent être induites notamment par un certain thème ou un certain style. Par exemple, dans un article de journal traitant de l'actualité politique, le bigramme `un parti` pourrait avoir une probabilité 0.9 d'être suivi du mot `politique` et 0.1 d'être suivi du mot `pris` ; dans un article du même journal traitant de sujets artistiques, la probabilité pourrait être de 0.3 pour `politique` et 0.7 pour `pris`. Mais sur l'ensemble du corpus d'apprentissage (section 1.5.6), qui peut être constitué de dizaines de milliers d'articles, la probabilité peut être de 0.5 pour chacun des deux. Pourtant, selon le thème traité, on voit que l'une ou l'autre des distributions peut être bien meilleure que l'autre.

Le modèle de cache vise à adapter dynamiquement les distributions de probabilité fournies par le modèle n-gramme selon le contexte. Il repose sur l'idée que certains mots, en particulier les mots les plus porteurs de sens (noms et verbes notamment, par opposition aux prépositions, déterminants, etc.) tendent à apparaître par « salves ». Le modèle de cache est donc un modèle unigramme calculé dynamiquement sur un historique de taille moyenne (de l'ordre du millier de mots). Combiné linéairement avec un modèle n-grammes classique, le modèle cache permet donc à ce dernier de s'adapter aux variations de style lexical et de sujet qui modifient localement les fréquences d'apparition des mots. Formellement :

$$\begin{aligned} P_{\text{cache}}(w|\mathbf{h}) &= \frac{\#_{\mathbf{h}} w}{\#\mathbf{h}} \\ P(w|\mathbf{h}) &= \lambda P_{\text{n-gramme}}(w|\mathbf{h}) + (1-\lambda) P_{\text{cache}}(w|\mathbf{h}) \end{aligned}$$

avec $\#_{\mathbf{h}} w$ le nombre d'occurrences de w dans l'historique $\mathbf{h}$, et $\#\mathbf{h}$ la taille de l'historique.

Les auteurs de [Kuhn 90] proposent un modèle cache plus sophistiqué : comme indiqué précédemment, le phénomène de « salve » concerne plutôt les mots les plus porteurs de sens que sont les verbes, les noms, à la rigueur les adverbes et les adjectifs. Les déterminants et autres prépositions ne subissent pas, ou marginalement ce phénomène. Afin de prendre ceci en compte, Kuhn *et al.* proposent de faire dépendre λ dans l'équation précédente de la catégorie grammaticale de w. Les poids sont ensuite déterminés automatiquement en optimisant la vraisemblance d'un corpus de développement. Ils montrent qu'en effet, le modèle cache obtient un poids plus important sur, notamment, les noms et les verbes.

Les modèles trigger

Les modèles de langage trigger fonctionnent sur le même principe que leurs homologues modèles de traduction (section 1.5.2) et peuvent être vus comme une généralisation des modèles de cache précédemment exposés. La différence est que les cooccurrences sont mesurées au sein d'une même phrase ou séquence de mots de longueur fixée : $P(x)$ est donc la probabilité qu'une phrase contienne le mot x et $P(x,y)$ est la probabilité qu'elle contienne le mot x et le mot y. Le score de la paire (x,y) dans ce modèle est toujours $I(x,y) = P(x,y)log(\frac{P(x,y)}{P(x)P(y)})$, éventuellement normalisée par un facteur ω pour respecter la contrainte suivante :

$$\forall x \in \mathcal{V}_S . \sum_{y \in \mathcal{V}_S} I(x,y) \times \omega = 1$$

Pour reprendre mon exemple précédent, le mot « crise » a de bonnes chances de déclencher le mot « économique » dans les articles de ces dernières années.

On peut imaginer différentes façons d'attribuer un score à une phrase ou un mot en utilisant ce modèle. Par exemple, nous utiliserons ce modèle comme mesure de confiance (partie II, sections 3.1.4 et 3.2.2) comme mesure de la cohérence globale du champ lexical de la phrase. Le score d'une phrase sera alors :

$$\frac{1}{Len(\mathbf{t}) \times (Len(\mathbf{t}) - 1)} \sum_{i=1}^{Len(\mathbf{t})} \sum_{1 \leqslant j \neq i \leqslant Len(\mathbf{t})} I(t_i, t_j)$$

Les auteurs de [Till 97a] proposent un autre modèle de triggers à interpoler avec un modèle n-gramme classique : dans un triplet $(x, y, q(y|x)) \in \mathcal{V}_T \times \mathcal{V}_T \times \mathbb{R}$, le score $q(y|x)$ est la fréquence d'apparition du terme y après un historique contenant x, observée sur le corpus d'apprentissage.

1.5.5 Le décodage

Les sections précédentes ont présenté les modèles de langue et de traduction, qui peuvent être vus comme les *connaissances* nécessaires à la traduction. Je vais ici présenter le processus qui mobilise ces connaissances pour effectuer le processus de traduction proprement dit. Ce processus s'appelle *décodage*. Le plus répandu des décodeurs pour la traduction automatique statistique est Moses [5] [Koeh 07a], successeur de Pharaoh. Outre les modèles IBM à base de mot ou de séquence, Moses permet aussi d'utiliser des modèles factorisés [Koeh 07b] ou des méthodes hybrides. Le système Moses est utilisé comme référence lors des campagnes WMT [6], et c'est ce système que nous utiliserons pour toutes nos expériences.

Le rôle du décodeur est, étant donnée une phrase source $\mathbf{s}$, de générer des hypothèses de traductions $\mathbf{t}$, de calculer la probabilité $P(\mathbf{s}|\mathbf{t}) \times P(\mathbf{t})$, et de calculer l'arg max (équation 1.2). Il serait évidemment absurde de prétendre énumérer toutes les traductions possibles pour ensuite choisir la meilleure. Non seulement il en existe une infinité, mais même en se restreignant à des traductions de longueur « réaliste » (par exemple, entre 0.25 et 4 fois la longueur de la phrase source), ce serait terriblement inefficace. Il faut donc développer une stratégie plus astucieuse. Moses met en œuvre des heuristiques et des algorithmes très sophistiqués, qu'il serait laborieux et inutile de présenter en détail ici. Je vais donc m'attacher à décrire l'idée générale de l'algorithme dans le cas d'une traduction basée sur les mots (et non les segments), en mentionnant les principales étapes et structures de données.

Moses va d'abord créer une liste *d'alternatives de traduction*, c'est-à-dire les éléments de la table de traduction qui sont des traductions possibles des éléments de la phrase source. Le tableau 1.1 illustre ce processus pour l'exemple `J'ai vendu mon bateau de pêche`.

En plus des traductions possibles, chaque mot peut-être omis lors de l'alignement. On voit que ces alternatives peuvent générer jusqu'à $(2+1)\times(4+1)\times(3+1)\times(2+1)\times(3+1)\times(3+1)\times(3+1) = 11520$ hypothèses, sans compter les différentes possibilités de réarrangement. Et ce n'est qu'un exemple jouet : une vraie table de traduction contient des dizaines d'alternatives par mot. On imagine aisément qu'il faut trouver une méthode efficace pour les représenter et les évaluer. Moses a pour cela recourt à $Len(\mathbf{s})$ piles, notons les $\mathbf{p}_1, .., \mathbf{p}_{Len(\mathbf{s})}$. $\mathbf{p}_k$ contient les hypothèses de traduction partielles (c'est-à-dire telles que tous les mots de la phrase source ne sont pas alignés avec un mot de la phrase cible ou avec `NULL`) couvrant k mots de $\mathbf{s}$. Les hypothèses de la pile $k+1$ sont construites à partir des hypothèses de la pile k, en les étendant mais aussi en les fusionnant et en supprimant les plus mauvaises afin de limiter la taille de l'espace de recherche.

5. http ://www.statmt.org/moses/
6. http ://www.statmt.org/wmt11/baseline.html

Mot de la phrase source	*Alternatives de traduction*	*Probabilité*
J’	I	0.8
	Me	0.2
ai	have	0.5
	had	0.3
	was	0.1
	were	0.1
vendu	sold	0.4
	traded	0.4
	betrayed	0.2
mon	my	0.8
	me	0.2
bateau	boat	0.4
	ship	0.4
	canoe	0.2
de	of	0.4
	from	0.3
	since	0.3
pêche	fishing	0.333
	fish	0.333
	peach	0.333

TABLE 1.1 – Exemples d’alternatives de traduction.

Plutôt qu’expliquer la représentation par piles de l’espace de recherche, à mon sens assez peu intuitive, je vais en présenter une strictement équivalente d’un point de vue formel, mais probablement plus familière aux membres de la communauté : le treillis de mots. En outre, cette explication sera nécessaire pour la compréhension des travaux exposés dans cette thèse, car je présenterai dans la partie III une méthode de traduction automatique de la parole reposant sur l’utilisation d’un *réseau de confusion*. Or, ce réseau peut être vu comme une simplification du treillis de mots.

Le treillis de mots est un graphe orienté acyclique dans lequel chaque arête représente un mot. Une phrase est un chemin entre le nœud le plus à gauche (début de la phrase, vide : nous le noterons `<s>`) et le nœud le plus à droite (fin de la phrase : il sera noté `</s>`). Un chemin partiel (c’est-à-dire un chemin du nœud `<s>` à un nœud qui n’est pas `</s>`) représente une hypothèse de traduction partielle. On peut calculer le coût d’un chemin partiel ou total σ grâce aux modèles d’alignement, de langue et de traduction en parcourant le chemin depuis le nœud `<s>` et en calculant le coût de chaque arête. Le coût d’une arête est $C(e) = -log(P(s_{a_i}|t_i) \times P(a_i|\mathbf{a}_1^{i-1}) \times P(t_i|\mathbf{t}_1^{i-1}))$, avec t_i le mot porté par l’arête e, qui porte également a_i. Le coût du chemin σ est la somme des coûts de ses arêtes, donc $P(\sigma) = exp(-C(\sigma)) = P(\mathbf{s}|\mathbf{t}_1^i) \times P(\mathbf{a}_1^i) \times P(\mathbf{t}_1^i) = P(\mathbf{t}_1^i|\mathbf{s})$.

Jusqu’ici, ce graphe est en réalité un arbre, si on ignore le nœud final `</s>`. Mais afin de réduire la complexité de la recherche de la meilleure hypothèse, on va réduire la taille du graphe en faisant se rejoindre certains chemins. Deux chemins partiels se rejoignent quand ils sont les traductions partielles de la même partie de la phrase source, et que les $n-1$ derniers mots de la traduction partielle sont identiques, avec n l’ordre du modèle de langue n-gramme utilisé pour le décodage. Sous ces conditions, on a en effet assez d’information pour distinguer les deux hypothèses et déterminer la meilleure des deux. Notons $\bar{\mathbf{t}}$ et $\bar{\mathbf{u}}$ les deux hypothèses partielles, et $\bar{\mathbf{s}}$ la partie de la phrase source couverte par ces traductions partielles (on a fait l’hypothèse qu’elle était la même pour les deux). Notons $\mathbf{v}$ leurs derniers mots communs. Quelle que soit la

suite $\mathbf{w} = w_1, .., w_k$ de la traduction, puisque $Len(\mathbf{v}) \geqslant n-1$, on a, pour le modèle de langue :

$$\forall i \in \{1, .., k\}, P(w_i|\bar{\mathbf{t}}\mathbf{w}_1^{i-1}) = P(w_i|\mathbf{v}\mathbf{w}_1^{i-1}) = P(w_i|\bar{\mathbf{u}}\mathbf{w}_1^{i-1})$$

En outre on connaît déjà à ce stade $P(\bar{\mathbf{s}}|\bar{\mathbf{t}})$ et $P(\bar{\mathbf{s}}|\bar{\mathbf{u}})$. On a donc toute l'information nécessaire pour déterminer quelle hypothèse est la meilleure, quelle que soit la façon dont on prolonge le chemin jusqu'à `</s>`, et il est inutile de continuer à les distinguer. On peut donc les faire converger. En pratique, la moins bonne des deux hypothèses partielles sera supprimée (on parle *d'élagage* ou en anglais *pruning*), en général immédiatement, par le décodeur. Ainsi, on minimise la taille du graphe en conservant toute l'information nécessaire à la détermination de la meilleure hypothèse. La figure 1.7 donne un aperçu de ce que serait un tel treillis de mots pour notre exemple, sans élagage, et sans considérer les réarrangements :

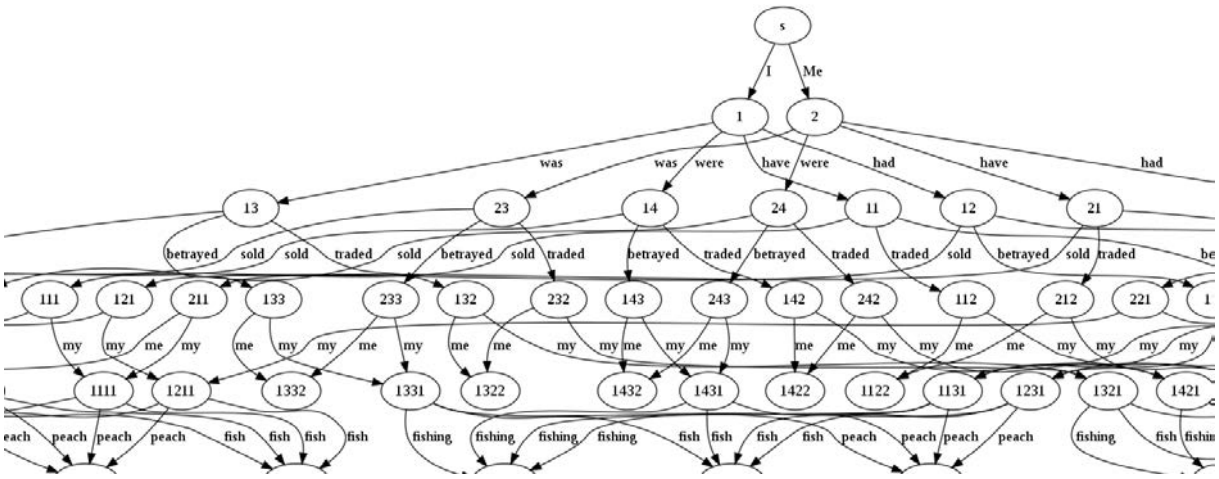


FIGURE 1.7 – Extrait d'un treillis non élagué

On voit que la taille du treillis, sur un exemple simpliste, reste très importante. Il va donc falloir avoir recourt à des stratégies de simplification pour que le problème reste abordable en terme de temps de calcul :

- *Élimination des hypothèses les moins bonnes.* Si plusieurs hypothèses partielles convergent, on ne conserve que la meilleure (le graphe devient un arbre).
- *Élagage.* Une hypothèse partielle dont la probabilité est inférieure à un seuil choisi par l'utilisateur n'est pas conservée.
- *Contrôle de la largeur.* On fixe une limite M au nombre de traductions partielles couvrant K mots qui sont représentées dans le graphe. Dit autrement, pour tout K, il existe au plus M chemins d'origine `<s>` et de longueur K dans le graphe (chaque pile a une capacité M).
- *Limitation du nombre d'alternatives de traduction.* On ne considère que les N meilleures alternatives de traduction pour chaque élément de la phrase source. Dit autrement, pour un alignement donné, chaque nœud a au plus N fils. On peut également ne conserver que les alternatives ayant une probabilité suffisante.

En appliquant ces stratégies, on peut se ramener à un treillis beaucoup plus petit. Mais il s'agit là d'approximations, et on n'est pas assuré de ne pas éliminer la meilleure hypothèse. Le treillis est donc construit de la façon suivante (algorithme très simplifié, ne prenant notamment pas en compte la fertilité, les séquences, etc.) :

1. On commence avec le seul nœud `<s>` (toutes les piles sont vides).

2. Pour toutes les possibilités d'alignement a_1 (alignement du premier mot de $\mathbf{t}$) données par le modèle d'alignement et le modèle de longueur $P(Len(\mathbf{S})|Len(\mathbf{T}))$, on cherche dans la table de traduction les alternatives $\mathcal{T}(s_{a_1})$ correspondant à s_{a_1}. Pour chaque valeur de a_1 on considère au plus N entrées. Chacune de ces alternatives constitue une nouvelle traduction partielle. On créé donc autant de fils du nœud `<s>`, les arcs portant tous les éléments des $\mathcal{T}(s_{a_1})$ qui n'ont pas été éliminés (c'est-à-dire que la pile p_1 contient $\cup_{a_1}\mathcal{T}(s_{a_1})$). Ensuite, seules les M meilleures traductions partielles sont conservées.
3. Le processus se poursuit récursivement : Chaque nœud n'ayant pas de fils est prolongé en créant, pour chaque position non encore couverte par les hypothèses arrivant à ce nœud, des fils correspondant aux alternatives de traductions. Le nombre de fils est contraint par les paramètres M et N et les probabilités d'alignement, du modèle de langage et du modèle de traduction. Si deux hypothèses convergent selon le critère évoqué précédemment, la moins bonne est éliminée.
4. Si un chemin ne peut être prolongé, il est considéré comme une impasse.
5. Quand un chemin couvre tous les mots de la phrase source, il est prolongé vers le nœud spécial `</s>` (fin de phrase).

Supposons qu'on fixe $M = 4$ et $N = 2$ et qu'on utilise un modèle trigramme (donc un historique de longueur deux) pour construire le treillis. La figure 1.8 montre ce que le treillis précédent pourrait devenir, avant élagage des hypothèses convergentes, et toujours en ne considérant que les alignements monotones (c'est-à-dire sans réarrangement) :

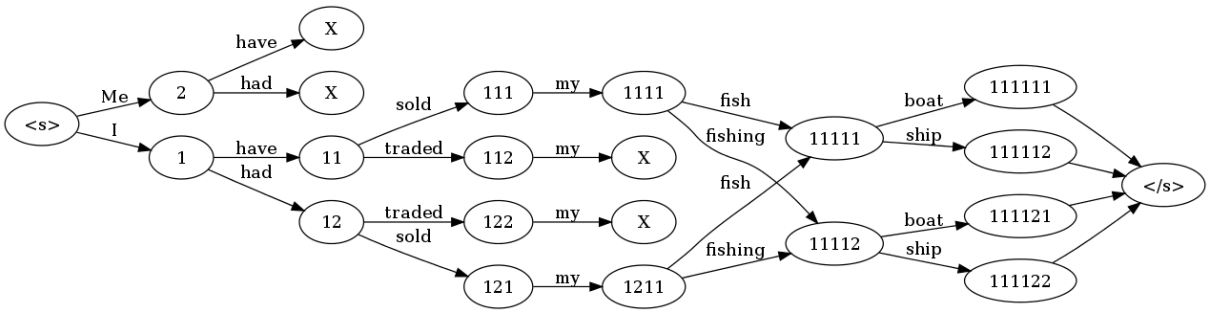


FIGURE 1.8 – Exemple de treillis élagué

La réduction de complexité est ici spectaculaire. Bien sûr, cela tient en partie au fait que cet exemple est simpliste. Un treillis réel compte aisément des dizaines ou des centaines de milliers de nœuds et d'arêtes. Une fois le treillis construit, la meilleure hypothèse de traduction est déterminée en cherchant le chemin de moindre coût entre `<s>` et `</s>` grâce à un algorithme de recherche de plus court chemin, par exemple l'algorithme A^* [Hart 68]. Pour certaines applications, on calculera non pas la meilleure hypothèse mais les n meilleures : l'ensemble obtenu est une *n-best list* (« liste des n meilleures »). Cela peut servir, par exemple, si on veut utiliser un système différent pour déterminer laquelle de ces hypothèses est vraiment la meilleure traduction [Shen 04]. Ou pour recombiner ces hypothèses pour en générer une nouvelle. Ou encore pour calculer des mesures de confiance (section 2.1.1).

1.5.6 Les corpora

À l'instar de n'importe quel problème d'apprentissage automatique, les paramètres des modèles de langue et de traduction sont estimés sur de grands corpora de données. La taille des

corpora doit être d'autant plus importante que le nombre de paramètres est élevé. Les modèles de langue sont estimés sur des corpora dits *monolingues* (un ensemble de texte dans la langue concernée), et les modèles d'alignement et de traduction sont estimés sur des corpora *bilingues alignés* : ceux-ci sont des collections de paires de phrases dont l'une est la traduction de l'autre. Ces corpora comptent quelques centaines de milliers à quelques centaines de millions de phrases. Plus les documents qui seront traduits par un système sont similaires aux corpora qui ont été utilisés pour l'estimation des modèles, meilleures seront les performances. Il existe des dizaines des corpora différents, chacun étant en général issu d'un besoin particulier. On peut citer :

Europarl [Koeh 05] : il comprend des corpora monolingues et bilingues, constitués de transcriptions de sessions du Parlement Européen et de leurs traductions dans différentes langues de l'Union Européenne (français, anglais, allemand, espagnol, tchèque, hongrois et d'autres). Il a été constitué dans le cadre du projet EuroMatrix [7] dont le but était de pourvoir les instances administratives et politiques de l'Union Européenne en outils de traduction automatique. Pour illustration, le corpus monolingue anglais de la version 6 d'Europarl compte environ deux millions de phrases (cinquante millions de mots), et le corpus bilingue français-anglais 1,8 millions de paires de phrases (45 millions de mots environ). Ce corpus est utilisé notamment pour les campagnes d'évaluations WMT [8]. Nous utiliserons ces corpora pour toutes nos expériences sur les mesures de confiance (partie II). En plus des transcriptions des sessions du Parlement Européen, des transcriptions d'émissions radiophoniques et leurs traductions ont été fournies lors de la campagne d'évaluation WMT 2011. Le corpus monolingue des transcriptions en anglais comprend environ trente millions de phrases (660 millions de mots), et le corpus bilingue 115 mille paires de phrases (2,9 millions de mots). Ces corpora seront utilisés dans la partie III pour l'entraînement de modèles de langue et de traduction de la parole.

TED : utilisé notamment lors de la campagne IWSLT 2011 [9], ces corpora monolingues et multilingues sont constitués des transcriptions des célèbres *TED Talks* [10]. Le corpus bilingue français-anglais comprend environ cent mille paires de phrases (1,7 millions de mots), et le corpus monolingue anglais environ 125 mille phrases (environ 2,4 millions de mots), ce qui en fait un corpus assez modeste. Nous l'utiliserons dans la partie III pour l'entraînement de modèles de langue et de traduction de la parole.

MultiUN [Eise 10] : un corpus multilingue tiré du site internet des Nations Unies [11]. La partie bilingue français-anglais contient environ douze millions de paires de phrases (346 millions de mots).

Google n-gram : ce corpus [12] consiste en une liste des n-grammes présents dans les livres numérisés dans le cadre du projet *Google Books* [13], et leur nombre d'occurrences. Ce corpus est donc intermédiaire entre un corpus textuel et un modèle de langue.

7. http ://www.euromatrix.net/ — http ://www.euromatrixplus.net/
8. http ://www.statmt.org/
9. http ://iwslt2011.org
10. http ://www.ted.com/
11. http ://ods.un.org/
12. http ://books.google.com/ngrams/datasets
13. http ://books.google.com/

GigaWord : ce corpus proposé par le *Linguistic Data Consortium*[14] est un imposant corpus de dépêches proposées par l'*Agence France Press English Service*, l'*Associated Press Worldstream English Service*, le *New York Times Newswire Service* et le *Xinhua News Agency English Service*, comprenant environ 1,8 milliards de mots.

Il en existe d'autres, et notamment des corpora plus petits destinés à des usages spécialisés (traduction de phrases utiles en voyage, documents médicaux, etc.).

1.6 L'évaluation des traductions automatiques

Évidemment, même si un système est capable de déterminer la traduction la plus probable de n'importe quelle phrase source, la qualité de la traduction proposée dépendra de la pertinence des modèles utilisés et de la précision de l'estimation des paramètres. Une traduction optimale au sens probabiliste, mais obtenue avec un modèle simpliste ou entraîné sur trop peu de données ne sera pas de qualité satisfaisante. Et les modèles les plus sophistiqués sont encore incapables de rendre compte de toute la complexité des langues naturelles et de leur évolution constante. Il est donc important, pour juger des progrès accomplis ou des améliorations à apporter, ou pour comparer entre eux les systèmes et les modèles, d'évaluer rigoureusement la qualité des traductions. Bien sûr, la notion même de qualité est difficile à définir dans un domaine aussi subjectif que la traduction (voir la discussion en partie II, section 1.1).

1.6.1 Évaluation humaine

L'idée la plus naturelle est de demander à un expert ou un ensemble d'experts d'évaluer la qualité des traductions. On fait souvent cette évaluation phrase par phrase selon deux critères [Koeh 06] :

- *Intelligibilité :* la traduction proposée est-elle compréhensible ? Est-elle une phrase bien formée ? En principe, l'intelligibilité s'évalue indépendamment de la phrase source.
- *Adéquation :* la traduction a-t-elle le même sens que la phrase source ? Ce critère est en principe évalué indépendamment de l'intelligibilité, même si les deux sont évidemment corrélées.

On distingue parfois aussi le *caractère informatif* de la traduction, en mesurant la capacité d'un juge humain à répondre à des questions portant sur le contenu de la phrase ou du document après avoir lu sa traduction. Pour certaines applications, on peut au contraire ne demander qu'une seule évaluation, synthétique. Parfois encore, on demande simplement aux juges de classer les traductions entre elles.

Si ces méthodes ont l'avantage de faire intervenir le bénéficiaire final de la traduction automatique, à savoir l'être humain, elles ont tout de même de sérieux désavantages : d'abord, elles sont extrêmement coûteuses. 391 heures de travail ont été nécessaires à 130 juges pour évaluer les résultats de la campagne WMT 2011 [Call 11] (quelques milliers de phrases au total). Ensuite, ces évaluations sont largement subjectives et ne sont pas reproductibles : les auteurs rapportent un accord inter-annotateurs assez modeste (l'interprétation de la métrique d'accord étant assez subtile, je ne la reprendrai pas ici afin d'éviter une trop grande simplification, et me contenterai de l'appréciation). Même l'accord intra-annotateur (c'est-à-dire la constance des jugements portés sur une même traduction à différents moments) n'est pas meilleure que « substantielle » selon les auteurs.

14. ressource LDC2003T05

Il peut donc être délicat d'utiliser ces évaluations pour comparer des systèmes ou pour juger de l'évolution des performances. Afin de résoudre ces problèmes, des techniques d'évaluation automatique ont été développées. Nous allons voir qu'en contrepartie d'un coût faible et de la reproductibilité de ces métriques, elles ne constituent toujours pas une panacée. La prudence reste donc de mise dans l'interprétation des évaluations.

1.6.2 Évaluation automatique

Les mesures d'évaluation automatique reposent sur la comparaison entre une traduction automatique et une traduction de référence, produite par un expert. De façon très simplifiée, toute différence entre la traduction automatique et la référence est considérée comme une erreur, et ces mesures comptent le nombre d'erreurs. C'est la façon de comparer les traductions automatiques et de référence qui diffèrent. Afin de ne pas pénaliser des traductions qui seraient correctes, mais différentes de la traduction de référence (il existe en général plusieurs façons correctes de traduire une même phrase source), certaines de ces mesures (notamment BLEU [Papi 01]) peuvent utiliser plusieurs traductions de référence.

Taux d'Erreur Mot : le *Taux d'Erreur Mot* ou *WER* (*Word Error Rate*) est simplement la distance de Levenshtein [Leve 66], d_L dans la formule ci-dessous, rapportée à la longueur de la référence :

$$WER(\mathbf{t}, \mathbf{t}_{ref}) = \frac{d_L(\mathbf{t}, \mathbf{t}_{ref})}{Len(\mathbf{t}_{ref})}$$

C'est la métrique utilisée en reconnaissance automatique de la parole. Si elle est bien adaptée pour ce problème car il a solution unique et bien définie, il n'en va pas de même pour la traduction : par exemple, un adjectif manquant rendra en général la traduction seulement imprécise, alors qu'une négation manquante peut en changer totalement le sens. De plus, cette mesure pénalise les différences dans l'ordre des mots, alors qu'il n'est pas toujours important. Les auteurs de [Till 97b] proposent une variante indépendante de la position des mots (ce qui évidemment pose problème dans certains cas), le *Taux d'Erreur Mot indépendant de la position* : si on considère l'hypothèse de traduction et la référence comme deux multi-ensembles (ensembles dans lesquels un élément peut apparaître plusieurs fois) et non comme des séquences, le *PER* (*Position-independant Error Rate*) est simplement le cardinal de leur différence symétrique rapporté au cardinal de la référence :

$$PER(\mathbf{t}, \mathbf{t}_{ref}) = \frac{|\mathbf{t} \Delta \mathbf{t}_{ref}|}{|\mathbf{t}_{ref}|}$$

Taux d'Édition de la Traduction : plus sophistiqué, le *Translation Edit Rate* (*TER*) [Snov 06] mesure le nombre minimal d'éditions élémentaires qu'un expert doit apporter à la traduction automatique pour qu'elle soit identique à une des traductions de référence. Les opérations élémentaires sont l'*insertion*, la *substitution* et la *suppression* (ces opérations sont également comptées par le WER), ainsi que le *déplacement d'un groupe de mots*. Notons $E(\mathbf{t}, \mathbf{t}_{ref})$ le nombre minimal d'opérations pour transformer $\mathbf{t}$ en $\mathbf{t}_{ref}$, déterminé par programmation dynamique ; notons $\mathbf{t}^1_{ref}, .., \mathbf{t}^K_{ref}$ l'ensemble des traductions de référence. Le TER s'exprime :

$$TER(\mathbf{t}; \mathbf{t}^1_{ref}, .., \mathbf{t}^K_{ref}) = \frac{min\{E(\mathbf{t}, \mathbf{t}^k_{ref})\}_{k=1..K}}{\frac{1}{K} \sum_{i=1..K} Len(\mathbf{t}^k_{ref})}$$

TER, et ses variantes HTER et TER plus [Snov 09] faisant intervenir des experts ou des connaissances expertes, figurent parmi les meilleures mesures actuellement disponibles pour l'évaluation automatique ou semi-automatique des traductions.

BLEU : je terminerai avec BLEU (BiLingual Evaluation Understudy, [Papi 01]), peut-être la métrique la plus populaire, utilisée notamment dans les campagnes WMT. C'est également celle que nous utiliserons dans nos expériences. Le principe de cette métrique est de comparer non seulement les mots composant les traductions, mais aussi les n-grammes. Plus précisément, BLEU est la moyenne géométrique pondérée des précisions n-grammes (c'est-à-dire, le nombre de n-grammes de l'hypothèse apparaissant effectivement dans une référence rapporté au nombre de n-grammes contenus dans l'hypothèse. Les n-grammes « manquants », c'est-à-dire apparaissant dans une référence et pas dans l'hypothèse, ne sont pas comptés), pour n compris entre 1 (les unigrammes, c'est-à-dire les mots) et N ($N = 4$ dans l'implémentation de référence) :

$$BLEU(\mathbf{t}; \mathbf{t}^1_{ref}, .., \mathbf{t}^K_{ref}) = BP(\mathbf{t}; \mathbf{t}^1_{ref}, .., \mathbf{t}^K_{ref}) \times e^{\sum_{n=1}^{N} w_n log(p_n)}$$

p_n est la précision n-gramme, w_n est son poids dans la moyenne, et BP (*Brevity Penalty*) est une pénalité aux traductions courtes (en effet, les n-grammes manquants ne sont pas pénalisés par la précision, les traductions tronquées sont donc artificiellement avantagées) :

$$BP(\mathbf{t}; \mathbf{t}^1_{ref}, .., \mathbf{t}^K_{ref}) = \begin{cases} 1 & \text{si } Len(\mathbf{t}) > r \\ e^{1-\frac{r}{Len(\mathbf{t})}} & \end{cases}$$

où r est la longueur de la traduction de référence la plus proche de $Len(\mathbf{t})$ parmi $Len(\mathbf{t}^1_{ref}), .., Len(\mathbf{t}^K_{ref})$.

BLEU a la réputation d'être une des meilleures métriques d'évaluation automatique des traductions. Notamment, lors de la campagne d'évaluation NIST MT, le classement des systèmes en compétition réalisé par BLEU a été le même que celui réalisé par des juges humains. Cependant, il faut noter BLEU excelle lorsque les précisions n-grammes sont calculées au niveau d'un *document* (un ensemble de phrases, un texte, une page web...), mais que l'utilisation de cette métrique pour évaluer les phrases séparément est sujette à caution. Les auteurs de [Tant 07] en proposent une variante permettant de faire la différence entre différents types d'erreurs (erreurs de vocabulaire, erreurs de flexion, etc.).

Cette présentation n'est pas exhaustive. On pourrait encore citer NIST [Dodd 02] et WNM [Baby 04], des améliorations de BLEU, ou METEOR (*Metric for Evaluation of Translation with Explicit Ordering*, [Bane 05]), qui utilise un modèle d'alignement entre une traduction automatique et la traduction de référence.

2

Les mesures de confiance

Quelle que soit la qualité des modèles implémentés, la taille des corpora sur lesquels leurs paramètres sont estimés, ou l'efficacité des décodeurs employés, les traductions proposées par un système de traduction automatique sont toujours loin d'être parfaites. Fondamentalement, un système de traduction génère l'hypothèse la plus probable selon un modèle étant donnée une phrase source ; or, aucun modèle ne rend compte de la variabilité, de la subtilité et de la complexité du langage naturel. Les meilleurs systèmes de traduction automatique font donc, et continueront de faire des erreurs. Après tout, même les humains en font. Les erreurs peuvent prendre des formes diverses. Un mot peut être incorrect, mal placé ou manquant. Une traduction peut n'avoir absolument aucun sens ou être légèrement moins précise que la phrase source. Un mot manquant (une négation, par exemple) peut bouleverser le sens, alors qu'une myriade de petites erreurs grammaticales, si elles rendent la lecture plus difficile, ne gênent pas forcément la compréhension. Le problème est donc complexe, mais les utilisateurs, et notamment les utilisateurs ne maîtrisant pas la langue source, pourraient tirer parti d'un outil de détection d'erreur dans les traductions. D'autres applications sont possibles : au sein des entreprises ou institutions déployant des systèmes de traduction, un document traduit automatiquement et destiné à la diffusion doit être relu et corrigé par un expert. On parle alors de post-édition, et un système permettant de détecter rapidement les erreurs pourrait bénéficier aux traducteurs [Ueff 05]. La post-édition automatique, telle qu'elle est par exemple décrite dans [Sima 07], pourrait aussi tirer parti de la détection d'erreur, en sélectionnant les phrases qui doivent être modifiées. Autre exemple, les auteurs de [Spec 11], par exemple, proposent de sélectionner parmi les résultats d'une traduction automatique les phrases assez bonnes pour être corrigées en post-édition, les autres étant retraduites intégralement à partir de la source par l'interprète. Cela évite à l'interprète la frustration de devoir comprendre et corriger des traductions particulièrement mauvaises. Les auteurs de [Gand 03] proposent également d'utiliser les mesures de confiance dans un système de prédiction de traduction. Ces applications potentielles seront décrites dans la section 2.2 de cette introduction. Dans cette thèse, je proposerai une application des mesures de confiance à la post-édition (section 4) et à la traduction de la parole (partie III).

Cependant, et peut-être à cause de telles attentes, l'estimation de qualité est un problème particulièrement délicat : en effet, si un utilisateur doit prendre une décision en se basant sur une mesure de confiance (comme accepter, éditer ou rejeter une proposition de traduction, s'y fier ou non, etc.), celle-ci doit être particulièrement fiable, afin de ne pas lui faire perdre de temps. Par exemple, si une mesure de confiance laisse passer trop d'erreurs, l'utilisateur devra de toutes façons vérifier toutes les traductions et il ne gagnera pas de temps (voir l'expérience de post-édition — section 4). Même pour une tâche d'assimilation (c'est-à-dire quand on ne

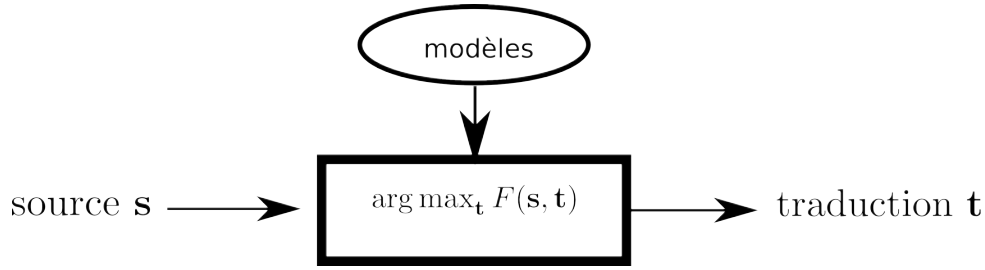


FIGURE 2.1 – Schéma très général d'un système de traduction automatique

cherche pas à obtenir une bonne traduction, mais seulement à comprendre le texte original), il est particulièrement important de ne pas induire l'utilisateur en erreur en lui laissant croire qu'une traduction est correcte alors que son sens est erroné.

2.1 Différentes approches des mesures de confiance

L'intérêt que la communauté scientifique de traduction automatique porte aux mesures de confiance est relativement récent. Une grande partie des articles traitant du sujet a été publiée après le début des années 2000. C'est donc encore un domaine jeune, et il est difficile d'y discerner une approche particulière qui constituerait sans conteste une référence, un point de comparaison stable. Dans cette section je vais donc m'attacher à présenter méthodiquement, mais sans prétendre à une exhaustivité illusoire, trois axes qui me semblent les plus représentatifs des différentes approches des mesures de confiance. Cette section n'est pas destinée à formaliser le problème, on se référera pour ceci à la partie II, section 1. Lorsque cela est nécessaire et ne nuit pas à la clarté, j'utiliserai ici quelques notations qui seront définies au début de la partie II, section 1.

2.1.1 Probabilité *a posteriori* des traductions

Très schématiquement, la plupart des systèmes de traduction automatique (section 1), et notamment les systèmes statistiques (section 1.5), fonctionnent en attribuant un score $F(\mathbf{s},\mathbf{t})$ à toute hypothèse de traduction $\mathbf{t}$ étant donnée une phrase à traduire $\mathbf{s}$, et en choisissant la traduction obtenant le meilleur score (figure 2.1). Ce score peut être basé par exemple sur des probabilités dans le cas d'un décodeur statistique, ou sur les similarités entre la traduction proposée et des traductions connues pour les systèmes par analogie.

Une première approche intuitive des mesures de confiance est de considérer que plus la traduction retenue a un score bas, plus sa qualité est basse, ou plus elle a de chance d'être erronée. C'est-à-dire :

$$F(\mathbf{s},\mathbf{t}) < F(\mathbf{s},\mathbf{t}') \Leftrightarrow \mathbf{t} \text{ est une moins bonne traduction de } \mathbf{s} \text{ que } \mathbf{t}'$$

Dans le cas des systèmes statistiques, on peut utiliser le score $P(\mathbf{s}|\mathbf{t}) \times P(\mathbf{t})$ fourni par le décodeur (section 1.5.5). Cependant, ce score n'est pas *normalisé*, car le facteur $P(\mathbf{s})$ est ignoré (il est constant donc inutile au calcul de la meilleure traduction — section 1.5.1). Il peut donc exister deux couples de phrases $(\mathbf{s}^1,\mathbf{t}^1)$ et $(\mathbf{s}^2,\mathbf{t}^2)$ tels que :

$$F(\mathbf{s}^1,\mathbf{t}^1) < F(\mathbf{s}^2,\mathbf{t}^2) \text{ et } P(\mathbf{t}^1|\mathbf{s}^1) > P(\mathbf{t}^2|\mathbf{s}^2)$$

Ce qui signifie que la traduction $\mathbf{t}^2$ obtient un meilleur score selon le système ($F(\mathbf{s}^2, \mathbf{t}^2)$) alors que $\mathbf{t}^1$ a plus de chances d'être une traduction correcte ($P(\mathbf{t}^1|\mathbf{s}^1) > P(\mathbf{t}^2|\mathbf{s}^2)$). On voit donc qu'on ne peut pas interpréter les scores fournis par le système directement comme une estimation de qualité.

Pour estimer correctement la probabilité qu'une traduction soit correcte, il faut donc calculer sa *probabilité a posteriori* $P(\mathbf{t}|\mathbf{s})$. Dans ce contexte, la mesure de confiance d'un mot sera :

$$P(t_i = w|\mathbf{s}) = \sum_{\tilde{\mathbf{t}} \in \mathcal{V}_T^* \setminus \tilde{t}_i = w} P(\tilde{\mathbf{t}}|\mathbf{s})$$

On peut calculer la probabilité *a posteriori* à partir des scores fournis par le système par la relation suivante, présentée dans la section 1.5.1 :

$$P(\mathbf{t}|\mathbf{s}) = \frac{P(\mathbf{s}|\mathbf{t}) \times P(\mathbf{t})}{P(\mathbf{s})}$$

Il faut donc estimer $P(\mathbf{s})$. Les approches les plus courantes consistent à utiliser pour cela les *n-best list* (liste des n meilleures hypothèses) et les treillis de mots fournis par les systèmes de traduction statistiques [Ueff 03, Blat 03] (voir la partie I, section 1.5.5 pour des explications sur les treillis et les n-best list). Cette approche est également utilisée pour l'estimation de confiance en transcription automatique [Razi 08, Razi 07, Kemp 97], les deux problèmes étant formellement identiques.

Les auteurs de [Ueff 05], par exemple, proposent une méthode pour déterminer la probabilité *a posteriori* d'un mot $P(t_i = w|\mathbf{s})$ en sommant les scores des chemins contenant ce mot. La normalisation est effectuée en divisant cette somme par la somme des probabilités de tous les chemins du graphe. Une méthode alternative pour calculer $P(t_i = w|\mathbf{s})$ est de sommer les probabilités de toutes les hypothèses $\mathbf{t}$ de la *n-best list* telles que $t_i = w$, et de normaliser en divisant par la somme des probabilités de toutes les hypothèses. Cette alternative est plus simple, mais la *n-best list* représente en général moins d'hypothèses que le graphe de mots. $P(\mathbf{s})$ est donc moins bien estimée.

Cette approche a l'avantage de ne nécessiter aucun modèle supplémentaire pour l'estimation de la mesure de confiance. Cette approche permet également de détecter les phrases hors domaine : par exemple, si des modèles statistiques ont été entraînés sur des corpora du Parlement Européen, et qu'on les utilise pour traduire des articles scientifiques, il y a fort à parier que les traductions seront de mauvaises qualité. Dans ce cas, les scores calculés par le système seront bas.

En revanche, cette approche nécessite l'accès aux scores fournis par le système de traduction, ce qui n'est pas toujours le cas lorsqu'on utilise des systèmes commerciaux. Cette méthode n'est pas non plus robuste aux modèles mal estimés (lorsqu'on ne dispose que de peu de données par exemple). Supposons par exemple que le modèle de traduction soit entraîné sur un corpus bilingue dans lequel il est question d'aviation, de telle sorte que le mot français *vol* est toujours traduit par le mot anglais *flight*. Dans ce cas, un modèle statistique ou par analogie prédirait toujours, avec un bon score, la traduction *flight* pour le mot *vol*. On voit donc que si on traduit une phrase dans lequel il a un sens différent, par exemple : *Le vol des plans du prototype*, on a toutes les chances d'obtenir une erreur de traduction et de ne pas détecter l'erreur.

Enfin, elle n'est pas robuste aux changements dans les systèmes : si on change les algorithmes, les paramètres ou les modèles utilisés (par exemple, si on passe d'un modèle généraliste à un modèle spécifique), le score d'une traduction donnée $\mathbf{t}$ pour une phrase source donnée peut être modifié, alors que naturellement, sa qualité ne change pas.

On voit donc les limites inhérentes à cette approche : d'une part, fondamentalement, la qualité de la traduction $\mathbf{t}$ d'une phrase source $\mathbf{s}$ dépend **uniquement** du couple $(\mathbf{s}, \mathbf{t})$ et de son contexte, pas du système utilisé pour la traduction. D'autre part, si les modèles ont échoué à fournir une traduction correcte, il y a peu de chances qu'ils puissent constituer un outil efficace pour détecter leurs propres erreurs.

2.1.2 Estimation directe

Afin d'éviter de devoir utiliser pour la détection des erreurs les modèles qui ont engendré ces mêmes erreurs, on peut développer un système d'estimation de qualité spécifique, distinct du système de traduction, utilisant ses propres modèles. Cela a en outre l'avantage de rendre inutile l'accès au fonctionnement interne du système de traduction, et de rendre l'estimation de qualité totalement indépendante de celui-ci. Formellement, cela revient à estimer un score $f(\mathbf{s}, \mathbf{t})$ avec $f \neq F$ (figure 2.1). Par exemple, f peut être la probabilité estimée selon un modèle de traduction estimé séparément, ou un modèle de langue, ou encore des modèles spécifiques comme décrits dans la section 3.

Les auteurs de [Ueff 05], par exemple, proposent d'utiliser un modèle IBM-1 pour estimer la probabilité de correction d'un mot. La probabilité du mot t_i dans ce modèle s'écrit en principe :

$$P(t_j|\mathbf{s}) = \frac{1}{Len(\mathbf{s})} \sum_{i=0..Len(\mathbf{s})} p_{IBM-1}(t_j|s_i)$$

Sous cette forme, elle est utilisée dans [Blat 03]. Les auteurs de [Ueff 05], constatant qu'une bonne approximation de cette somme est le maximum, proposent de la remplacer par :

$$\tilde{P}(t_j|\mathbf{s}) = max\{p_{IBM-1}(t_j|s_i)\}_{i=0..Len(\mathbf{s})}$$

Ils montrent cependant que cette méthode, si elle est plus simple, donne de moins bons résultats que celle exposée dans la section précédente.

Cette méthode a le gros avantage d'apporter à l'estimation de qualité des informations nouvelles, qui n'ont pas été utilisées pour la traduction elle-même. On peut ainsi espérer plus efficacement compenser les défaillances des modèles utilisés pour la traduction. Les scores ainsi produits seront en outre aisément interprétables : il s'agit de juger de la qualité d'une traduction à l'aune d'un certain critère ou d'un certain modèle. Cependant des limitations persistent : d'une part, les modèles utilisés (modèles de traduction IBM, modèles de langage, etc.) sont souvent très corrélés, voire identiques, aux modèles utilisés pour la traduction. On a donc le même problème qu'avec la méthode exposée précédemment : ces modèles risquent de commettre les mêmes erreurs que ceux utilisés pour la traduction, et d'être donc impuissants à détecter les erreurs. D'autre part, on peut se dire que si on avait un modèle effectivement capable de distinguer un mot erroné d'un mot correct, on pourrait l'utiliser dans le système de traduction lui-même, pour éviter les erreurs. Enfin, un score, c'est-à-dire une seule source d'information, s'avère dans bien des cas une représentation trop fruste pour détecter une erreur étant donnée la complexité des langues naturelles.

2.1.3 Classification et Régression multivariées

Nous en arrivons donc à la dernière approche, qui sera celle développée dans cette thèse, et qui est déjà adoptée dans de nombreuses équipes de recherche. Si une source d'information n'est pas une représentation assez riche pour représenter la traduction dans toute sa complexité, nous

allons multiplier les sources d'information. Plus de détails seront donnés dans la partie II, section 1, mais le principe est de représenter une paire $(\mathbf{s}, \mathbf{t})$ d'une phrase source et d'une traduction automatique par un vecteur de $\mathbb{R}^d$, avec $d \in \mathbb{N}$ le nombre de « descripteurs » ou « paramètres prédictifs » utilisés (partie II, section 3), et de voir le problème d'estimation de qualité comme un problème d'apprentissage statistique, qui sera traité par des techniques classiques de classification et régression multivariées (partie II, section 1.4). Si cette technique présente les inconvénients de temps de calculs souvent élevés, et de scores difficilement interprétables (pris individuellement, les paramètres prédictifs n'ont pas d'interprétation univoque en terme d'estimation de qualité, et les algorithmes de classification et de régression sont utilisés comme des boîtes noires), on peut en revanche aisément ajouter de nouvelles sources d'informations, et même utiliser celles présentées dans les sections précédentes 2.1.1 et 2.1.2.

L'objectif de cette section n'est pas de formaliser le problème, ni de détailler les paramètres prédictifs et les algorithmes d'apprentissage automatique utilisés. Ce sera fait en partie II, section 3. Je vais ici faire un tour d'horizon des différentes approches rencontrées dans la littérature sur le sujet. Je ne prétends pas à l'exhaustivité : l'intérêt majeur de l'estimation de confiance par classification ou régression multivariée est justement de pouvoir s'accommoder d'une très grande diversité de paramètres prédictifs, et il existe des dizaines d'algorithmes d'apprentissage automatique différents pouvant être utilisés pour résoudre le problème, chacun avec leurs spécificités. Les passer tous en revue serait laborieux, sinon impossible. Je me contenterai donc de présenter les approches à ma connaissance les plus efficaces et les plus représentatives.

Cette approche a très tôt été explorée par la communauté de traduction automatique lorsque l'intérêt pour les mesures de confiance s'est développé au début des années 2000. Un bon point de départ (il fut le mien) pour la compréhension de l'estimation de confiance par classification ou régression multivariée est le compte-rendu de l'atelier de la *Johns Hopkins University* sur les mesures de confiance [Blat 03]. Cet article donne un bon aperçu de la diversité des paramètres prédictifs utilisés dans ce domaine. Les participants à cet atelier ont utilisé des modèles de langue n-gramme (partie II, sections 3.1.1 et 3.2.1), des modèles de traduction IBM-1 (section 1.5.2 ; partie II, sections 3.1.6 et 3.2.3), des estimations de probabilité *a posteriori* (section 2.1.1), et pléthore d'autres paramètres (utilisant par exemple l'alignement, ou une analyse superficielle des hypothèses de traduction). La classification est ensuite effectuée en utilisant un classifieur bayésien naïf ou un perceptron multicouches (partie II, section 1.4.3). Les résultats présentés sont cependant assez délicats à interpréter, car ils sont très variables notamment selon la façon dont est construite la classification de référence utilisée pour l'entraînement et l'évaluation des classifieurs. Cette classification de référence est en effet produite automatiquement en comparant les hypothèses de traduction avec des traductions de référence, et il est difficile de s'assurer de la qualité de cette annotation automatique, qui en plus dépend de plusieurs paramètres.

L'auteur de [Quir 04], quant à lui, utilise des scores produits par le système de traduction utilisé (MSR MT [Mene 01]), ainsi que la perplexité et la longueur de la phrase source et de la traduction produite comme paramètres prédictifs. La classification est ensuite effectuée en utilisant un réseau de neurones ou une *machine à vecteurs de supports* (*SVM* pour *Support Vectors Machine* — voir partie II, section 1.4.2). L'auteur montre que les SVM donnent des résultats légèrement meilleurs que les réseaux de neurones. Il montre également qu'utiliser une classification de référence de petite taille, mais réalisée par des experts, permet une meilleure classification que l'utilisation d'un corpus important, mais annoté automatiquement. Cependant, je proposerai dans la partie II, section 2.3 une autre méthode pour produire automatiquement de grandes quantités de données automatiquement annotées dont l'utilisation donne de meilleurs résultats que celle d'un corpus restreint annoté par des experts.

[Spec 09] présente une estimation de la qualité des traductions selon des critères et sur

une échelle similaire à ceux utilisés par un juge humain (section 1.6.1). Les auteurs utilisent pour cela l'essentiel des paramètres présentés dans [Blat 03], et effectuent une *Régression des Moindres Carrés Partiels* (*Partial Least Square Regression, PLSR* [Abdi 03] — voir partie II, section 1.4.4). Les auteurs proposent également une liste des paramètres qu'ils ont déterminés être les plus pertinents, essentiellement les paramètres utilisant les modèles de langues source et cible, les longueurs de la phrase source et de la traduction, la correspondance entre leurs signes de ponctuation, des informations sur l'alignement et des indications sur la cohérence de la phrase source avec les corpora entraînés pour l'estimation du modèle de traduction.

Enfin, je mentionnerai [Xion 10] et [Bach 11], qui proposent d'utiliser des informations linguistiques (reposant sur des connaissances expertes) comme paramètres prédictifs pour l'estimation de confiance : le premier estime la probabilité de correction d'un mot en utilisant, entre autres les classes grammaticales de ses voisins et une analyse grammaticale de la phrase. Le second utilise, entre autres paramètres, une grammaire de dépendance pour vérifier les accords à longue distance entre les mots (c'est-à-dire, hors de la portée des historiques utilisés par les modèles de langue traditionnels, notamment n-gramme). Ces sources d'information supplémentaires permettent d'améliorer quelque peu la détection d'erreur, mais les performances restent similaires à celles des systèmes purement statistiques.

Dans les sections 1 et 3 de la partie II, je proposerai une formalisation de l'estimation de confiance par classification et régression multivariée, ainsi que des expériences autour des algorithmes d'apprentissage automatique de l'état-de-l'art, des paramètres prédictifs déjà utilisés dans la littérature, d'autres adaptés de l'estimation de confiance pour la reconnaissance automatique de la parole, ainsi que plusieurs paramètres originaux (sections 3.1.4, 3.1.5, 3.2.2).

2.2 Applications des mesures de confiance

On imagine sans peine que connaître la correction d'un mot ou d'une phrase est une information précieuse. Les applications sont bien sûr nombreuses. Les auteurs de [Blat 03] proposent d'utiliser des mesures de confiance pour déterminer, au sein d'une n-best list (section 1.5.5), laquelle est la meilleure. Cette opération s'appelle *reclassement* (*reranking*), car les hypothèses d'une n-best list sont déjà classées par le système de traduction, et cette méthode modifie le classement, pour ensuite choisir la nouvelle meilleure hypothèse.

Les auteurs de [Jaya 05] décrivent un système, *Multi-Engine Machine Translation (MEMT)*, permettant de combiner les hypothèses de traductions produites par différents systèmes, pour en générer de nouvelles. Chaque système propose une estimation de confiance pour l'hypothèse qu'il produit. Ces hypothèses sont ensuite recombinées, en utilisant les estimations de confiance pour sélectionner les mots qui reviennent le plus souvent dans les hypothèses ayant le meilleur score. Ainsi, les auteurs obtiennent une amélioration de 17% du score METEOR [Bane 05] par rapport à la meilleure hypothèse fournie par un des systèmes.

[Spec 10, Spec 11] présentent une méthode pour prédire le temps que demandera la correction des hypothèses à un expert humain. En effet, lorsqu'un document traduit est destiné à être publié, il doit systématiquement être relu par un humain, et certaines phrases doivent être corrigées. Ce processus s'appelle *post-édition*. Mais certaines traductions sont trop mauvaises pour être corrigées. Cela représente une perte de temps et une source frustration pour un traducteur, qui préférera la traduire entièrement (voir également les conclusions de notre expérience de post-édition, partie II, section 4.3). Lucia Specia *et al.* proposent donc d'utiliser des mesures de confiance pour détecter les traductions qui sont trop mauvaises pour être corrigées, et soumettre alors au traducteur la phrase source pour qu'il la traduise lui-même.

He *et al.* proposent également dans [He 10] d'utiliser les mesures de confiance pour écarter des traductions trop mauvaises pour être corrigées par un humain, mais dans un contexte un peu différent : les traducteurs professionnels utilisent parfois des *mémoires de traduction*, des logiciels qui enregistrent les phrases, segments ou mots déjà traduits, et proposent à l'utilisateur la traduction qu'il a utilisée par le passé (*l'historique* de traduction). Lorsque l'élément traduit n'a jamais été rencontré, le logiciel ne propose rien. He *et al.* proposent d'étendre ce système avec des *prédictions de traduction* : ils y incorporent un système de traduction automatique statistique, qui fera des propositions lorsque l'élément traduit n'est pas dans l'historique. Afin de ne pas proposer de traductions de trop mauvaise qualité, source de perte de temps et de frustration, une estimation de confiance est calculée. La traduction automatique ne sera proposée que si la confiance dépasse un certain seuil, qui peut être choisi par l'utilisateur. Toujours dans le domaine de l'aide à la traduction, les auteurs de [Gand 03] proposent d'utiliser plusieurs systèmes de prédiction de traduction, et d'utiliser des mesures de confiance pour choisir la meilleure prédiction.

On pourrait aussi imaginer que des systèmes grand public, comme Google Translate ou Babelfish, proposent une estimation de confiance à leurs utilisateurs : en effet, beaucoup des utilisateurs de ces systèmes les utilisent pour traduire des phrases ou des pages web rédigées dans des langues qu'ils ne maîtrisent pas ou mal. Ils n'ont donc aucun moyen de savoir si la traduction proposée est correcte ou non. Une estimation automatique de la fiabilité des résultats pourrait permettre d'éviter les contresens.

2.3 Apports de la thèse

Dans les chapitres qui suivent j'exposerai les mesures de confiance de l'état de l'art qui me semblent les plus représentatives et les plus importantes (parmi la multitude existant), ainsi que certaines mesures originales (sections 3.1 et 3.2). Je présenterai également les résultats de la combinaison de plusieurs mesures de confiance et autres paramètres prédictifs en utilisant diverses méthodes : régression logistique, réseau de neurones, machine à vecteurs de support et régression des moindres carrés partiels (sections 3.1.8 et 3.2.6). L'article [Blat 03] propose déjà un tour d'horizon des mesures de confiance existantes et m'a servi de point de départ pour la formalisation du problème, la création d'un système de référence et le développement de nouvelles mesures, basées sur l'information mutuelle et les *Parties-du-Discours* (que nous abrégerons en *PoS* pour l'anglais *Part of Speech*). J'ai également développé une méthode pour générer des données d'apprentissage pour les classifieurs, permettant d'obtenir à faible coût une grande quantité de données de bonne qualité.

On évalue les performances d'une mesure de confiance à l'aune des *taux d'erreurs* ou de l'*information qu'elle apporte* (section 1.5). Afin de les mesurer, j'ai implémenté les techniques décrites dans [Siu 99]. Dans les sections 3.1.8 et 3.2.6, je montrerai qu'utiliser un classifieur pour combiner différentes mesures de confiance et paramètres prédictifs permet d'obtenir une amélioration d'1,3 points en termes de taux d'égale erreur (section 1.5.1) par rapport à la meilleure mesure.

Nous présentons dans la section 4 les résultats d'une expérience de post-édition : nous avons demandé à des volontaires de corriger des traductions dans lesquelles les erreurs détectées par les mesures de confiance étaient soulignées, afin de mesurer d'éventuels gains de productivité. Les résultats ne montrent pas d'amélioration de la vitesse ni de la qualité de la post-édition, mais l'expérience nous a permis d'identifier les défauts du système et les caractéristiques vraiment importantes pour les utilisateurs.

La section 5 décrira notre participation à la tâche d'estimation automatique de qualité de la campagne d'évaluation WMT 2012 (`http://www.statmt.org/wmt12`).

Enfin, la partie III, la dernière de ce manuscrit, est consacrée à la présentation d'un système de traduction de la parole spontanée tirant parti des mesures de confiance pour améliorer la robustesse aux erreurs transcription, et pour offrir à l'utilisateur une estimation de la qualité des traductions qui lui sont proposées.

Deuxième partie

Mesures de confiance

1

Théorie des mesures de confiance

1.1 Estimation de la qualité d'une phrase

Nous savons intuitivement distinguer une traduction correcte d'une traduction erronée. Si une traduction n'a pas le même sens que la phrase source, si elle n'a aucun sens ou si elle contient des erreurs, la traduction doit être corrigée. Mais la notion de « traduction erronée » dépend de l'usage qu'on veut en faire : si les traductions des débats au Parlement Européen, par exemple, doivent être parfaites, une documentation technique pourra elle souffrir des tournures peu idiomatiques, et un utilisateur qui veut simplement saisir le sens général d'une page web pourra accepter des erreurs de grammaire. De plus, le sens d'une phrase dépend parfois de l'interprétation du lecteur (particulièrement si la phrase est sortie de son contexte). Une phrase n'est donc pas « correcte » ou « incorrecte » de façon univoque. Il faut donc abandonner tout espoir d'obtenir un classifieur parfait.

Vus ces obstacles à une définition formelle de la correction d'une traduction, et en cohérence avec notre approche de la traduction automatique, nous avons suivi une approche empirique de l'estimation de confiance : nous avons développé des modèles statistiques de mesures de confiance dont les paramètres sont estimés sur des exemples créés par des humains. Cet ensemble d'exemples est appelé *corpus*. Nous avons donc demandé à six volontaires d'annoter des phrases traduites automatiquement afin de constituer notre corpus, avec le barème suivant :

1 : La traduction est inutilisable.

2 : La traduction est assez mauvaise, mais une partie de l'information de la phrase source se retrouve dans la traduction.

3 : On comprend globalement le sens, mais la traduction pourrait être nettement améliorée.

4 : La traduction a des défauts mineurs.

5 : La traduction est très bonne.

Nous avons entraîné le système de mesure de confiance à, d'une part, approcher ces scores (tâche de régression numérique), et d'autre part, classifier comme incorrectes les phrases ayant un score de 1 ou 2 (tâche de classification automatique), ce qui correspond à un objectif d'assimilation (le but est de comprendre le document, pas de publier la traduction automatique telle quelle), et non de diffusion (la traduction proposée doit être publiable). Plutôt que chercher à modéliser le sens d'une phrase ou estimer sa correction grammaticale, ce qui semble hors de portée des méthodes actuelles de traitement des langues naturelles dans le cas général, nous calculons de nombreuses mesures de confiance « simples » (comme la probabilité selon un modèle de langue n-gramme) et des paramètres prédictifs — c'est-à-dire des indicateurs numériques

n'ayant pas forcément d'interprétation directe en termes de confiance — comme le nombre de mots dans la phrase source. Ces indicateurs forment une représentation numérique de la phrase source et de la traduction proposée, plus simple et plus accessible aux méthodes numériques que les phrases elles-mêmes, mais dont on espère qu'elle conserve assez d'information pour qu'un classifieur puisse estimer la qualité de la traduction.

1.2 Estimation de la fiabilité d'un mot

Définir la correction (c'est-à-dire le caractère correct ou non) d'un mot est encore plus délicat. Parfois, il est clair qu'un mot de la traduction n'a aucun rapport avec la phrase source, le problème est alors simple. Cela peut être le cas lorsqu'elle contient des homonymes, par exemple dans la traduction suivante :

`I love rock'n roll.` → `J'aime rocher et rouler.`

Il ne fait aucun doute que `rock'n roll` ne doit pas ici être traduit mot-à-mot mais par `le rock'n roll`, et que cette traduction est complètement erronée. Mais ce n'est pas toujours aussi simple. Considérons l'exemple suivant :

$$\texttt{These words have close meanings} \rightarrow \begin{cases} 1: & \texttt{Ces mots ont des sens proches} & \textit{(correct)} \\ 2: & \texttt{Ces mots sont presque synonymes} & \textit{(correct)} \\ 3: & \texttt{Ces mots sont presque proches} & \textit{(incorrect)} \end{cases}$$

La troisième hypothèse est indubitablement une mauvaise traduction, mais quel mot exactement est faux ? `presque` ? `proches` ? les deux ? Cet exemple montre une fois de plus qu'il est impossible d'obtenir une détection parfaite des mots erronés, tout simplement parce que cela n'a pas de sens. Néanmoins, comme nous le verrons plus loin, il est possible, en utilisant des paramètres prédictifs et des classifieurs, d'approcher le jugement humain.

1.3 Formalisation

Le problème peut sembler inextricable. La difficulté de modéliser la sémantique, la variabilité du vocabulaire, la complexité de la syntaxe et l'ambiguïté de la classification semblent placer l'estimation de correction hors de notre portée. Pourtant, une formalisation rigoureuse va permettre de la réduire à des problèmes connus, et accessibles au calcul. L'idée est de simplifier la représentation des phrases et des mots en remplaçant leur représentation graphémique par une représentation numérique. Cette simplification permet d'utiliser les techniques d'apprentissage statistique classique. Afin d'être pertinente, la représentation numérique doit s'attacher à conserver autant que possible toute l'information portée par la représentation textuelle. En plus de rendre le problème accessible aux méthodes numériques, définir rigoureusement un cadre formel de l'estimation de correction permet d'interpréter plus précisément les résultats, et de développer, tester et comparer facilement de nouvelles mesures de confiance. Nous allons donc détailler dans cette section une formalisation rigoureuse des mesures de confiance, en partant du concept élémentaire de *vocabulaire* pour arriver, par développements successifs, à celui de traduction et d'estimation de confiance.

L'objectif de la traduction automatique est de générer une phrase dans une langue (hypothèse) étant donnée une phrase dans une autre langue (source). Une phrase est une suite finie de

mots et de signes de ponctuation, qui font partie d'un ensemble dénommé *vocabulaire*. Tous les éléments du vocabulaire seront appelés *graphèmes*. Les phrases et les graphèmes sont modélisés par des variables aléatoires. J'utiliserai les conventions suivantes (en plus des notations mathématiques habituelles) : une variable aléatoire sera représentée par une lettre majuscule (X) et une réalisation de cette variable par une lettre minuscule (x). Les vecteurs et autres variables multidimensionnelles sont en gras, ainsi que les fonctions à valeurs non scalaires ; les scalaires et les fonctions à valeurs scalaires sont en maigre. Ainsi, par exemple, le vecteur $\mathbf{x} \in \mathbb{R}^d$ est une réalisation de la variable aléatoire $\mathbf{X}$, et le scalaire $y \in \mathbb{R}$ une réalisation de Y. Les lettres cursives ($\mathcal{V}$) sont des ensembles, et l'ajout d'une étoile ($\mathcal{V}^*$) désigne l'ensemble des séquences finies d'éléments de cet ensemble (les mots, au sens de la théorie des langages), sauf $\mathbb{R}^*$, $\mathbb{N}^*$ et $\mathbb{Z}^*$ qui gardent leur sens habituel. Une séquence pourra être notée $x_1, .., x_n$, $(x_i)_{i=1..n}$ ou, par économie de place, $\mathbf{x}_1^n$. $Len(\mathbf{x})$ désigne la longueur d'un mot, d'une phrase ou d'une séquence, ou la dimension d'un vecteur.

$\mathcal{V}_S$	:	vocabulaire de la langue source.
$\mathcal{V}_T$	:	vocabulaire de la langue cible (T pour « target »).
$\mathbf{S} \in \mathcal{V}_S^*$	:	phrase (au sens de suite de mots) dans la langue source.
$\mathbf{T} \in \mathcal{V}_T^*$	:	phrase dans la langue cible.

De ces deux variables aléatoires de base, nous dérivons deux autres :

$Len(\mathbf{S}) \in \mathbb{N}$	:	longueur de $\mathbf{S}$ (nombre de mots)
$Len(\mathbf{T}) \in \mathbb{N}$	:	longueur de $\mathbf{T}$
$S_i \in \mathcal{V}_S$	:	i-ème mot de $\mathbf{S}$
$T_j \in \mathcal{V}_T$	:	j-ème mot de $\mathbf{T}$

Calculer des mesures de confiance, cela signifie qu'on a des réalisations des variables ci-dessus, et que l'on doit estimer les valeurs de :

$C_{\mathbf{S},\mathbf{T}} \in \{0,1\}$	:	correction de la phrase $\mathbf{T}$ en tant que traduction de $\mathbf{S}$
$C_{\mathbf{S},\mathbf{T},j} \in \{0,1\}$	:	correction du j-ème mot de $\mathbf{T}$

À cette fin il faut estimer les distributions de probabilité suivantes :

$$P(C_{\mathbf{S},\mathbf{T}} = 1|\mathbf{S},\mathbf{T}) \quad : \quad \text{probabilité que } \mathbf{T} \text{ soit une traduction correcte de } \mathbf{S} \tag{1.1}$$

$$P(C_{\mathbf{S},\mathbf{T},j} = 1|\mathbf{S},\mathbf{T}) \quad : \quad \text{probabilité que le j-ème mot de } \mathbf{T} \text{ soit correct} \tag{1.2}$$

$\mathbf{S}$ et $\mathbf{T}$ peuvent être n'importe quelle phrase. L'estimation directe de ces probabilités est donc hors de la portée des méthodes actuelles de traitement des langues naturelles. On va donc passer par une représentation numérique plus simple. La paire $(\mathbf{S},\mathbf{T})$ est représentée par un vecteur $\mathbf{x}^{\mathbf{s}}(\mathbf{S},\mathbf{T}) \in \mathbb{R}^{d_s}$ de *paramètres prédictifs*. d_s est la dimensionalité de la représentation numérique d'une paire de phrases, c'est-à-dire le nombre de paramètres prédictifs utilisés. Ces paramètres seront décrits dans la section 3.2. De même, le j-ème mot de l'hypothèse $\mathbf{T}$, c'est-à-dire formellement le triplet $(\mathbf{S},\mathbf{T},j)$ est représenté par $\mathbf{x}^{\mathbf{w}}(\mathbf{S},\mathbf{T},j) \in \mathbb{R}^{d_w}$, avec d_w le nombre de paramètres prédictifs utilisés (section 3.1).

$$\mathbf{x}^{\mathbf{s}} : (\mathbf{S},\mathbf{T}) \in \mathcal{V}_S^* \times \mathcal{V}_T^* \rightarrow \mathbf{x}^{\mathbf{s}}(\mathbf{S},\mathbf{T}) \in \mathbb{R}^{d_s}$$

$$\mathbf{x}^{\mathbf{w}} : (\mathbf{S},\mathbf{T},j) \in \mathcal{V}_S^* \times \mathcal{V}_T^* \times \mathbb{N} \rightarrow \mathbf{x}^{\mathbf{w}}(\mathbf{S},\mathbf{T},j) \in \mathbb{R}^{d_w}$$

Ces paramètres peuvent être, par exemple, la longueur de la phrase source, le score de $\mathbf{T}$ selon un modèle de traduction ou un modèle de langage, etc. Certains paramètres prédictifs sophistiqués (par exemple, la vraisemblance d'une phrase selon un modèle de langage n-gramme) peuvent s'interpréter directement comme des mesures de confiance. Dans le cas contraire, ou si on veut utiliser plusieurs paramètres, on estime les distributions de probabilité suivantes, plutôt que celles exprimées dans les équations 1.1 et 1.2, dont elles sont des approximations :

$$p(C_{\mathbf{S},\mathbf{T}}; \mathbf{S}, \mathbf{T}) \stackrel{def}{=} P(C_{\mathbf{S},\mathbf{T}} | \mathbf{x}^{\mathbf{s}}(\mathbf{S}, \mathbf{T})) \tag{1.3}$$

$$p(C_{\mathbf{S},\mathbf{T},j}; \mathbf{S}, \mathbf{T}, j) \stackrel{def}{=} P(C_{\mathbf{S},\mathbf{T},j} | \mathbf{x}^{\mathbf{w}}(\mathbf{S}, \mathbf{T}, j)) \tag{1.4}$$

Bien que cela n'apparaisse pas explicitement dans la notation, p dépend de la fonction $\mathbf{x}$, qui d'une part n'est pas la même pour les mots et pour les phrases, d'autre part diffère dans les différentes expériences qui seront présentées par la suite. Ces distributions seront entraînées sur de vastes corpora (section 2) en utilisant des algorithmes classiques (Section 1.4), comme les Machines à Vecteurs de Support (*Support Vector Machine, SVM*) [Cort 95], les Réseaux de Neurones (*Neural Networks, NN*) [Faus 94], la Régression Logistique (*Logistic Regression, LR*) [Mena 02] ou la Régression des Moindres Carrés Partiels (*Partial Least Square Regression, PLSR*) [Tobi 95].

1.3.1 Classification

Une fois les distributions de probabilité estimées (Équations 1.3 et 1.4), on peut les utiliser comme mesures de confiance. On peut ensuite déterminer une classification des mots ou des phrases :

$$\hat{c} : (\mathbf{T}, \mathbf{S}) \rightarrow \hat{c}(\mathbf{T}, \mathbf{S}) \in \{0, 1\}$$

ou au niveau des mots :

$$\hat{c} : (\mathbf{T}, \mathbf{S}, j) \rightarrow \hat{c}(\mathbf{T}, \mathbf{S}, j) \in \{0, 1\}$$

Afin de minimiser le nombre d'erreurs de classification, celle-ci doit se faire selon la règle :

$$\hat{c}(\mathbf{T}, \mathbf{S}) \stackrel{def}{=} \arg\max_{c \in \{0,1\}} p(c; \mathbf{S}, \mathbf{T}) \tag{1.5}$$

$$\hat{c}(\mathbf{T}, \mathbf{S}, j) \stackrel{def}{=} \arg\max_{c \in \{0,1\}} p(c; \mathbf{S}, \mathbf{T}, j) \tag{1.6}$$

Cependant cette classification ne prend en compte ni un éventuel biais des estimations de probabilité (par exemple, si la probabilité de correction est systématiquement surévaluée, il faut le prendre en compte et la décision optimale n'est pas l'`argmax` des équations 1.5 et 1.6), ni la différence entre les coûts d'un faux rejet et d'une fausse acceptation (en effet, la gêne pour un utilisateur n'est pas forcément la même quand un mot correct est indûment rejeté et quand un mot faux est détecté à tort comme correct ; voir les sections sur l'évaluation des classifieurs — section 1.5.1 — et sur les conclusions de l'expérience de post-édition — section 4.3). Nous introduisons donc un *seuil d'acceptation* δ :

$$\hat{c}(\mathbf{T},\mathbf{S};\delta) \stackrel{def}{=} \begin{cases} 1 & \text{si } p(1;\mathbf{S},\mathbf{T}) \geqslant \delta \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

$$\hat{c}(\mathbf{T},\mathbf{S},j;\delta) \stackrel{def}{=} \begin{cases} 1 & \text{si } p(1;\mathbf{S},\mathbf{T},j) \geqslant \delta \\ 0 & \text{sinon} \end{cases} \tag{1.8}$$

Si $\delta = 0.5$, les formules 1.7 et 1.8 sont équivalentes aux formules 1.5 et 1.6. Mais choisir un seuil plus élevé permet de compenser un biais positif des estimations de probabilité, ou de pénaliser plus sévèrement les fausses acceptations que les faux rejets ; au contraire, un seuil plus bas permet de compenser un biais négatif, et de diminuer le nombre de faux rejets (au prix d'un nombre plus élevé de fausses acceptations).

1.3.2 Biais

Les estimations de probabilités des formules 1.3 et 1.4 peuvent être biaisées, notamment parce que les données d'apprentissage ne viennent pas forcément du même corpus que le corpus de test ou les phrases soumises à un système en production, nécessairement inconnues voire indisponibles lors de la phase d'apprentissage. En général, cela n'a pas d'influence sur la classification, pour deux raisons : d'abord, lorsque le biais est uniforme ($p_{biais} = p + b_0$ avec b_0 une constante), ou plus généralement lorsqu'il est monotone (c'est-à-dire $p_{biais} = p + b(p)$ avec b la fonction représentant le biais, et $p_{biais} \geqslant p'_{biais} \Leftrightarrow p \geqslant p'$), le biais est compensé très simplement en choisissant un seuil adapté ; mais surtout, ces distributions de probabilités sont estimées en minimisant l'erreur de classification. Il n'est donc pas surprenant que même si on observe un biais, les exemples corrects obtiennent une meilleure « probabilité » que les exemples incorrects.

En revanche, le biais est néfaste au regard d'autres critères d'évaluation, comme l'Information Mutuelle Normalisée (section 1.5.2) [Siu 99]. Le biais sera donc estimé sur un corpus dédié à cet usage, comme décrit par Siu *et al.*, et corrigé. Pour ce faire, l'intervalle $[0,1]$ est divisé en 1000 segments disjoints $\mathcal{B}_i$ de même taille (quoi qu'il puisse être plus judicieux, selon la nature du biais, d'adopter un échantillonage non uniforme). Le biais est considéré comme constant sur chacun de ces intervalles et estimé en utilisant le corpus dédié.

$$\begin{aligned} [0,1] &= \cup_{i=1}^{1000} \mathcal{B}_i \text{avec } \forall i,j \in \{1,1000\}\ i \neq j \Rightarrow \mathcal{B}_i \cap \mathcal{B}_j = \emptyset \\ b(\mathcal{B}_i) &= \frac{\sum_{j|\hat{p}_j \in \mathcal{B}_i} (\hat{p}_j - c_j^*)}{\sum_{j|\hat{p}_j \in \mathcal{B}_i} 1} \ \forall i \in \{1,..,1000\} \end{aligned} \tag{1.9}$$

où les $\hat{p}_j$ sont les probabilités de correction estimées (donc biaisées) des éléments du corpus d'estimation du biais, et les c_j^* leurs véritables classes (on ne dispose en effet que de cette information, pas des probabilités non biaisées). On obtient ainsi une fonction de compensation du biais :

$$\text{si } p \in \mathcal{B}_i : unbias(p) = p - b(\mathcal{B}_i) \tag{1.10}$$

Si $\hat{p}$ est une probabilité de correction estimée par notre système, dans nos applications nous compenserons le biais :

$$p(1;\mathbf{S},\mathbf{T}) = unbias(\hat{p})$$

1.3.3 Estimation de la qualité d'une traduction

Certaines applications ne demandent pas une classification des phrases en deux catégories « correct/incorrect », mais plutôt une estimation globale de leur qualité (par exemple, lorsqu'on doit décider quelle phrase retravailler dans un document avant de le publier : même correcte, une phrase peut être trop maladroite pour être acceptable). C'est la raison pour laquelle, lors des campagnes d'évaluation, les traductions sont notées sur une échelle de 1 à 5 (section 1.1), au lieu d'êtres étiquetées « correcte » ou « incorrecte ». C'est l'usage qu'on veut faire des traductions automatiques qui détermine ensuite le seuil d'acceptabilité. Le but de l'estimation automatique de la qualité d'une traduction est de fournir sans intervention humaine une telle évaluation, un peu à la manière de la métrique BLEU [Papi 01] ou du Taux d'Édition de la Traduction (*Translation Edit Rate, TER*) [Snov 06], mais sans nécessiter de traduction de référence. Dans la section 3.2, je présenterai donc des méthodes permettant l'estimation soit d'une distribution de probabilités (Équation 1.3), soit de scores reflétant la qualité de la traduction proposée. Le corpus d'apprentissage utilisé pour cette tâche peut être décrit comme suit :

$$\left\{ \left(\mathbf{x}^{\mathbf{s}}(\mathbf{s}^n, \mathbf{t}^n); q^*_{\mathbf{s}^n,\mathbf{t}^n} \right) \right\}_{n=1..N} \subset \mathbb{R}^{d_s} \times \mathbb{R}^+$$

où $q^*_{\mathbf{s}^n,\mathbf{t}^n}$ est un score de référence (provenant d'une évaluation humaine, ou calculé en comparant la traduction à une référence, à l'aide de BLEU ou TER par exemple). L'objectif est d'apprendre une fonction $f_\Theta : \mathbb{R}^{d_s} \rightarrow \mathbb{R}^+$ associant la représentation d'une phrase (le vecteur des caractéristiques numériques) à un score. Θ désigne l'ensemble des paramètres contrôlant le comportement de la fonction, qui seront optimisés par des techniques de régression (linéaire, moindres carrés partiels, vecteurs de support, etc.) pour minimiser l'erreur quadratique moyenne :

$$\frac{1}{N} \sum_{n=1}^{N} |f_\Theta \left(\mathbf{x}^{\mathbf{s}}(\mathbf{s}^n, \mathbf{t}^n) \right) - q^*_{\mathbf{s}^n,\mathbf{t}^n}|^2 \tag{1.11}$$

La qualité des scores prédits sera évaluée en mesurant leur corrélation avec les scores réels (section 3.2), la racine carrée de l'erreur quadratique moyenne ou l'erreur absolue moyenne pour la campagne WMT (section 5).

1.3.4 Corpora pour l'apprentissage

Les distributions de probabilité 1.3 et 1.4 doivent être apprises sur de grands corpora pour être précises. Les données sont de la forme :

- n phrases source $\mathbf{s}^1, .., \mathbf{s}^n$, c'est-à-dire des réalisations de la variable aléatoire $\mathbf{S}$.
- Les traductions automatiques correspondantes $\mathbf{t}^1, .., \mathbf{t}^n$, c'est-à-dire des réalisations de $\mathbf{T}$.
- Des classifications de référence (c'est-à-dire des réalisations de $C_{\mathbf{S},\mathbf{T}}$) et/ou des scores de référence
  $(c^*_1, q^*_1), .., (c^*_n, q^*_n) \in \{0, 1\} \times \mathbb{R}^+$; ces scores peuvent provenir d'une évaluation humaine (Section 2.1), être calculés automatiquement en comparant la traduction automatique à une ou plusieurs traductions de référence (Section 2.2), ou être calculés de façon entièrement automatique sur des corpora spéciaux (Section 2.3).
- Classifications de référence pour les mots, c'est-à-dire des réalisations de la variable $C_{\mathbf{S},\mathbf{T},j}$, provenant d'une évaluation de référence ou calculés automatiquement :

$$\left((c^*_{k,1}, .., c^*_{k,Len(\mathbf{t}^k)}) \in \{0, 1\}^{Len(\mathbf{t}^k)} \right)_{k \in \{1,..,n\}}$$

1.4 Algorithmes de classification et de régression

La question de l'estimation de confiance est donc maintenant formalisée comme un problème standard de classification et de régression. Nous utiliserons donc les outils disponibles implémentant les techniques bien connues de Machine à Vecteurs de Support, Régression Logistique, Réseaux de Neurones et Régression des Moindres Carrés Partiels. Afin d'alléger la lecture et lorsque ce n'est pas ambigu, nous identifierons un ensemble de paramètres prédictifs et le classifieur (ou la régression) qui en dérive (par exemple, nous écrirons « les mesures basées sur les n-grammes fonctionnent bien » plutôt que « le réseau de neurones utilisant les mesures basées sur les n-grammes etc. »).

1.4.1 Régression Logistique

L'objectif est de prédire la valeur de $C \in \{0, 1\}$ étant donné un vecteur de paramètres $\mathbf{X} \in \mathbb{R}^d$. On cherche donc à estimer la distribution de probabilité $P(C = 1|\mathbf{X})$. La Régression Logistique [Mena 02] consiste à supposer qu'elle peut s'écrire sous la forme :

$$P(C = 1|\mathbf{X}) \quad = \quad \frac{1}{1 + e^{\langle \Theta, \mathbf{X} \rangle + b}} \tag{1.12}$$

Pour un certain $\Theta \in \mathbb{R}^d$ et un certain $b \in \mathbb{R}$. On optimise donc Θ et b sur les données selon le critère de maximum de vraisemblance. La Régression Logistique peut être utilisée pour combiner plusieurs mesures de confiance, ou faire correspondre un score à une probabilité (dans ce cas, les vecteurs $\mathbf{X}$ et Θ sont de dimension 1).

1.4.2 Machines à Vecteurs de supports

Les Machines à Vecteur de Support (SVM) [Hsu 03] ont des propriétés intéressantes qui en font des classifieurs très adaptés à notre problème. Outre la possibilité d'estimer la probabilité qu'a un échantillon d'appartenir à une classe, on peut s'en servir pour distinguer des classes non linéairement séparables dans un espace euclidien, et elles permettent également d'effectuer une régression [Smol 04]. Nous avons utilisé LibSVM [Chan 11] pour la normalisation des caractéristiques, la classification et la régression.

Normalisation des paramètres : Étant donnée une fonction $\mathbf{x} : \mathcal{V}_S^* \times \mathcal{V}_T^* \rightarrow \mathbb{R}^{d_S}$ (c'est-à-dire un vecteur de paramètres prédictifs), il convient d'effectuer une normalisation préalable. Par exemple, si un des paramètres choisis est à valeur dans $[0, 1]$ et un autre à valeur dans $[0, 0.001]$, le second risque d'être ignoré au profit du premier, même s'il est plus informatif. Même si cela n'est pas en théorie nécessaire, pour des raisons numériques, normaliser les paramètres permet d'éviter ce problème. En pratique, les paramètres de la normalisation sont appris sur le corpus d'apprentissage (c'est-à-dire, comment modifier la valeur de chaque composante du vecteur), et la normalisation est effectuée sur les autres corpora (développement et test) en utilisant les mêmes paramètres. Nous utilisons pour cela l'outil `svm-scale` de LibSVM.

Classification par SVM : Les SVM sont entraînées pour estimer la probabilité de correction, plutôt que pour générer directement une classification. Cela permet d'adapter le seuil d'acceptation et de compenser le biais (section 1.3.1 et 1.3.2). Nous utilisons le noyau Radial Basis Function (RBF), qui est simple et donne en général de bons résultats [Zhan 01] :

$$K_\gamma(\mathbf{x}(\mathbf{s}, \mathbf{t}, j), \mathbf{x}(\mathbf{s}', \mathbf{t}', j')) = e^{-\gamma \|\mathbf{x}(\mathbf{s},\mathbf{t},j) - \mathbf{x}(\mathbf{s}',\mathbf{t}',j')\|^2}$$

Estimation de qualité par SVM : On utilise le même noyau, mais cette fois pour effectuer une régression sur le score BLEU de chaque phrase [Papi 01]. Nous sommes bien conscients que BLEU n'est pas une métrique très adaptée au niveau des phrases, et que TER serait plus indiqué. Cependant, nous avons observé qu'en pratique, beaucoup de traductions obtenaient un score TER très bas, alors que le jugement humain correspondant était de 2 ou 3 selon les critères donnés à la section 1.1. Tout en étant bien conscients de ses limites, nous avons donc décidé d'utiliser malgré tout la mesure BLEU, bien connue de la communauté et dont les performances sont suffisantes pour les besoins de nos expériences.

Optimisation des méta-paramètres : L'entraînement d'une SVM nécessite, outre l'apprentissage du modèle lui-même (les vecteurs de support) l'optimisation de deux méta-paramètres, le coefficient γ du noyau RBF ci-dessus, et le coût C d'une erreur. Nous les avons optimisés au regard du taux d'égale erreur (voir section 1.5.1) pour les tâches d'estimation de probabilité, ou de l'erreur quadratique moyenne pour les tâches de régression, en effectuant une grid search (« recherche par quadrillage ») sur un corpus de développement. Nous avons observé que les performances des SVM sur notre tâche d'estimation de probabilité (section 3.1.8) sont très sensibles aux variations de ces méta-paramètres.

1.4.3 Réseaux de neurones

Nous utilisons la bibliothèque FANN (*Fast Artificial Neural Network*, [Niss 03]) pour implémenter des réseaux de neurones. Après une phase d'expérimentation, nous avons opté pour le perceptron multicouche, une structure classique : une couche d'entrée contenant autant de neurones que nous calculons de paramètres prédictifs, une couche cachée en contenant moitié moins, et une couche de sortie n'en contenant qu'un seul. Nous avons choisi un taux de connectivité de 0.5 afin de ne pas augmenter déraisonnablement le temps de calcul. Nous utilisons la fonction d'activation sigmoïde classique. L'optimisation des poids se fait par l'algorithme classique de rétro-propagation du gradient.

1.4.4 Régression des Moindres Carrés Partiels

La Régression des Moindres Carrés Partiels (Partial Least Squares Regression, PLSR) [Wold 84, Spec 09] est une technique d'analyse de données multivariées qui détermine une relation bilinéaire entre les variables observées (nos paramètres prédictifs, c'est-à-dire le vecteur $\mathbf{X}$) et les variables réponse, en l'occurrence la probabilité de correction $p(1;\mathbf{X})$ ou le score de qualité. Le principe est de projeter les observations et les réponses dans un sous-espace vectoriel et d'y effectuer une régression par les moindres carrés. Cette technique a l'avantage d'être robuste aux observations corrélées, propriété très utile dans notre cas car beaucoup de nos paramètres prédictifs sont très similaires.

1.5 Évaluation des classifieurs

Le *taux d'erreur* est la métrique la plus évidente pour mesurer les performances d'un classifieur. Et pourtant, elle est en fait assez inadaptée, car très sensible aux *probabilités a priori* des différentes classes [Kono 91, Siu 99]. Prenons l'exemple d'un système de traduction automatique générant environ 15% de mots incorrects, et d'une mesure de confiance triviale p^0 telle que :

$$\forall \mathbf{s}, \mathbf{t}, j \quad p^0(1; \mathbf{s}, \mathbf{t}, j) = 1$$

Si on choisit un seuil non dégénéré dans l'équation 1.7 ($\delta = 0.5$ par exemple), tous les mots seront toujours acceptés. Donc, elle ne fait aucune erreur sur les mots effectivement corrects (85% du total), mais échoue complètement à détecter le moindre mot incorrect (15%). Son taux d'erreur est donc $0\% \times 0.85 + 100\% \times 0.15 = 15\%$.

Considérons maintenant un autre mesure de confiance p^1 qui détecte correctement tous les mots erronés :

$$p^1(1; \mathbf{s}, \mathbf{t}, j) = 0 \quad \text{si } t_j \text{ est erroné}$$

mais assigne une probabilité nulle ($p^1(1; \mathbf{s}, \mathbf{t}, j) = 0$) à environ 20% des mots corrects, et une probabilité de 1 à tous les autres. Pour $\delta = 0.5$ ou tout seuil non dégénéré, son taux d'erreur est donc :

$$0\% \times 0.15 + 20\% \times 0.85 = 17\%$$

p^0 semble donc meilleure que p^1. Et pourtant, il est évident que p^0 est une mesure complètement inutile, car constante : elle n'apporte aucune information. En revanche, p^1 est une bonne mesure, car elle permet de détecter à coup sûr un mot erroné, ce qui est très utile, par exemple, pour la post-édition. On voit donc qu'une bonne façon de mesurer l'efficacité d'une mesure de confiance n'est pas selon le nombre d'erreurs de classifications qu'elle commet mais selon *la quantité d'information qu'elle fournit*. Nous utiliserons donc *l'Information Mutuelle Normalisée* (Normalised Mutual Information, NMI) [Siu 99] comme mesure de l'information apportée par une mesure de confiance (Section 1.5.2).

Nous utiliserons également la courbe DET (*Discrimination Error Trade-off*, section 1.5.1), une représentation graphique de l'évolution des taux de faux négatifs et de faux positifs en fonction du seuil d'acceptation, et le Taux d'Égale Erreur (*Equal Error Rate, EER*) qui correspond au point de la courbe où ces deux taux sont identiques.

1.5.1 Compromis Détection-Erreur

On peut distinguer deux types d'erreurs commises par un classifieur : *fausse acceptation*, si un mot ou une phrase incorrecte reçoit un score de confiance $\hat{c} = 1$ (on parle aussi d'erreur de type 1), et *faux rejet*, si un mot ou une phrase correcte reçoit le score $\hat{c} = 0$ (erreur de type 2). Le nombre d'erreurs de type 1 et de type 2 varie en fonction du seuil d'acceptation δ.

Lorsqu'on veut évaluer les performances d'un classifieur, on connaît les prédictions $\hat{c}(\mathbf{t}; \mathbf{s}; \delta)$ (équations 1.5 et 1.6), et les véritables classes, c'est-à-dire les réalisations $c^*(\mathbf{t}; \mathbf{s}; \delta)$ des variables $\mathbf{C}$, déterminées par des annotateurs ou en utilisant des traductions de référence (Section 2). Le taux de fausses acceptations (ici, pour les phrases, mais adapter ces formules pour les mots est immédiat) est défini par :

$$e_1(\mathbf{s}, \mathbf{t}; \delta) = \begin{cases} 1 & \text{si} \quad \hat{c}(\mathbf{t}; \mathbf{s}; \delta) = 1 \text{ et } c^*_{\mathbf{s},\mathbf{t}} = 0 \\ 0 & \text{sinon} \end{cases} \tag{1.13}$$

$$err_1(\delta) = \frac{\sum_{\mathbf{s},\mathbf{t}} e_1(\mathbf{s}, \mathbf{t}; \delta)}{\sum_{\mathbf{s},\mathbf{t}} (1 - c^*_{\mathbf{s},\mathbf{t}})} \tag{1.14}$$

err_1 est donc la proportion de phrases (ou de mots) erronées qui sont à tort classifiées comme correctes ($\sum_{\mathbf{s},\mathbf{t}} (1 - c^*_{\mathbf{s},\mathbf{t}})$ est le nombre de phrases ou mots incorrects).

Le taux de faux rejets (pour les phrases) est défini par :

$$e_2(\mathbf{s},\mathbf{t};\delta) = \begin{cases} 1 & \text{si} \quad \hat{c}(\mathbf{t};\mathbf{s};\delta) = 0 \text{ et } c^*_{\mathbf{s},\mathbf{t}} = 1 \\ 0 & \text{sinon} \end{cases} \tag{1.15}$$

$$err_2(\delta) = \frac{\sum_{\mathbf{s},\mathbf{t}} e_2(\mathbf{s},\mathbf{t};\delta)}{\sum_{\mathbf{s},\mathbf{t}} c^*_{\mathbf{s},\mathbf{t}}} \tag{1.16}$$

err_2 est donc la proportion de phrases ou mots corrects qui sont à tort rejetés par le classifieur.

Un classifieur laxiste aura un taux d'erreurs de type 1 élevé et un taux d'erreur de type 2 bas (utile, idéalement, si on veut corriger seulement les erreurs graves), contre l'inverse pour un classifieur sévère (nécessaire si on veut s'assurer que les traductions sont vraiment de bonne facture). Les auteurs de [Siu 99] montrent que err_1 et err_2 sont indépendants des probabilités a priori des classes.

Si on fait varier δ de 0 à 1, de plus en plus d'éléments corrects sont rejetés à tort, mais de moins en moins d'éléments incorrects échappent à la détection. err_1 décroît donc de façon monotone de 1 à 0, et err_2 croît de façon monotone de 0 à 1. Le lieu des points $(err_2(\delta), err_1(\delta))$, est appelé *courbe DET*, pour *Discrimination-Error Trade off*, c'est-à-dire « compromis détection-erreur » : en effet, on peut voir chaque point (qui correspond à un certain δ) comme un compromis entre les deux types d'erreurs. Nous en verrons des exemples dans la section 3.1. Plus la courbe est « basse », meilleur est le classifieur. La courbe d'un classifieur parfait serait réduite au point $(0,0)$. Tous les points de la courbe devraient se trouver sous la diagonale $[(0,1),(1,0)]$, qui est la courbe d'un classifieur attribuant des probabilités au hasard.

err_1 et err_2 peuvent être vus comme des approximations de fonctions continues de δ [15]. On peut donc déterminer un certain seuil δ_{EER} tel que :

$$err_1(\delta_{EER}) \simeq err_2(\delta_{EER}) = EER \tag{1.17}$$

EER est appelé *Taux d'Égale Erreur* (*Equal Error Rate*). On peut le voir comme un compromis « équilibré » entre les erreurs de types 1 et 2. On s'en servira pour comparer les classifieurs, mais il faut garder à l'esprit que selon l'application qu'on veut faire des mesures de confiance, un EER plus bas ne signifie pas forcément un classifieur plus adapté.

1.5.2 Information Mutuelle Normalisée

L'*Information Mutuelle Normalisée* (*Normalised Mutual Information*, NMI) mesure la quantité d'information apportée par une mesure de confiance, de la façon la plus générale possible [Siu 99]. Elle ne dépend pas d'une classification particulière. Intuitivement, la NMI mesure la réduction de l'entropie de la distribution de probabilité de la vraie classe $C \in \{correct, incorrect\}$ lorsqu'on connaît la mesure de confiance.

Soit $\mathbf{x}(\mathbf{S},\mathbf{T})$ un vecteur de paramètres prédictifs ; alors :

15. Ce n'est pas tout à fait exact : même avec une infinité d'échantillons, la fonction n'est pas continue si l'ensemble des scores n'est pas dense dans $[0,1]$. C'est le cas par exemple des mesures de confiance à valeurs dans des ensembles discrets, dont nous verrons des exemples plus loin. C'est cependant le cas de la plupart des mesures réalistes, au moins dans le voisinage des valeurs de δ « raisonnables », et quand ce n'est pas le cas, nous nous contenterons d'approximations.

$$
\begin{aligned}
NMI(C,\mathbf{x}) &= \frac{I(C;\mathbf{x})}{H(C)} = \frac{H(C) - H(C|\mathbf{x})}{H(C)} \\
H(C) &= -p^* log(p^*) - (1-p^*) log(1-p^*) \\
H(C|\mathbf{x}) &= \int_{\mathbf{v}} \Big(P(\mathbf{x}(\mathbf{S},\mathbf{T}) = \mathbf{v}) \times \\
&\qquad \sum_{c \in \{0,1\}} P(C = c|\mathbf{x}(\mathbf{S},\mathbf{T}) = \mathbf{v}) log(P(C = c|\mathbf{x}(\mathbf{S},\mathbf{T}) = \mathbf{v})) \Big) d\mathbf{v}
\end{aligned}
\tag{1.18}
$$

avec I l'information mutuelle, H l'entropie et p^* la vraie probabilité a priori qu'un échantillon (phrase ou mot) soit correct. On ne dispose pas des véritables distributions $P(\mathbf{x}(\mathbf{S},\mathbf{T}))$ et $P(C|\mathbf{x}(\mathbf{S},\mathbf{T}))$, qui sont donc remplacées par des estimations. On a donc :

– NMI pour les phrases :

$$
\begin{aligned}
H(C|\mathbf{x}) \simeq \frac{1}{N} \sum_{(\mathbf{s},\mathbf{t}) \in \mathcal{S}} \Big(& p(1;\mathbf{s},\mathbf{t}) log(p(1;\mathbf{s},\mathbf{t})) \\
& + (1 - p(1;\mathbf{s},\mathbf{t})) log(1 - p(1;\mathbf{s},\mathbf{t})) \Big)
\end{aligned}
\tag{1.19}
$$

– Pour les mots :

$$
\begin{aligned}
H(C|\mathbf{x}) \simeq \frac{1}{N_w} \sum_{(\mathbf{s},\mathbf{t}) \in \mathcal{S}} \sum_{j=1}^{Len(\mathbf{t})} \Big(& p(1;\mathbf{s},\mathbf{t},j) log(p(1;\mathbf{s},\mathbf{t},j)) \\
& + (1 - p(1;\mathbf{s},\mathbf{t},j)) log(1 - p(1;\mathbf{s},\mathbf{t},j)) \Big)
\end{aligned}
\tag{1.20}
$$

$H(C|\mathbf{x})$ est nécessairement positif, ou nul quand pour tout échantillon on a $p(c_{\mathbf{s},\mathbf{t}};\mathbf{s},\mathbf{t}) = 1$, c'est-à-dire que la bonne classe est prédite sans incertitude. De plus, $H(C|\mathbf{x})$ est nécessairement plus petit que $H(C)$, et on a l'égalité si la mesure de confiance n'apporte aucune information. $NMI(\mathbf{x})$ est donc en théorie un nombre réel compris entre 0 et 1. En pratique, à cause des approximations, il peut être négatif.

2

Corpora

Les paramètres des modèles statistiques de correction et de qualité des traductions seront déterminés automatiquement. Des corpora bien fournis sont donc nécessaires pour estimer correctement les distributions de probabilités exprimées dans les équations 1.3 et 1.4. Dans l'idéal, un interprète professionnel devrait annoter les traductions proposées par un système de traduction automatique et assigner à chaque mot et chaque phrase l'étiquette « correct » ou « incorrect ». On obtiendrait ainsi un corpus d'excellente qualité, mais à un coût très élevé. On peut aussi utiliser des méthodes automatiques ou semi-automatiques pour annoter les corpora d'entraînement, de développement et de test. Nous allons dans cette section discuter ces différentes approches.

2.1 Annotation manuelle

C'est une approche « de luxe » : coûteuse et de bonne qualité. Des interprètes professionnels analysent chaque phrase et chaque mot proposé par le système, et décident s'il est correct ou non. L'annotation dépend en principe de l'application (assimilation, diffusion ou post-édition, par exemple). Pour nos expérimentations, nous avons choisi des règles correspondant à une tâche d'assimilation :

1. Un mot est incorrect s'il n'est pas sémantiquement cohérent avec la phrase source ou le contexte (par exemple, le cas où « vol » en parlant d'un trajet aérien serait traduit par « theft »), s'il souffre d'une erreur de grammaire nuisant à la compréhension (erreur d'accord en temps, par exemple), ou s'il est très mal placé.
2. Les phrases sont annotées selon le barème décrit dans la section 1.1 (un score entre 1 et 5 selon la qualité de la traduction).

Cette méthode permet d'obtenir un corpus d'excellente qualité, car très peu d'erreurs d'annotation. Elle a cependant deux inconvénients majeurs. Le premier est qu'elle est lente et coûteuse. La seconde est que, comme le font clairement apparaître ces règles, la classification est largement subjective. On a donc des problèmes sérieux d'accord inter-annotateurs et de reproductibilité. À titre d'exemple, lors d'une expérience menée dans notre équipe au cours de laquelle des volontaires devaient annoter une centaine de phrases, le taux de mots corrects variait entre 50% et 85% selon les annotateurs, alors que les phrases provenaient du même corpus. Nous avons ensuite fait une nouvelle expérience en donnant non seulement des règles, mais également des exemples, et en réduisant le nombre d'annotateurs. Nous avons obtenu des annotations plus cohérentes, mais le problème subsistait.

Nous avons ainsi généré un petit corpus de 150 phrases annotées manuellement. Nous avons pour ce faire utilisé le système de traduction décrit comme référence de la campagne WMT08 [16] : il repose sur un modèle de langage 5-gramme entraîné avec SRILM [Stol 02] sur un corpus d'environ 35 millions de mots, un modèle de traduction IBM-5 entraîné avec Giza++ [Och 03] sur un bitexte d'environ 40 millions de mots (corpus EUROPARL [Koeh 05]), et le décodeur Moses [Koeh 07a]. Un corpus de quarante mille paires de phrases est mis de côté et sera utilisé pour l'entraînement des mesures de confiance.

Nous avons traduit 150 phrases extraites de transcriptions d'émissions radiophoniques à l'aide du système ainsi entraîné. Le style spontané et le vocabulaire ne correspondant pas tout à fait au corpus d'apprentissage du système, le score BLEU est assez faible (21.8, avec une seule référence. Le taux d'erreur — WER, partie I, section 1.6.2 — est lui d'environ 16%). Cependant, la plupart des traductions était, au minimum, compréhensible en y réfléchissant un peu. Voici quelques exemples d'annotations (les mots incorrects sont soulignés) :

Phrase source	*Traduction automatique*	*Classe de la phrase*
je vous remercie monsieur le commis--saire pour votre déclaration.	thank you mr commissioner for your question.	incorrecte
j'ai de nombreuses questions à poser à m. le commissaire.	i have some questions to ask to the commissioner.	correcte
les objectifs de la stratégie de lisbonne ne sont pas les bons.	the lisboa strategy mistaken.	incorrecte

2.2 Annotation automatique

Les traductions proposées sont cette fois comparées aux traductions de référence. Un mot est considéré comme correct s'il est aligné (au sens de Levenshtein) avec un mot de la référence [Ueff 04]. Lorsqu'une seule traduction de référence est utilisée, cet algorithme identifie indûment comme erronés beaucoup de mots corrects ; en effet, il existe souvent de nombreuses façons de traduire une phrase, qui peuvent n'avoir aucun mot commun. Si le système de traduction en propose une correcte, mais différente de la traduction de référence, elles sera considérée à tort comme erronée. On peut compenser en partie cet inconvénient en utilisant plusieurs traductions de référence (c'est qui a été fait lors de l'atelier [Blat 03]), mais la production de ces références supplémentaires est lente et coûteuse, et même ainsi, il reste des doutes sur la qualité de l'annotation.

2.3 Corpus Artificiel

Nous allons présenter dans cette section un algorithme réalisant la synthèse (presque) idéale des deux méthodes présentées ci-dessus. La méthode consiste à générer automatiquement des traductions dans lesquelles les erreurs sont connues. On a ainsi la possibilité de générer à faible coût (car sans intervention humaine hormis l'utilisation d'une traduction de référence) un corpus assez grand, avec une annotation quasiment parfaite, car c'est l'algorithme qui insère les erreurs, donc leurs positions sont connues de façon exacte. Nous pourrons donc entraîner les classifieurs sur des corpora de bonne taille. Cette méthode cependant ne permet pas d'obtenir un corpus d'entraînement automatiquement annoté pour l'estimation de correction des phrases : en effet, contrôler la position des erreurs insérées ne suffit pas à déterminer si la phrase modifiée est toujours une traduction acceptable ou non si on ne maîtrise pas la nature des erreurs. Par

16. http ://statmt.org/wmt08/baseline.html

exemple, si un adjectif est supprimé, la phrase peut toujours être une traduction acceptable. En revanche, si on supprime une négation, elle devient erronée. Au lieu d'utiliser cette méthode pour produire une classification de référence des phrases, nous allons l'utiliser pour produire une *estimation de qualité* de référence, en calculant le score BLEU (partie I, section 1.6.2) des phrases dégradées. Cette métrique est perfectible, mais elle a l'avantage d'être bien connue de la communauté, et les scores produits seront facilement interprétables. La méthode est décrite dans la section 3.2.6.

Le principe de l'algorithme est simple : il insère des erreurs dans les traductions de référence, de façon à approcher les erreurs réellement commises par un système de traduction. Étant donnée une phrase $\mathbf{t}$ dans la langue cible, qui est une traduction correcte de la phrase source $\mathbf{s}$, on choisit d'abord où insérer les erreurs. Les erreurs commises par un véritable système de traduction ont tendance à arriver « en grappe » (une erreur en entraînant une autre, etc.) et non à être distribuées uniformément dans la phrase. Nous avons donc implémenté et testé deux modèles de distribution d'erreur pour recréer ces grappes. Les paramètres de ces modèles ont été estimés sur de petits corpora de phrases annotées manuellement (50 paires de phrases). Ces modèles sont décrits dans les sections suivantes.

2.3.1 Modèle d'Erreurs Bigramme

Nous implémentons un simple modèle bigramme $P(C_i|C_{i-1})$ qui, étant donnée la correction (la vraie classe) d'un mot, estime la probabilité de correction du mot suivant. Ce modèle est basé sur l'idée qu'un mot erroné tend à en engendrer un autre (notamment à cause du modèle n-gramme utilisé lors du décodage). Le premier mot de la phrase a une probabilité a priori $P(C_0 = 1)$ d'être correct. Ce modèle est donc contrôlé par trois paramètres :

$$\begin{aligned} p_{start} &\stackrel{def}{=} P(C_0 = 1) \\ p_{incorrect|incorrect} &\stackrel{def}{=} P(C_i = 0|C_{i-1} = 0) \\ p_{correct|correct} &\stackrel{def}{=} P(C_i = 1|C_{i-1} = 1) \end{aligned}$$

En utilisant un petit corpus manuellement annoté, nous avons déterminé qu'environ neuf phrases sur dix commencent par un mot correct ($p_{start} = 0.9$), qu'un mot correct a environ 90% de chances d'être suivi d'un autre ($p_{correct|correct} = 0.9$) et qu'un mot incorrect avait 50% de chance d'être suivi d'un autre mot incorrect ($p_{incorrect|incorrect} = 0.5$).

2.3.2 Modèle d'Erreurs en Grappes

Cette seconde approche modélise explicitement les grappes d'erreur. On part du principe qu'une phrase est une suite de groupes de mots corrects et de groupes de mots erronés. La séquence des annotations est de la forme : $\mathbf{C}_1, .., \mathbf{C}_n$, où les $\mathbf{C}_.$ sont des séquences d'annotations identiques. Par définition, un groupe ne contient que des mots corrects ou que des mots incorrects (on parlera de « groupe correct » et « groupe incorrect ») ; si un groupe est correct, le suivant est incorrect, et inversement. Soit $\bar{c}_i \in \{0, 1\}$ la correction des mots du i-ème groupe. Le modèle est contrôlé par les paramètres suivants :

$P(\bar{c}_0 = 1)$	probabilité que le premier groupe soit correct
$P(Len(\mathbf{C}_.) = l \vert \bar{c}_. = 1)$ pour $l \in \mathbb{N}^*$	distribution des longueurs des clusters corrects
$P(Len(\mathbf{C}_.) = l \vert \bar{c}_. = 0)$ pour $l \in \mathbb{N}^*$	distribution des longueurs des clusters incorrects

On voit qu'en théorie, ce modèle est contrôlé par une infinité de paramètres. En pratique, l est borné par la longueur maximale des phrases de notre corpus. Ces paramètres sont estimés sur cinquante paires de phrases annotées manuellement. Nous n'allons pas les énumérer tous, mais pour donner une idée, nous avons déterminé que la longueur moyenne d'un groupe de mots corrects est 12 ($\sum_{k \geqslant 1} k \times P(Len(\mathbf{C}_.) = k \vert \bar{c}_. = 1) = 12$) celle d'un groupe incorrect est 2.

Une fois que les groupes (c'est-à-dire les positions des erreurs) sont générés, on insère les erreurs elles-mêmes. Elles sont de quatre types :

- déplacement
- erreur de grammaire
- substitution
- insertion

Les erreurs de type « déplacement » sont simples : le mot choisi est déplacé à une distance choisie aléatoirement (entre 1 et 4 dans nos expériences). Les « erreurs de grammaire » sont générées en modifiant la terminaison du mot (par exemple, « many bikes » peut devenir « many bike »), en s'assurant à l'aide de WordNet [Mill 95] que le mot modifié est bien un mot du vocabulaire. Les « substitutions » et les « insertions » sont un peu plus complexes. Étant donnée la position du mot à insérer ou remplacer, la probabilité de chaque mot du vocabulaire est calculée selon un modèle de traduction IBM-1 et un modèle de langage pentagramme :

$$\forall t' \in \mathcal{V}_T \ . \ p(t') = p_{IBM-1}(t' \vert \mathbf{s}) \times p_{5-gram}(t' \vert t_{i-4}, .., t_{i-1})$$

Le nouveau mot t' est ensuite choisi au hasard selon la distribution obtenue, en s'assurant à l'aide de WordNet qu'on ne remplace pas un mot par un synonyme (sinon, on n'est pas assuré qu'il s'agisse d'une erreur). On obtient ainsi des erreurs réalistes. Cet algorithme est contrôlé par les probabilités des différentes erreurs, qui ont été choisies « à la main » :

- Taux de déplacements P_d
- Taux de substitutions P_r
- Taux d'insertions P_i
- Taux d'erreurs grammaticales P_g

Nous obtenons finalement un corpus contenant environ 16% de mots erronés, ce qui correspond approximativement au taux d'erreur des traductions proposées par notre système de traduction automatique. Nous présentons ci-dessous un exemple de phrase générée par cet algorithme à partir d'une traduction correcte :

phrase source	Quant à eux, les instruments politiques doivent s'adapter à ces objectifs.
traduction de référence	Policy instruments, for their part, need to adapt to these goals.
traduction modifiée	Policy instruments, for the part, must to adapt to these goals.

Nous avons modifié ainsi quarante mille phrases extraites du corpus EUROPARL utilisé pour la campagne WMT 2008. Nous avons observé que le **modèle bigramme** donnait de meilleurs résultats, dans le sens où les classifieurs entraînés sur le corpus modifié en suivant ce modèle

tendent à donner une meilleure classification au sens de la courbe DET. Sans doute le nombre réduit de paramètres le rend-il plus robuste à la petite taille du corpus d'entraînement — 50 paires de phrases. C'est donc ce modèle que nous avons utilisé pour toutes nos expériences. Le corpus obtenu sera appelé $\mathcal{A}$ (pour « Artificiel »).

Le score BLEU du corpus modifié était de 56.5. C'est un score bien plus élevé que celui du corpus de test $\mathcal{T}$ traduit par le système basé sur Moses (21.8). Mais cette différence ne signifie pas que les phrases sont de meilleure qualité dans le corpus artificiel : en effet, la qualité de la traduction de $\mathcal{T}$ est sous-estimée car son score BLEU est calculé avec une seule référence, et ne prend donc pas du tout en compte la variabilité des traductions possibles ; or, $\mathcal{A}$ étant généré à partir des traductions de référence, la variabilité des traductions n'entre pas en compte : d'une part, on sait que toute déviation par rapport à la référence est une erreur, par construction ; d'autre part, le score BLEU, qui repose sur des cooccurrences de n-grammes, est naturellement haut puisque la référence et la référence dégradée restent similaires.

L'utilisation d'un corpus artificiel n'est pas une nouveauté. Mais d'habitude, c'est plutôt l'annotation qui est automatique, et non la phrase, comme décrit dans la section 2.2. Les auteurs de [Quir 04], par exemple, obtiennent de meilleurs résultats en entraînant leurs classifieurs sur un petit corpus annoté manuellement que sur un grand corpus annoté automatiquement (en comparant les hypothèses de traduction avec une traduction de référence). Nos conclusions sont opposées : les performances sont légèrement meilleures en entraînant nos classifieurs sur un corpus de dix milles phrases générées automatiquement que sur cent cinquante phrases traduites par Moses et annotées manuellement (section 3.1.8, figure 3.5). Ces conclusions s'expliquent sans doute par le fait qu'en utilisant un corpus généré automatiquement, l'annotation obtenue ne contient pas ou peu d'erreurs, contrairement à ce qui se passe quand le corpus est réel et l'annotation automatique. En revanche, le corpus automatiquement généré est moins réaliste que le corpus $\mathcal{T}$, ce qui explique que l'amélioration ne soit pas énorme. Mais si cette amélioration est modeste, il faut tout de même noter que le coût d'obtention de $\mathcal{T}$ est bien plus élevé que celui de $\mathcal{A}$ (une semaine environ, en comptant l'élaboration des instructions et le temps mis par les volontaires pour s'atteler à la tâche et nous renvoyer les phrases annotées, contre quelques heures).

3

Expériences autour des mesures de confiance

On peut utiliser certains paramètres prédictifs seuls (c'est-à-dire que la représentation numérique du mot ou de la phrase est de dimension 1), par exemple ceux basés sur les n-grammes. Nous les appelons alors « mesures de confiance » (on confond le paramètre prédictif et le classifieur), car ils reflètent directement la qualité ou la probabilité de correction de la traduction, sans qu'il soit besoin d'utiliser un algorithme de classification ou de régression. La classification se fait alors simplement en choisissant un seuil, et il est simple d'en évaluer les performances en termes de taux d'erreur ou de courbe DET (*Discrimination Error Tradeoff*, section 1.5.2). Pour calculer la NMI (*Normalised Mutual Information*, section 1.5.2) associée, en revanche, il faut transformer cette mesure en une véritable probabilité. Pour cela, nous utiliserons la régression logistique (section 1.4.1) en optimisant les paramètres sur le corpus artificiel.

Si on veut combiner le pouvoir prédictif de plusieurs paramètres, en revanche, la procédure est plus compliquée. Nous procéderons comme suit (selon le classifieur utilisé, certaines étapes peuvent être ignorées ; nous n'en ferons pas systématiquement mention) :

Nous générons d'abord un corpus artificiel $\mathcal{A}$ de quarante mille paires de phrases. Il est scindé en plusieurs corpora : un premier corpus d'entraînement $\mathcal{A}^{T_1}$ et un corpus de développement $\mathcal{A}^D$ sont utilisés pour déterminer, par grid search, les meilleurs méta-paramètres au regard du taux d'égale erreur (taux d'apprentissage, connectivité et nombre de neurones cachés pour les réseaux de neurones, γ et C pour les SVM — voir section 1.4). Une fois ces méta-paramètres optimisés, on apprend à nouveau le modèle sur un autre corpus d'entraînement $\mathcal{A}^{T_2}$. Si nécessaire, le biais est estimé sur un corpus $\mathcal{A}^B$. Enfin, le modèle est testé sur un corpus $\mathcal{U}$ constitué, cette fois, de véritables traductions automatiques manuellement annotées. $\mathcal{A}^{T_1}$, $\mathcal{A}^{T_2}$, $\mathcal{A}^D$, et $\mathcal{A}^B$ comprennent chacun dix mille paires de phrases (environ deux cent mille mots). $\mathcal{U}$ comprend cent cinquante paires de phrases soit environ trois mille mots (section 2.1). Dans le reste de ce chapitre, les expériences seront menées dans le cadre d'une traduction du français vers l'anglais de phrases extraites du corpus EUROPARL. Tous les résultats sont résumés en annexe A (page 125).

3.1 Mesures de confiance pour les mots

Nous allons maintenant passer en revue les différents paramètres prédictifs (c'est-à-dire les composantes du vecteur $\mathbf{x}(\mathbf{S}, \mathbf{T}, j)$) que nous avons utilisés pour estimer les probabilités de correction des mots. Ces composantes seront notées x_{index}, avec *index* l'indice de l'équation, afin qu'il soit plus facile de s'y référer par la suite. Considérées ensemble, ces composantes sont

une représentation du mot de la langue cible t_j, de son contexte (la phrase $\mathbf{t}$ dans son ensemble) et de la phrase source $\mathbf{s}$. Bien sûr, cette représentation n'est pas aussi expressive que le véritable triplet $(\mathbf{s}, \mathbf{t}, j)$, mais elle est plus accessible aux techniques d'apprentissage numérique, tout en conservant suffisamment d'information pour permettre une estimation efficace de la correction d'un mot.

Certaines de ces composantes (par exemple celles basées sur les modèles n-grammes) peuvent être utilisées seules (le vecteur $\mathbf{x}$ n'a alors qu'une seule composante). Ces paramètres prédictifs sont alors appelés des mesures de confiance, et nous présentons leurs performances. Pour d'autres (comme le PoS ou les caractéristiques de la phrase source), cela n'a pas de sens ou est très inefficace, et nous omettrons l'évaluation des performances.

3.1.1 Mesures basées sur les n-grammes

Les probabilités estimées par un modèle n-gramme et les replis nécessaires au calcul de ces probabilités fournissent des informations utiles à l'estimation de la correction des mots.

On peut tout d'abord utiliser la probabilité n-gramme classique :

$$x_{3.1}(\mathbf{S}, \mathbf{T}, j) = P(t_j | t_{j-1}, .., t_{j-n+1}) \tag{3.1}$$

Intuitivement, on s'attend à ce qu'un mot erroné ait une probabilité plus basse qu'un mot correct. Cependant, un modèle n-gramme étant déjà utilisé lors du décodage, même un mot erroné risque d'avoir une probabilité élevée. Afin de compenser ce phénomène, on peut aussi utiliser un modèle à rebours, comme proposé dans [Duch 02]. Les probabilités calculées sont ainsi moins corrélées avec les scores déjà utilisés lors du décodage :

$$x_{3.2}(\mathbf{S}, \mathbf{T}, j) = P(t_j | t_{j+1}, .., t_{j+n-1}) \tag{3.2}$$

Mais en pratique, la figure 3.1 et le tableau 3.1 montrent que les résultats de cette mesure de confiance sont moins bons que ceux des n-grammes « à l'endroit ». Enfin, on utilise les replis qu'un modèle de langage trigramme effectue pour l'estimation des probabilités évoquées. En effet, un n-gramme absent du modèle de langage peut dans certains cas être le signe d'une erreur de traduction. Cette mesure de confiance attribue donc un score en fonction des replis que le modèle a dû effectuer, comme proposé pour la reconnaissance automatique de la parole dans [Uhri 97] :

$$x_{3.3}(\mathbf{S}, \mathbf{T}, j) = \begin{cases} 1.0 \text{ si } t_{j-2}, t_{j-1}, t_j \text{ est présent dans le modèle} \\ 0.8 \text{ si } t_{j-2}, t_{j-1} \text{ et } t_{j-1}, t_j \text{ sont présents, mais pas le trigramme} \\ 0.6 \text{ si seul } t_{j-1}, t_j \text{ est présent} \\ 0.4 \text{ si seul } t_{j-2}, t_{j-1} \text{ et } t_j \text{ existent} \\ 0.3 \text{ si } t_{j-1} \text{ et } t_j \text{ sont présents séparément (comme unigrammes)} \\ 0.2 \text{ si seul } t_j \text{ est présent} \\ 0.1 \text{ si } t_j \text{ est un mot inconnu} \end{cases} \tag{3.3}$$

La figure 3.1 présente la courbe DET de classifieurs basés sur ces mesures de confiance utilisées isolément. Nous présentons ici les performances pour des modèles trigrammes, mais nous utiliserons également des modèles pentagrammes lors de la combinaison de plusieurs mesures de confiance. On observe que les mesures basées sur le repli sont nettement meilleures que celles

basées sur les probabilités trigrammes, classiques ou à rebours, qui sont presque indiscernables. Cela s'explique peut-être par le fait que les mesures basées sur le repli sont moins corrélées que les probabilités trigrammes avec les scores utilisés pour le décodage. Les résultats numériques sont repris dans le tableau 3.1.

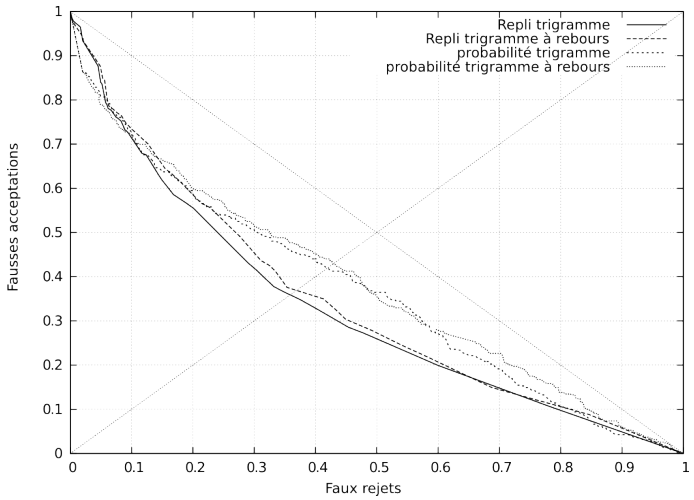


FIGURE 3.1 – Courbes DET des classifieurs utilisant les paramètres basés sur les trigrammes.

Mesure	Taux d'égale erreur	Information Mutuelle Normalisée
trigramme	42.1	4.86×10^{-3}
trigramme à rebours	42.9	-3.93×10^{-3}
repli	37.0	6.11×10^{-2}
repli à rebours	38.1	1.09×10^{-2}

TABLE 3.1 – Performances des mesures utilisant les trigrammes pour la détection des mots erronés.

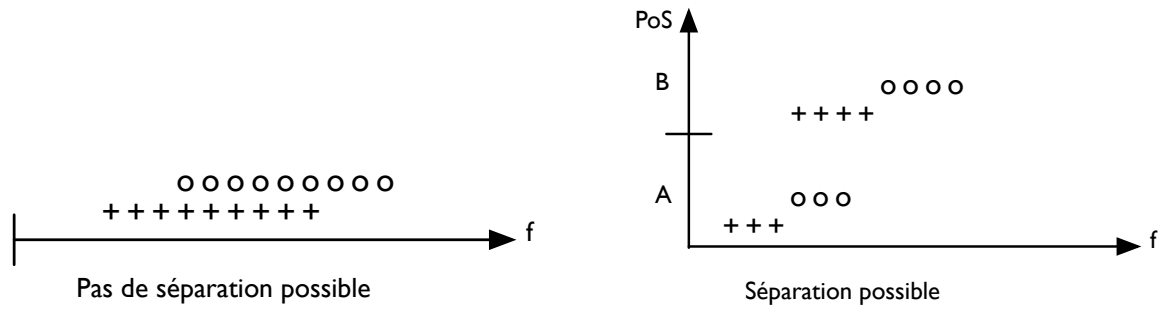
On observe que l'Information Mutuelle Normalisée du score trigramme à rebours est négative. Bien que cela ne soit théoriquement pas possible, on peut l'expliquer par les approximations effectuées lors du calcul de la NMI et par un biais dans les estimations de probabilités qui serait imparfaitement corrigé.

3.1.2 Utilisation des Parties du Discours

Nous utiliserons les Parties du Discours syntaxiques (en anglais *Part of Speech, PoS*) de deux façons. D'une part, comme un paramètre prédictif simple : s'il n'apporte directement aucune information quant à la correction d'un mot, il permet de prendre en compte le fait que d'autres paramètres prédictifs n'ont pas les mêmes distributions sur les différentes PoS. Ainsi, l'ajout d'un paramètre, même s'il n'apporte pas directement d'information sur la correction d'un mot, peut permettre de rendre séparables deux classes qui ne le sont pas (figure 3.2).

$$x_{3.4}(\mathbf{S}, \mathbf{T}, j) \quad = \quad POS(t_j) \tag{3.4}$$

D'autre part, nous calculons le score trigramme des catégories PoS, afin de détecter d'éventuelles erreurs de grammaires :



(a) Deux classes non séparables par le paramètre f (b) Deux classes rendues séparables

FIGURE 3.2 – Comment l'ajout d'un paramètre peut rendre séparables deux classes

$$x_{3.5}(\mathbf{S}, \mathbf{T}, j) \quad = \quad P(POS(t_j)|POS(t_{j-2}), POS(t_{j-1})) \tag{3.5}$$

Les catégories PoS sont calculées automatiquement en utilisant GPoSTTL, une alternative libre à TreeTagger [Schm 94, Schm 95]. La catégorie PoS peut donc être vue comme une caractéristique à valeur dans un ensemble de 44 éléments, que nous noterons $\{\pi_1, .., \pi_{44}\}$. Selon la méthode suggérée dans [Hsu 03], afin d'être efficacement utilisées par les classifieurs, ces valeurs sont transformées en un vecteur binaire à 44 composantes :

$$\pi(t_j)[i] = \begin{cases} 1 & \text{si } POS(t_j) = \pi_i \\ 0 & \text{sinon} \end{cases}$$

Le paramètre $x_{3.4}$ est donc à valeur dans $\{0, 1\}^{44}$.

3.1.3 Prendre en compte les erreurs dans le contexte

Toutes les mesures basées sur les n-grammes sont sensibles aux erreurs dans le contexte : supposons que l'on veuille évaluer le mot t_j ; si un mot est faux dans son contexte gauche (ou droit, pour les n-grammes à rebours), t_j risque d'avoir une probabilité n-gramme basse, et un mauvais score de repli, même s'il est correct. Afin de prendre en compte ce phénomène, pour chaque mesure basée sur les n-grammes, nous calculerons aussi le score moyen sur le voisinage de chaque mot. Intuitivement un score bas sur le voisinage peut indiquer que la probabilité n-gramme du mot considéré est mal estimée. Par exemple, supposons que la traduction correcte soit :

$$a\ b\ c$$

mais le système propose :

$$a\ Z\ c$$

supposons, par exemple, que le bigramme $Z\ c$ n'existe pas : dans ce cas, la probabilité n-gramme $P(c|Z)$ sera très basse. Et pourtant, c est correct. En revanche, $P(Z|a)$ risque aussi d'être très basse, puisque Z est erroné. Donc, si on ajoute l'information qu'un ou plusieurs mots du contexte

gauche de c a une probabilité n-gramme basse (en calculant la valeur $P(a) \times P(Z|a)$), le classifieur peut distinguer le cas où un mot a une probabilité basse parce qu'il est erroné, et le cas où il a une probabilité basse parce qu'un mot de son contexte gauche est erroné.

Ainsi, pour chaque mesure concernée $x_.$ définie dans la section 3.1.1 ainsi que pour la mesure n-gramme PoS $x_{3.5}$, on calcule aussi :

$$\begin{aligned} x_.^{gauche}(\mathbf{S},\mathbf{T},j) &= x_.(\mathbf{S},\mathbf{T},j-2) * x_.(\mathbf{S},\mathbf{T},j-1) * x_.(\mathbf{S},\mathbf{T},j) \\ x_.^{centre}(\mathbf{S},\mathbf{T},j) &= x_.(\mathbf{S},\mathbf{T},j-1) * x_.(\mathbf{S},\mathbf{T},j) * x_.(\mathbf{S},\mathbf{T},j+1) \\ x_.^{droite}(\mathbf{S},\mathbf{T},j) &= x_.(\mathbf{S},\mathbf{T},j) * x_.(\mathbf{S},\mathbf{T},j+1) * x_.(\mathbf{S},\mathbf{T},j+2) \end{aligned}$$

Ces paramètres sont ensuite combinés avec les paramètres « sans contexte ». La figure 3.3 et le tableau 3.2 (obtenus en utilisant un réseau de neurones) montrent une amélioration notable des performances (significative pour $p_{value} = 0.01$).

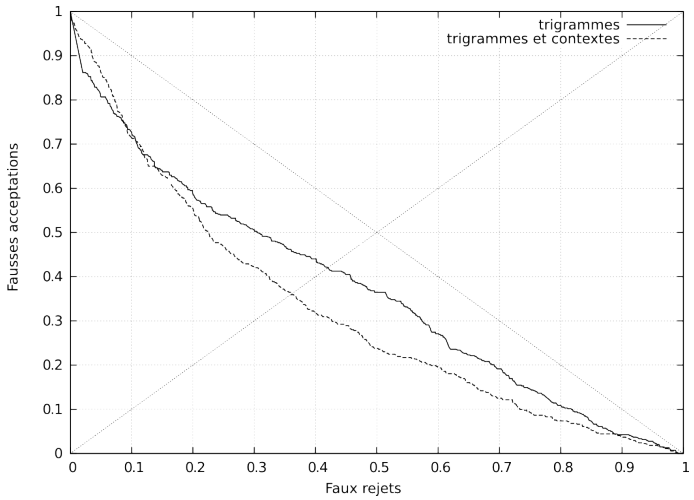


FIGURE 3.3 – Courbes DET des mesures niveau mot basées sur les trigrammes, avec et sans prise en compte des erreurs dans le contexte.

	Taux d'Égale Erreur	Information Mutuelle Normalisée
trigrammes	42.1	4.86×10^{-3}
trigrammes et contexte	36.3 **(-5.7)**	4.57×10^{-3} (-0.29×10^{-3})

TABLE 3.2 – Influence des erreurs du contexte sur l'estimation de confiance.

Si le taux d'égale erreur est largement amélioré, l'information mutuelle normalisée est légèrement dégradée. Nous expliquons cela par le fait que les probabilités n-grammes moyennées sur le contexte ne sont pas de véritables probabilités de correction d'un mot.

3.1.4 Information Mutuelle Intra-Langue

Nous avons présenté dans [Rayb 09a, Rayb 09b] des mesures de confiance originales utilisant *l'information mutuelle* entre les mots. L'information mutuelle I est une mesure de la quantité

d'information qu'une variable aléatoire fournit à propos d'une autre. L'idée d'utiliser l'information mutuelle pour calculer des mesures de confiance a été utilisée auparavant dans [Guo 04], et les auteurs de [Lave 07] l'utilisent pour construire des tables de traduction.

Nous considérons ici deux variables aléatoires dont les réalisations sont deux mots présents dans une phrase, W_1 et W_2 :

$$I(W_1, W_2) = \sum_{w_1, w_2} P(W_1 = w_1, W_2 = w_2) \times log\left(\frac{P(W_1 = w_1, W_2 = w_2)}{P(W_1 = w_1)P(W_2 = w_2)}\right)$$

$P(W = w)$ est ici la probabilité d'apparition du mot w dans une phrase, et $P(W_1 = w_1, W_2 = w_2)$ la probabilité de cooccurrence des mots w_1 et w_2 au sein de la même phrase. Ces distributions sont apprises sur le corpus EUROPARL. Nous nous intéresserons ici à l'information mutuelle mot-à-mot, c'est-à-dire chaque terme de la somme, qui peut être vue comme la contribution de chaque paire de mots à l'information mutuelle totale :

$$IMI(w_1, w_2) = P(W_1 = w_1, W_2 = w_2) log\left(\frac{P(W_1 = w_1, W_2 = w_2)}{P(W_1 = w_1)P(W_2 = w_2)}\right)$$

Le couple (w_1, w_2) est appelé *trigger* et $IMI(w_1, w_2)$ est son *score*. Intuitivement, un trigger avec un fort score indique des mots qui « ont tendance » à apparaître dans les mêmes phrases, c'est-à-dire que l'apparition de l'un déclenche (en anglais *triggers*) l'apparition de l'autre. Les triggers sont appris sur un bitexte aligné. La mesure de confiance (appelée *IMI* pour *Intra-lingual Mutual Information*) qui en découle est la moyenne, pour chaque mot de la traduction proposée, des informations mutuelles entre ce mot et les autres mots de la phrase :

$$x_{3.6}(\mathbf{s}, \mathbf{t}, j) = \frac{1}{Len(\mathbf{t}) - 1} \sum_{1 \leqslant i \neq j \leqslant Len(\mathbf{t})} IMI(t_i, t_j) \tag{3.6}$$

Cette mesure estime la cohérence de la présence d'un mot dans la traduction, car une information mutuelle plus élevée indique que des mots ont tendance à apparaître dans le même contexte. On estime souvent que les mots les plus éloignés dans la phrase sont moins liés entre eux. Les expériences visant à prendre en compte ce phénomène, en pondérant la moyenne par des coefficients décroissant avec la distance $|i - j|$ ou en ne calculant la moyenne que sur une fenêtre autour de j, ont cependant montré que cela n'améliorait pas les performances des mesures utilisant l'information mutuelle.

3.1.5 Information Mutuelle Inter-Langues

L'information Mutuelle Inter-Langues (*CMI* pour *Cross-lingual Mutual Information*) est similaire à la mesure IMI précédemment décrite, mais mesure la cohérence des mots formant la traduction avec les mots formant la phrase source. Une information mutuelle $CMI(s, t)$ élevée signifie cette fois que la présence du mot s dans une phrase source a tendance à déclencher la présence du mot t dans sa traduction. On a donc :

$$CMI(w_1, w_2) = P(W_1 = w_1, W_2 = w_2) log\left(\frac{P(W_1 = w_1, W_2 = w_2)}{P(W_1 = w_1)P(W_2 = w_2)}\right) \tag{3.7}$$

$$x_{3.8}(\mathbf{s}, \mathbf{t}, j) = \frac{1}{Len(\mathbf{s})} \sum_{1 \leqslant i \leqslant Len(\mathbf{s})} MI(s_i, t_j) \tag{3.8}$$

Avec $P(W_1 = w_1)$ la probabilité que w_1 apparaisse dans une phrase source, $P(W_2 = w_2)$ la probabilité que w_2 apparaisse dans une phrase traduite, et $P(W_1 = w_1, W_2 = w_2)$ la probabilité que w_1 apparaisse dans une phrase source et w_2 dans sa traduction.

Le tableau 3.3 présente les évaluations des performances des mesures basées sur l'information mutuelle. Bien que fort modeste, nous verrons qu'elles apportent tout de même une amélioration lorsqu'elles sont utilisées en combinaison avec d'autres (section 3.1.8).

Paramètre	Taux d'Égale Erreur	Information Mutuelle Normalisée
intra-langue (IMI)	45.8	9.46×10^{-4}
inter-langues (CMI)	45.7	-2.21×10^{-1}

TABLE 3.3 – Performances des mesures de confiance niveau mots basées sur l'information mutuelle.

3.1.6 Modèle IBM-1

L'utilisation des probabilités données par un modèle IBM-1 comme mesure de confiance a été proposée dans [Blat 03] et [Ueff 05]. Cette mesure s'écrit :

$$x_{3.9}(\mathbf{s}, \mathbf{t}, j) = \frac{1}{Len(\mathbf{s}) + 1} \sum_{i=0}^{Len(\mathbf{s})} p_{IBM-1}(t_j | s_i) \tag{3.9}$$

Les résultats (tableau 3.4) sont là encore assez décevants. Les auteurs de [Ueff 05] suggèrent d'approcher la somme par un maximum, mais cela n'a rien changé au résultat sur cette tâche. Il est surprenant de remarquer que sur cette tâche d'évaluation de traduction, aucune mesure impliquant la phrase source ne donne de meilleurs résultats que les mesures utilisant uniquement les n-grammes de l'hypothèse de traduction.

Mesure	Taux d'Égale Erreur	Information Mutuelle Normalisée
IBM-1	45.0	-1.84×10^{-3}

TABLE 3.4 – Performances de la mesure de confiance niveau mot basée sur un modèle IBM-1.

3.1.7 Indicateurs à base de règles

Nous proposons cinq paramètres prédictifs binaires supplémentaires utilisant des règles simples :

$$x_{3.10}(\mathbf{S}, \mathbf{T}, j) \quad = \quad \begin{cases} 1 & \text{si } t_j \text{ est un mot outil} \\ 0 & \text{sinon} \end{cases} \tag{3.10}$$

sert à identifier les mots outils, c'est-à-dire, en simplifiant, les mots qui ne désignent ni un objet, ni une personne, et ne sont ni des verbes, ni des adverbes ni des adjectifs (*the*, *it*, etc.). Ce paramètre est strictement moins informatif que les PoS, mais il est plus simple, et le nombre réduit de classes le rend plus robuste au manque de données d'apprentissage. Il pourrait donc donner de meilleurs résultats. La liste des mots outils est obtenue par une heuristique simple : il s'agit des mots courts (trois lettres ou moins) les plus fréquents.

Enfin, nous implémentons quatre indicateurs binaires indiquant si un « mot » est un signe de ponctuation, un nombre, une URL ou un nom propre (en utilisant une liste de noms propres).

Ces indicateurs permettent de mieux prendre en compte des distributions de probabilité différentes selon les catégories.

3.1.8 Classification et régression multivariées des mesures de confiance niveau mots

Nous avons développé au total 66 paramètres prédictifs. Beaucoup d'entre eux sont très corrélés (par exemple, les probabilités n-gramme et et les probabilités n-grammes moyennées sur le contexte). Certains sont très simples (les indicateurs à base de règles — section 3.1.7), d'autres plus sophistiqués (modèle IBM-1 — section 3.1.6). Nous allons utiliser cette représentation numérique des mots pour assigner une probabilité de correction à chacun à l'aide de différents classifieurs (Section 1.4). Nos expériences utiliseront la Régression Logistique, les Réseaux de Neurones, les Machines à Vecteurs de Support (SVM) et la Régression des Moindres Carrés Partiels (PLSR). Seuls les Réseaux de Neurones ont permis d'obtenir une meilleure estimation de probabilité, en combinant les différents paramètres, que la meilleure mesure de confiance utilisée seule (le repli trigramme — section 3.1.1). Cette amélioration reste toutefois modeste (1.3 points de Taux d'Égale Erreur). La structure du réseau est un perceptron à trois couches avec un taux de connectivité de 0.5, ayant 66 neurones d'entrée (un par paramètre), 33 neurones cachés et un neurone de sortie. La fonction d'activation est une sigmoïde. La courbe DET de ce classifieur est présentée dans la figure 3.4 et les résultats de tous les autres dans le tableau 3.5.

Classifieur	Taux d'Égale Erreur	NMI	Entraînement	Test
Régression Logistique	36.8	-2.61×10^{-2}	13”	5”
PLSR	37.5	-5.84×10^{-2}	15’	1”
SVM	36.7	-1.87×10^{-1}	12h	500”
Réseau de Neurones	35.0	6.06×10^{-2}	10’	2”

TABLE 3.5 – Classification des mots utilisant tous les paramètres

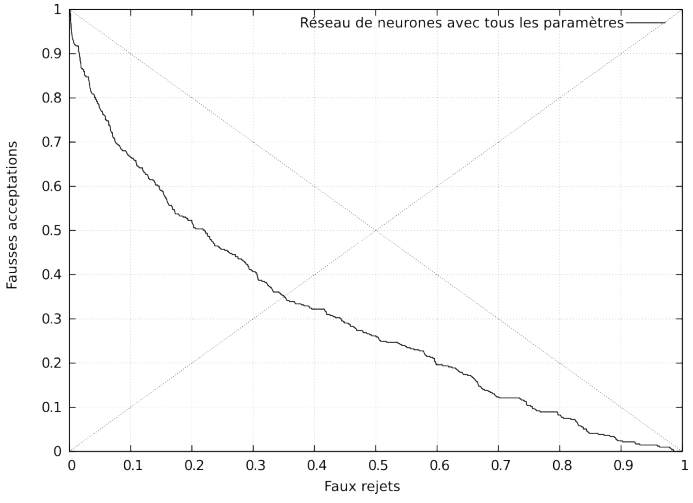


FIGURE 3.4 – Classification par un réseau de neurones utilisant tous les paramètres

Évalués à l'aune de l'Information Mutuelle Normalisée (NMI), les résultats sont particulièrement décevants. Une mauvaise estimation du biais en est probablement la cause : comme expliqué dans la section 1.3.2, le biais pénalise la mesure NMI. Or, la méthode que nous implémentons pour le corriger estime le biais sur des données artificielles (faute de données annotées en quantité suffisante). Il est possible qu'il soit en fait différent sur le corpus de test (constitué de « vraies » traductions automatiques annotées) et que la correction du biais ne soit donc pas efficace.

Le tableau 3.6 présente la modeste, mais néanmoins significative à 1%, contribution des mesures de confiance originales basées sur l'information mutuelle (section 3.1.4)

Mesure	Taux d'Égale Erreur	Information Mutuelle Normalisée
Sans IMI et CMI	35.6	5.32×10^{-2}
Avec IMI et CMI	35.0	6.06×10^{-2}
amélioration	**-0.60**	$\mathbf{+7.4 \times 10^{-3}}$

TABLE 3.6 – Contribution des mesures utilisant l'information mutuelle.

Enfin, nous avons également voulu comparer les performances obtenues en entraînant le classifieur sur un corpus artificiel et sur un corpus annoté. Le corpus annoté dont nous disposons est de taille très réduite, ce qui réduit la significativité des résultats. Il faut cependant noter que bien que de taille réduite, il a été beaucoup plus coûteux à obtenir que le corpus artificiel, en termes de temps de travail. Nous avons scindé ce corpus en un corpus d'entraînement (70 phrases), un corpus de développement (30 phrases) et un corpus de test (50 phrases). Nous avons entraîné et optimisé le réseau de neurones sur les deux premiers, et l'avons testé sur le dernier. La courbe DET obtenue est présentée figure 3.5 et les résultats numériques dans le tableau 3.7. Les deux classifieurs ont un comportement très semblable, mais les données artificielles sont obtenues à bien moindre coût.

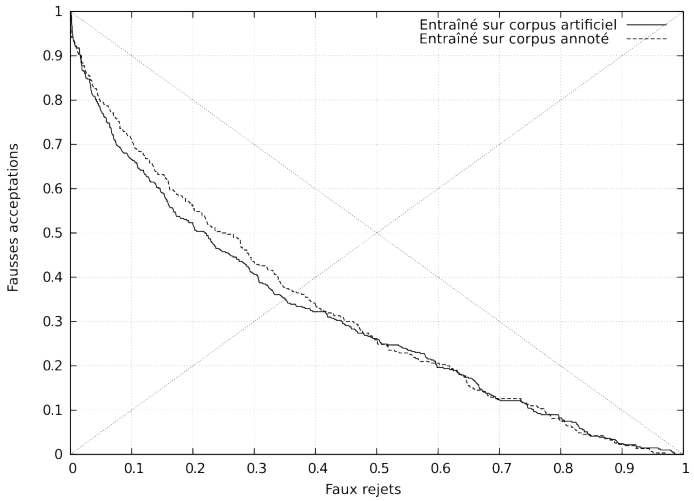


FIGURE 3.5 – Comparaison de classifieurs entraînés sur des données annotées ou artificielles.

Classifieur	Taux d'Égale Erreur	Information Mutuelle Normalisée
Réseau de neurone (corpus artificiel)	35.0	6.06×10^{-2}
Réseau de neurone (corpus annoté)	36.8	5.79×10^{-2}

TABLE 3.7 – Comparaison de classifieurs entraînés sur des données annotées ou artificielles.

3.2 Mesures de confiance pour les phrases

Les paramètres prédictifs présentés dans cette section ont pour but de générer une représentation numérique des phrases, afin d'utiliser les techniques de classification et de régression évoquées précédemment. Comme pour les mots, l'estimation automatique de qualité sera effectuée en estimant une distribution de probabilité sur un sous espace de $\mathbb{R}^{d_S}$ correspondant à l'ensemble des valeurs que peuvent prendre les paramètres prédictifs. La méthode que nous proposions pour générer un corpus d'apprentissage artificiel annoté pour la classification des mots ne fonctionne pas pour les phrases : en effet, on ne peut pas dire qu'un certain nombre d'erreurs rend une traduction incorrecte (cela dépend de la nature de l'erreur). Nous avons donc utilisé un corpus dégradé comme pour la génération d'un corpus d'apprentissage pour la classification des mots, mais au lieu de déterminer automatiquement des classes, nous avons utilisé la métrique BLEU pour estimer la qualité des phrases dégradées. Si le besoin s'en fait sentir, une classification est ensuite effectuée en choisissant un seuil sur cette métrique (voir la section 3.2.6).

Beaucoup des paramètres utilisés pour les mots peuvent être adaptés de façon simple pour les phrases. Je ne m'en priverai donc pas, et passerai rapidement sur les paramètres prédictifs qui sont des adaptations des paramètres utilisés pour les mots.

3.2.1 Mesures utilisant un modèle de langage

Je commencerai par les paramètres utilisant les n-grammes. Il s'agit de la *perplexité*, de la *perplexité selon un modèle à rebours*, et de la moyenne du *repli* des mots de la phrase (une simple moyenne du paramètre $x_{3.3}$ utilisé pour les mots — section 3.1.1). Les vraisemblances sont normalisées afin que les phrases longues ne soient pas pénalisées :

$$x_{3.11}(\mathbf{s},\mathbf{t}) = \left(\prod_{j=1}^{Len(\mathbf{t})} P(t_j|t_{j-1},..,t_{j-n+1})\right)^{\frac{1}{Len(\mathbf{t})}} \tag{3.11}$$

$$x_{3.12}(\mathbf{s},\mathbf{t}) = \left(\prod_{j=1}^{Len(\mathbf{t})} P(t_j|t_{j+1},..,t_{j+n-1})\right)^{\frac{1}{Len(\mathbf{t})}} \tag{3.12}$$

$$x_{3.13}(\mathbf{s},\mathbf{t}) = \frac{1}{Len(\mathbf{t})} \sum_{j=1}^{Len(\mathbf{t})} x_{3.3}(\mathbf{S},\mathbf{T},j) \tag{3.13}$$

Chacun peut être utilisé comme une mesure de confiance à part entière, et les résultats sont présentés dans le tableau 3.8, et leurs courbes DET dans la figure 3.6 (page 64), à côté de celle de la mesure utilisant l'information mutuelle intra-langue, qui peut être vue comme une forme de modèle de langage. Les performances de la mesure utilisant le repli se distinguent nettement de celles des trois autres, qui sont très proches.

Mesures	Taux d'Égale Erreur	Information Mutuelle Normalisée
perplexité trigramme	41.7	4.02×10^{-3}
perplexité trigramme à rebours	41.3	3.97×10^{-3}
repli moyen	34.2	4.15×10^{-3}

TABLE 3.8 – Performances des mesures utilisant des modèles n-grammes.

Nous utiliserons aussi comme paramètre prédictif la perplexité de la phrase source. Intuitivement, plus la perplexité de la phrase source est faible, plus elle sera difficile à traduire. Bien sûr, ce paramètre ne peut être utilisé comme une mesure de confiance à lui tout seul.

$$x_{3.14}(\mathbf{s},\mathbf{t}) = \left(\prod_{i=1}^{Len(\mathbf{s})} P(s_i|s_{i-1},..,s_{i-n+1}) \right)^{\frac{1}{Len(\mathbf{s})}} \tag{3.14}$$

3.2.2 Mesures utilisant l'information mutuelle

Continuons avec les mesures utilisant l'information mutuelle intra- et inter-langues. Là encore, nous nous contenterons de faire la moyenne des mesures de confiance sur les mots :

$$\begin{aligned} x_{3.15}(\mathbf{s},\mathbf{t}) &= \frac{1}{Len(\mathbf{t}) \times (Len(\mathbf{t})-1)} \sum_{i=1}^{Len(\mathbf{t})} \sum_{1 \leqslant j \neq i \leqslant Len(\mathbf{t})} MI(t_i,t_j) \qquad (3.15) \\ &= \frac{1}{Len(\mathbf{t})} \sum_{j=1}^{Len(\mathbf{t})} x_{3.6}(\mathbf{s},\mathbf{t},j) \\ x_{3.16}(\mathbf{s},\mathbf{t}) &= \frac{1}{Len(\mathbf{s}) \times Len(\mathbf{t})} \sum_{i=1}^{Len(\mathbf{s})} \sum_{j=1}^{Len(\mathbf{t})} MI(s_i,t_j) \qquad (3.16) \\ &= \frac{1}{Len(\mathbf{t})} \sum_{j=1}^{Len(\mathbf{t})} x_{3.8}(\mathbf{s},\mathbf{t},j) \end{aligned}$$

Malheureusement, la mesure utilisant l'information mutuelle inter-langues donne ici des résultats encore moins bons qu'au niveau des mots, avec un taux d'égale erreur d'environ 46%. Les performances pour l'information mutuelle intra-langue sont, eux, raisonnablement proches de ceux des mesures de confiance utilisant les n-grammes (tableau 3.9 et figure 3.6).

Mesure	Taux d'Égale Erreur	Information Mutuelle Normalisée
IMI	39.0	9.46×10^{-4}
CMI	46.1	-0.30×10^{-4}

TABLE 3.9 – Mesure de confiance utilisant l'information mutuelle intra-langue

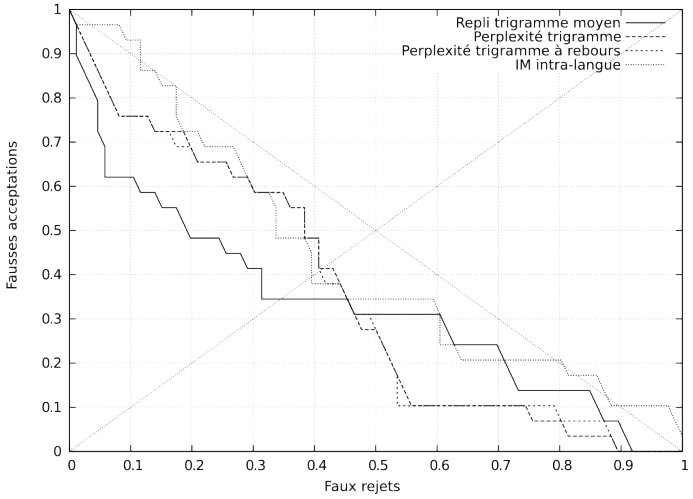


FIGURE 3.6 – Courbes DET des mesures de confiance utilisant des modèles de langage (n-gramme et IMI)

3.2.3 Mesure utilisant un modèle IBM-1

Nous utilisons la probabilité de traduction donnée par un modèle IBM-1 comme mesure de confiance, à ceci près qu'elle est normalisée afin de ne pas pénaliser les phrases longues :

$$x_{3.17}(\mathbf{s},\mathbf{t}) \quad = \quad P_{IBM-1}(\mathbf{s}|\mathbf{t})^{\frac{1}{Len(\mathbf{s})}} = \left(\prod_{i=1}^{Len(\mathbf{s})} \sum_{j=0}^{Len(\mathbf{t})} P(s_i|t_j) \right)^{\frac{1}{Len(\mathbf{s})}} \tag{3.17}$$

Là encore, cette mesure donne d'étonnamment mauvais résultats. Sans doute le modèle IBM-1 est-il trop rudimentaire pour cette application.

3.2.4 Vérification basique de la syntaxe

Le rôle de ce paramètre prédictif est de faire des vérifications élémentaires de syntaxe : les parenthèses vont-elles bien par paires, les points, points d'exclamation, d'interrogation, etc. sont-ils bien situés à la fin de la phrase, cette dernière s'achève-t-elle par un signe de ponctuation approprié, etc. On peut trouver une description plus complète dans [Blat 03].

$$x_{3.18}(\mathbf{s},\mathbf{t}) \quad = \quad \begin{cases} 1 & \text{si } \mathbf{t} \text{ est bien formée} \\ 0 & \text{sinon} \end{cases} \tag{3.18}$$

Ce paramètre, bien sûr, ainsi que ceux qui suivent, ne sont pas destinés à être des mesures de confiance à part entière.

3.2.5 Paramètre prédictifs utilisant la longueur

Ce paramètre proposé dans [Blat 03], très basique lui aussi, mesure la cohérence entre la longueur de la phrase source et la longueur de la traduction. Des longueurs très différentes indiquent peut-être une traduction erronée, du moins pour des langues ayant des notions de « mot » comparables, comme c'est le cas pour le français et l'anglais par exemple.

$$x_{3.19}(\mathbf{s},\mathbf{t}) = Len(\mathbf{s}) \quad (3.19)$$

$$x_{3.20}(\mathbf{s},\mathbf{t}) = Len(\mathbf{t}) \quad (3.20)$$

$$x_{3.21}(\mathbf{s},\mathbf{t}) = \frac{Len(\mathbf{t})}{Len(\mathbf{s})} \quad (3.21)$$

3.2.6 Classification et régression multivariées des mesures de confiance niveau phrase

Comme je l'ai expliqué précédemment, notre algorithme de génération de corpus d'apprentissage n'est pas adapté pour l'entraînement d'un classifieur au niveau des phrases, car on ne peut pas déterminer de façon automatique si une phrase générée reste compréhensible ou non. Au lieu d'une classification binaire des phrases du corpus généré, nous avons donc calculé leurs scores BLEU, au niveau de chaque phrase, en les comparant à la phrase d'origine. Nous avons ensuite utilisé les SVM, les Réseaux de Neurones et la PLSR pour effectuer une régression des paramètres prédictifs sur le score BLEU des phrases. La classification sera ensuite effectuée en choisissant un seuil pour ce score.

BLEU n'est certes pas une bonne métrique au niveau des phrases, et est mieux corrélé avec les jugements humain au niveau d'un document. Mais une partie des raisons qui en font une mauvaise métrique ne s'appliquent pas ici, puisque **la traduction générée est issue de la traduction de référence**. De plus, cette métrique a l'avantage d'être bien connue, et des outils efficaces existent pour la calculer. Nous avons également mené des expériences utilisant le *Translation Edit Rate*, sans succès, à cause de la grande variabilité des scores et d'un nombre élevé de phrases obtenant un score nul.

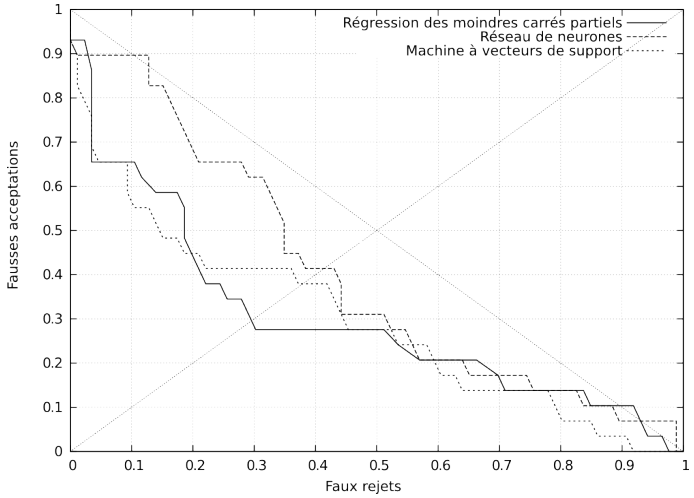


FIGURE 3.7 – Courbes DET des classifieurs niveaux phrases utilisant la Régression des Moindres Carrés Partiels, les Réseaux de Neurones ou les Machines à Vecteurs de Support.

Seule la PLSR permet une amélioration, de 5,2 points, par rapport à la meilleure mesure de confiance prise isolément (le repli moyen, Section 3.2.1). Le score ainsi obtenu présente une corrélation de 0,358 avec les annotations humaines (section 1.1).

Classifieur	Taux d'Égale Erreur	NMI
PLS	29.0	8.14×10^{-2}
SVM	38.0	-2.56×10^{-1}
Réseau de Neurones	41.3	-2.44×10^{-2}

TABLE 3.10 – Performances des classifieurs PLSR, SVM Réseaux de Neurones niveau phrase.

3.3 Conclusion

Dans les sections précédentes, je me suis employé à décrire une approche des mesures de confiance fondée sur des techniques de classification et de régression, à donner une description précise et formelle de paramètres prédictifs de l'état-de-l'art aussi bien que d'autres orginaux, et à présenter une évaluation des performances de différentes techniques évoquées. J'ai également proposé une méthode originale de génération automatique de données d'apprentissage qui permet, à faible coût, de générer de grands corpus annotés, facilitant l'estimation des paramètres des algorithmes de classification et de régression précédemment évoqué, avec de bons résultats, et notamment des résultats meilleurs qu'en utilisant de petits corpora manuellement annotés.

L'approche par classification et régression offre un cadre de travail robuste, fondé sur des algorithmes bien connus pour lesquels il existe des outils efficaces, et dans lequel il est aisé d'ajouter de nouvelles sources d'informations. L'utilisation de réseaux de neurones nous permet d'obtenir un taux d'égale erreur de classification de 35% pour la détection de mots erronés (soit 1,3 points d'amélioration par rapport au meilleur paramètre prédictif utilisé seul, en l'occurence le comportement de repli d'un modèle de langage trigramme), tout en étant peu coûteux en termes de temps de calcul, et l'utilisation de la régression des moindres carrés partiels nous permet d'otbenir un taux d'égale erreur de 29% pour la détection de phrases erronées (6 points d'amélioration par rapport au meilleur paramètre prédictif, là aussi le comportement de repli d'un modèle trigramme).

Les techniques présentées, cependant, ont certaines limitations. D'abord, tous les paramètres utilisés utilisent les formes de surface de mots, et aucun n'intègre de notion sémantique. Or, de petites erreurs de surface peuvent engendrer de grandes erreurs de sens. On pense par exemple à l'omission d'une négation. Ensuite, nous avons développé des paramètres prédictifs pour les phrases et les mots, mais aucun à un niveau intermédiaire, comme les groupes nominaux ou verbaux, ou plus simplement les séquences, comme le font certains modèles de traduction et notamment ceux que nous utilisons. Or, comme je l'ai expliqué précédemment, le mot est une unité trop petite, et sa correction trop ambigüe pour que cette approche soit pleinement satisfaisante, et à l'inverse, une phrase est une entité trop longue, qui peut être partiellement incorrecte sans qu'il soit nécessaire de la retraduire entièrement. L'annotation de confiance au niveau des segments reste une piste prometteuse et largement inexplorée. Enfin, comme nous le verrons dans le chapitre suivant (chapitre 4) qui présentera une expérience de post édition, la précision d'une mesure de confiance ne reflète pas son utilité : en effet, l'utilisateur ne s'intéresse pas au taux d'égale erreur, il veut que le système d'estimation de confiance lui fasse gagner du temps. Nous verrons qu'il est très sensible aux erreurs du système, et notamment aux faux rejets.

4

Une expérience de Post-Édition

Dans la section précédente, nous avons passé en revue de façon détaillée le principe, la définition mathématique et les performances d'un certain nombre de mesures de confiance. Nous savons donc estimer le nombre d'erreurs qu'on peut espérer détecter, les compromis faux positifs/faux négatifs réalisables, et l'information qu'elles nous apportent. Mais cela ne nous renseigne pas sur leur utilité réelle, dans le cadre d'une application de post-édition par exemple. Nous avons donc mis en place une telle expérience. L'objectif était d'obtenir un retour qualitatif de la part d'utilisateurs, afin de déterminer si le système que nous proposons, bien qu'imparfait, est utile en pratique ou non. Du fait du faible nombre d'utilisateurs sur lesquels porte notre étude, et du caractère expérimental de beaucoup des mesures de confiance utilisées, les résultats sont strictement qualitatifs, et ne doivent être interprétés que comme des indications des besoins réels des utilisateurs, et des pistes à suivre afin d'augmenter l'utilité des mesures de confiance dans un système de post-édition. Le protocole suivi est inspiré de celui décrit dans [Plit 10].

Nous avons implémenté un outil de post-édition utilisant des mesures de confiance pour indiquer à l'utilisateur les mots potentiellement erronés. Chaque utilisateur a ensuite eu pour tâche de corriger un certain nombre de traductions automatiques, parfois avec et parfois sans l'aide des mesures de confiance. Nous avons mesuré, entre autres, la vitesse de post-édition dans les deux cas.

4.1 L'outil de Post-Edition

Nous avons développé un système de post-édition simple (voir la capture d'écran figure 4.1). Il peut être vu comme une version très simplifiée d'un logiciel de traduction assistée par ordinateur. Il affiche une phrase source (dans notre cas, en français) et la traduction (en anglais) proposée par un logiciel de traduction automatique (l'outil Moses). Si la fonction est activée, les erreurs détectées par le système sont affichées en rouge. L'utilisateur peut ajuster le seuil de détection et modifier comme il le souhaite la traduction proposée, avant de valider. Pour les besoins de l'expérience, un fois une traduction validée, un utilisateur ne peut plus y revenir.

En haut de la fenêtre, dans la barre d'outil, on trouve les boutons permettant de charger les fichiers nécessaires à l'expérience. Ils étaient manipulés par l'expérimentateur et les volontaires n'avaient pas à s'en occuper, seulement à appuyer sur « Start ». La phrase source est affichée dans la champ du haut, et la traduction dans celui du bas, qui est éditable (à la différence du premier). À gauche on trouve un curseur permettant de modifier le seuil de détection (Section 1.4). Tous les mots dont le score est inférieur à celui choisi par l'utilisateur sont affichés en rouge. Les scores sont normalisés entre 0 et 1 : lorsque le curseur est tout en bas, le seuil est 0 (cela

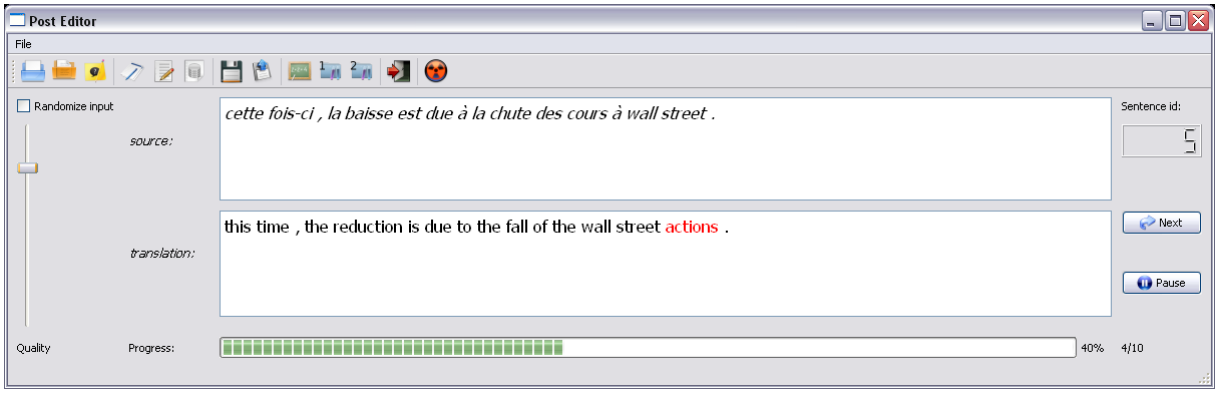


FIGURE 4.1 – Aperçu du logiciel de post-édition.

correspond au point d'abscisse 0 de la courbe DET : dans ce cas, tous les mots sont considérés comme corrects). Lorsqu'il est tout en haut, le seuil est 1 (c'est la point à l'extrême droite de la courbe DET : tous les mots sont considérés comme faux). L'utilisateur peut modifier la sévérité de la classification automatique, s'il le désire, en déplaçant ce curseur.

La traduction automatique et les mesures de confiance sont calculées hors-ligne, c'est-à-dire avant l'expérience, et non à la volée. La forte puissance de calcul requise par les outils de traduction et d'estimation de confiance rend nécessaire ce précalcul, sans quoi le programme serait beaucoup trop lent, et son déploiement sur des machines diverses beaucoup plus difficile. C'est un inconvénient majeur : en effet, le système ne peut donc pas prendre en compte les modifications apportées par l'utilisateur. Il ne peut ni estimer la correction de ses éditions, ni mettre à jour les estimations sur les parties non éditées de la phrase (qui peuvent avoir évolué, puisque la correction dépend du contexte).

On donne aux utilisateurs des explications superficielles sur le fonctionnement du système : non seulement un cours détaillé sur les mesures des confiance ne les intéresse pas, mais une connaissance trop poussée des détails de leur fonctionnement risquerait d'influer sur leur interaction avec le logiciel, invalidant les résultats de l'expérience [17]. Une fois que l'utilisateur a corrigé, s'il le souhaitait, la traduction automatique proposée, il peut cliquer sur « next » (ou « pause » s'il souhaite se reposer quelques instants — ce bouton a pour effet de ne pas afficher la nouvelle traduction à corriger et de suspendre le décompte du temps passé. En pratique, aucun des utilisateurs n'a fait de pause). Tout le reste de l'interface (la barre de progression, etc.) n'est que cosmétique.

Au cours de l'expérience on enregistre aussi le temps total passé à éditer les phrases (c'est-à-dire la durée séparant l'affichage d'une phrase et l'activation du bouton « next », moins le temps de pause). Ce temps est subdivisé en trois composantes : la saisie (le temps pendant lequel l'utilisateur presse des touches du clavier) ; la manipulation de l'interface (c'est-à-dire le temps que l'utilisateur passe à déplacer le curseur) ; et le temps de réflexion (tout le reste).

Les utilisateurs ont dans l'ensemble affirmé que le programme était facile à utiliser. Cela dit, il faudrait étudier l'ergonomie de l'interface, afin d'être certain que ce qui est mesuré est bien l'influence des mesures de confiance, et non celle de nos choix de conception.

17. Lors des tests, j'ai moi même observé que je tentais de comprendre pourquoi la mesure de confiance avait donné tel score à tel mot, et que cela avait une influence directe sur le jugement de correction que je portais sur celui-ci.

4.2 Protocole expérimental

Nous ne nous attendions pas (avec raison) à un afflux massif de volontaires. Nous avons donc voulu que leur niveau d'anglais soit aussi homogène que possible. Nous avons donc recruté parmi des enseignants d'anglais dont la langue maternelle est le français, et des étudiants de master d'anglais. Sur sept volontaires, deux n'ont pas suivi les consignes, et les données recueillies ont été ignorées. Chaque expérimentation durait environ deux heures, divisées en quatre étapes :

Première étape, introduction et entraînement : nous avons commencé par donner aux volontaires des explications rapides sur le domaine et sur le travail demandé. Nous leur avons ensuite donné dix traductions à corriger à l'aide de notre logiciel. Ces traductions n'étaient destinées qu'à les familiariser avec le logiciel et aucune mesure n'était réalisée à ce stade.

Seconde étape, première variante de l'expérience : lorsqu'ils se sentaient prêts, les volontaires pouvaient commencer l'expérience proprement dite. Trente paires d'une phrase source et d'une traduction automatique étaient affichées successivement. L'utilisateur pouvait modifier le seuil d'acceptation des mesures de confiance et éditer la traduction proposée.

Troisième étape, seconde variante de l'expérience : cette étape se déroule comme la précédente, sauf que les utilisateurs n'ont cette fois pas accès aux mesures de confiance. Pour la moitié des volontaires, l'ordre des étapes 2 et 3 est inversé, afin de compenser les effets opposés de la fatigue et de l'habitude : en effet, on peut s'attendre à ce que la réalisation du travail demandé soit plus facile lors de la seconde partie de l'expérience, puisque l'utilisateur a eu le temps de s'habituer, mais on peut aussi craindre qu'il soit fatigué et que cela ait un impact sur la vitesse et la qualité de son travail.

Quatrième étape, le recueil des impressions de l'utilisateur : pour terminer, les volontaires devaient remplir un questionnaire (annexe B, page 127) dont le but était de recueillir leurs impressions sur l'ergonomie du programme, la difficulté de la tâche, l'utilité et la qualité des mesures de confiance, etc.

Voici les instructions que nous avons données aux volontaires ; intuitivement, l'objectif était d'obtenir des traductions assez bonnes pour être compréhensibles sans trop d'efforts, mais pas forcément du « bel anglais » :

- Le but est d'obtenir une traduction correcte, pas une traduction parfaite : corrigez les erreurs, pas le style.
- Vous pouvez utiliser toute l'aide que vous souhaitez, tant que vous ne copiez pas sur vos voisins (l'idée étant d'éviter que certains ne perdent du temps sur des difficultés de langue, brouillant ainsi les mesures ; en pratique, tous ou presque ont utilisé des dictionnaires en ligne ou imprimés). Les seules contraintes sont :
  - N'utilisez pas d'outil de traduction automatique en ligne.
  - Ne perdez pas trop de temps sur des détails.
  - N'embêtez pas vos voisins.
  - Ne posez pas de questions à l'expérimentateur.

Les deux ensembles de trente phrases ont été tirés aléatoirement parmi les données de test de la campagne WMT 2009. Il s'agissait de transcriptions d'émissions radiodiffusées. Le caractère aléatoire de ces données permet d'éviter d'avoir un ensemble de traductions particulièrement facile ou au contraire trop difficile, faussant ainsi l'évaluation. Cela dit, ce choix est discutable

car en général, une tâche « réelle » consiste en la traduction de documents cohérents, et non d'un suite de phrases sans rapport les unes avec les autres.

4.3 Analyse des résultats

Les résultats les plus importants de cette expérience sont présentés dans le tableau 4.1. Les trois premières mesures (temps par phrase, taux d'édition et qualité du résultat) sont dupliquées : on compare d'abord ces mesures selon l'utilisation ou non de mesures de confiance, puis selon l'ordre dans lequel les étapes 2 et 3 exposées dans la section précédente sont effectuées. La plupart des métriques sont aisément compréhensibles, mais certaines demandent un complément d'explications :

Perplexité de la séquence des erreurs : la perplexité peut-être interprétée comme le nombre d'alternatives qui sont présentées à un utilisateur pour chaque mot. En l'occurrence : éditer, ou ne pas éditer, telle est la question. Cette métrique mesure la capacité de la mesure de confiance à détecter les mots qui devront être édités. L'hypothèse de traduction proposée par Moses est alignée au sens de Levenshtein avec la traduction éditée par l'utilisateur. À chaque mot de l'hypothèse initiale, on associe 1 s'il est aligné avec un mot de l'hypothèse éditée (c'est-à-dire qu'il n'a pas été modifié), 0 sinon (l'utilisateur y a apporté une correction, l'a supprimé ou remplacé). le corpus est ainsi remplacé par une séquence de 0 et de 1. On calcule ensuite la perplexité de séquence selon deux modèles : la probabilité a priori des deux classes, d'une part ; dans ce cas la perplexité reflète la difficulté intrinsèque de la classification (elle est maximale quand les probabilités a priori des deux classes sont égales, et minimale quand l'une est nulle et l'autre certaine) ; et d'autre part, la perplexité donnée par les probabilités estimées à l'aide des mesures de confiance ; dans ce cas, la perplexité reflète la capacité des mesures de confiance à modéliser les erreurs. La différence entre les perplexités données par les deux modèles peut être interprétée comme l'aide à la décision qu'apportent les mesures de confiance (la désambiguïsation, ou réduction du choix). La perplexité varie donc entre 1 (le modèle a exactement détecté sans ambiguïté les mots qui devaient être corrigés) et 2 (le modèle n'a apporté aucune information utile à la prise de décision).

Qualité des traductions éditées : après l'expérience, toutes les traductions éditées ont été manuellement évaluées par un membre de l'équipe. Chaque traduction a reçu un score entre 1 et 5, d'une façon similaire à l'évaluation réalisée pour les campagnes WMT :

1. La traduction est inutilisable.
2. La traduction est assez mauvaise, mais une partie de l'information de la phrase source se retrouve dans la traduction.
3. On comprend globalement le sens, mais la traduction pourrait être nettement améliorée.
4. La traduction ne présente que quelques erreurs mineures.
5. La traduction est très bonne.

Cette évaluation manuelle permet de mesurer l'influence des mesures de confiance sur la qualité de la post-édition. Comme on pouvait s'y attendre, aucune des traductions éditées n'obtient le score 1 ou 2.

Corrélation entre les mesures de confiance et les éditions : on s'intéresse ici à la corrélation qui existe entre les mesures de confiance calculées par le système, et les décisions prises par l'utilisateur d'éditer ou non un mot. Les traductions proposées par le système de traduction automatique sont transformées en une séquence de 0 et de 1 selon la méthode précédemment exposée. On calcule ensuite la corrélation entre ces « scores » et les probabilités de correction fournies par le système.

Rapport entre le nombre d'édition et le nombre d'erreurs détectées : il s'agit du rapport du nombre d'éditions effectivement réalisées par l'utilisateur, et le nombre d'erreurs détectées par les mesures de confiance étant donné le seuil d'acceptation choisi par l'utilisateur. Si les mesures de confiance étaient parfaites et si l'utilisateur ajustait correctement le seuil d'acceptation, ce rapport vaudrait 1 (mais la réciproque n'est pas vraie). Si ce rapport est inférieur à 1, cela signifie que l'utilisateur a choisi un seuil tel que la détection automatique est trop sévère (cela peut être une façon de détecter un maximum d'erreurs, quitte à éventuellement avoir des mots injustement rejetés) ; s'il est supérieur à 1, l'utilisateur a choisi de rendre la classification plus laxiste, ce qui peut être une façon de ne mettre en valeur que les erreurs les plus graves.

	Sans mesures de confiance	Avec mesure de confiance
Temps moyen par traduction (secondes)	77	87
Taux d'édition moyen	30%	32%
Qualité moyenne des traductions éditées	4.3	4.2
	1ère expérience	2nde expérience
Temps moyen par traduction (secondes)	84.22	80.12
Taux d'édition moyen	29%	33%
Qualité moyenne des traductions éditées	4.2	4.3
Rapport éditions/détections	1.76	
Perplexité de la séquence d'erreurs	a priori : 1.71	a posteriori : 1.49
Corrélation mesures de confiance/éditions	0.23	

TABLE 4.1 – Influence des mesures de confiance sur la post-édition.

Les résultats en terme de rapidité d'éditions sont décevants : l'utilisation des mesures de confiance semble en fait ralentir les utilisateurs ! Ceux-ci nous ont en effet confirmé que les prédictions n'étaient pas assez fiables pour qu'ils leur fassent entièrement confiance, et qu'ils devaient donc passer du temps à vérifier les indications du système. Ces indications représentant une information supplémentaire à appréhender, cela peut expliquer que la présence d'indications de correction automatiques ralentisse les utilisateurs. Cette interprétation semble confirmée par une analyse plus fine du temps de post-édition : en effet, la différence observée entre les éditions avec et sans mesures de confiance est entièrement due au temps de « réflexion ». On observe aussi indépendamment de l'utilisation des mesures de confiance, les utilisateurs sont plus rapides dans la seconde partie de l'expérience que dans la première. Cela peut traduire l'acquisition d'une certaine aisance, ou une certaine impatience. La qualité des traductions éditées n'étant pas dégradée d'une étape à l'autre, cette dernière hypothèse peut être écartée. Cela suggère tout de même que la phase d'apprentissage devrait être plus longue. Dans toutes les configurations, le taux d'éditions est assez élevé, ce qui montre que les traductions n'étaient dans l'ensemble pas assez bonnes pour être compréhensibles telles quelles. Le rapport très élevé du nombre de corrections sur le nombre d'erreurs détectées suggère que les utilisateurs n'ont pas pu trouver de compromis

acceptable entre la détection des erreurs les plus importantes et l'augmentation du nombre du mots injustement rejetés lorsqu'on augmente le seuil d'acceptation. Enfin, dans le questionnaire final, la plupart des volontaires a indiqué qu'ils préféraient retraduire complètement une phrase plutôt que corriger une traduction automatique erronée.

4.4 Conclusion

Cette expérience nous a permis d'identifier un certain nombre de points qui doivent être améliorés en priorité :

- Les mesures de confiance doivent être calculées dynamiquement.
- Il faut proposer aux utilisateurs d'éditer un document, et non un ensemble aléatoire de phrases.
- La phase d'entraînement doit être beaucoup plus longue : en effet, on observe que les utilisateurs étaient beaucoup plus à l'aise lors de la seconde partie de l'expérience. De plus, à la fin de l'expérience, certains n'avaient pas bien saisi l'intérêt du curseur servant à modifier le seuil d'acceptation.
- L'interface du programme doit être plus sérieusement réfléchie, en gardant à l'esprit les aspects d'ergonomie, afin d'être certain que les utilisateurs peuvent pleinement tirer parti des fonctionnalités du programme, et que l'influence mesurée est bien celle des mesures de confiance et non celle de l'interface.
- Bien entendu, la précision des mesures de confiance doit être grandement améliorée. Tout d'abord, si l'utilisateur a le moindre doute sur leur précision, il perdra du temps à vérifier les prédictions du système. Ensuite, nous avons observé que les faux rejets (mots injustement considérés comme faux) étaient plus gênants pour les utilisateurs que les fausses acceptations. La précision de la classification, plus encore que le rappel, doit donc être améliorée.

Enfin, nous pensons qu'il est important de calculer des mesures de confiance non plus sur les mots ou les phrases, mais à un niveau intermédiaire, des segments (par exemple, un groupe nominal ou un groupe verbal). Cela permettrait d'attirer l'attention de l'utilisateur sur des groupes cohérents de mots à modifier dans leur ensemble, tout en gardant une analyse plus fine qu'au niveau de la phrase. De plus, des mesures d'ordre sémantique devraient être introduite, par exemple pour détecter une négation manquante, ce qui est délicat avec des mesures de type n-gramme ou trigger.

5

La campagne WMT 2012

Nous avons participé à la tâche d'estimation de qualité de la campagne d'évaluation WMT 2012 [Call 12] [18]. Il s'agissait d'utiliser des mesures de confiance pour approcher les évaluations humaines sur des traductions automatiques. L'approche du système de référence proposé est similaire à la nôtre : une régression est effectuée sur les évaluations en se basant sur différents paramètres prédictifs. Nous avons utilisé les paramètres proposés dans le système de référence, et certains de ceux décrits dans le présent manuscrit et dans [Rayb 11], soit un total de 67 paramètres. Nous avons également utilisé un algorithme de sélection de paramètres pour écarter les mesures les moins performantes, qui allongent inutilement le temps de calcul et risquent d'apporter plus de bruit que d'information. Notre système est décrit dans l'article [Lang 12].

Le système est présenté dans son ensemble dans la section 5.1. Les paramètres prédictifs utilisés sont décrits dans la section 2.3.2 et l'algorithme de sélection de paramètres dans la section 5.3. Enfin, je présenterai et commenterai les résultats dans la section 5.4.

5.1 Présentation du système

Les données fournies pour la campagne consistent en des phrases source en anglais, leurs traductions automatiques en espagnol, et pour chaque paire, une estimation de qualité entre 1 et 5 (moyenne des notes attribuées par trois experts), comme décrit dans la section 1.1. 1 signifie que la traduction est incompréhensible et 5 qu'elle est claire et intelligible. Avec chaque paire de phrases est fourni un vecteur de paramètres prédictifs de référence. En plus de ceux-ci, nous avons calculé d'autres paramètres prédictifs, et avons utilisé une SVM pour effectuer une régression de ces paramètres sur les scores fournis, comme décrit dans la section 1.4.2. L'objectif était de prédire au mieux l'évaluation humaine, par des moyens automatiques.

5.2 Paramètres prédictifs utilisés pour la campagne

Je vais commencer par décrire les paramètres prédictifs du système de référence. Certains ont été décrits dans la section 3.2 :

- Longueurs de la phrase source et de la traduction (section 3.2.5).
- Longueur moyenne des mots de la phrase source. L'idée est que les mots plus longs ont tendance à être plus rares (donc moins bien modélisés), et les mots les plus courts sont souvent des mots outils.

18. `http://www.statmt.org/wmt12/quality-estimation-task.html`

– Vraisemblances de la phrase source et de la traduction selon un modèle n-gramme (section 3.2.1).
– Nombre moyen d'occurrences des mots de la traduction au sein de l'hypothèse de traduction. Des mots répétés peuvent suggérer une erreur de traduction.
– Nombre moyen de traductions par mot de la phrase source. Soit $\mathcal{T}_1(s) \subset \mathcal{V}_T$ l'ensemble des mots de la langue cible qui sont une traduction potentielle du mot de la langue source s, et tels que

$$\forall t \in \mathcal{T}_1(s) \ . \ P_{IBM-1}(s|t) \geqslant 0.2$$

avec $P_{IBM-1}(s|t)$ la probabilité de traduction donnée par un modèle IBM-1. Alors, le paramètre est :

$$\frac{1}{Len(\mathbf{s})} \sum_{i=1}^{Len(\mathbf{s})} |\mathcal{T}_1(s_i)|$$

– Moyenne pondérée du nombre de traductions par mot source. Ce paramètre est similaire au précédent, mais les mots fréquents ont un poids plus faible dans la moyenne.
– Paramètres utilisant les fréquences des n-grammes : les fréquences sont divisées en quartiles. Ces paramètres indiquent combien d'unigrammes, bigrammes et trigrammes de la phrase source sont dans les quartiles 1 (les moins fréquents) et 4 (les plus fréquents).
– Nombre de signes de ponctuations dans la phrase source et dans la traduction.

Le système de référence était composé de 17 paramètres prédictifs, auquel nous ajoutons 49 autres, décrits dans la section 3.2.

5.3 Sélection de paramètres

Nous utilisons dans nos expériences un grand nombre de paramètres prédictifs. Certains ne sont pas très efficaces, comme le montrent les expériences de la section 3. Même si, en théorie, un paramètre qui n'apporte pas d'information utile sera ignoré par le classifieur, il vaut tout de même mieux l'écarter, afin de diminuer la complexité du calcul, d'atténuer les problèmes de surapprentissage, de manque de données et de dimensionalité, et de favoriser la convergence des méthodes d'optimisation numérique. Nous avons donc, pour cette campagne, développé un algorithme de sélection de paramètres. Il s'agit d'un algorithme « dégressif » (*backward*), c'est-à-dire qu'il part de l'ensemble de tous les paramètres et écarte les moins efficaces, par opposition aux algorithmes « progressifs » (*forward*), qui partent d'un ensemble vide et ajoutent les paramètres les plus efficaces [Guyo 03]. Plus précisément, il s'agit d'un algorithme dégressif glouton, c'est-à-dire qu'il supprime des paramètres un à un de l'ensemble des paramètres sélectionnés jusqu'à ce qu'un certain critère de qualité cesse de s'améliorer. Ce paramètre est *l'erreur moyenne* (EM) estimée sur un corpus de développement :

$$EM_{\mathcal{F}}(\mathbf{s}, \mathbf{r}) = \frac{\sum_{i=1}^{n} |s_i - r_i|}{n} \tag{5.1}$$

avec $\mathcal{F}$ l'ensemble des paramètres prédictifs utilisés, $\mathbf{s}$ la liste des scores prédits par le système, $\mathbf{r}$ la liste des scores de référence, et n la longueur de ces listes. L'algorithme est formalisé page 75 (algorithme **??**).

L'algorithme semble simple, mais plusieurs étapes sont nécessaires pour calculer l'EM : entraîner la régression des paramètres prédictifs sur les scores, effectuer la régression sur le corpus de développement, puis calculer l'erreur.

Algorithme 1: Algorithme de sélection de paramètres

```
begin
    ÉTAT INITIAL :
    F ← ensemble de tous les paramètres.
    err ← EM_F(s, r)
    repeat
        Δ ← 0
        forall the f ∈ F do
            F' ← F \ f
            err' ← EM_F'(s, r)
            if err − err' > Δ then
                Δ ← err − err'
                f* ← f
        if Δ > 0 then
            F ← F \ f*
            err ← err'
    until Δ < ε;
```

5.4 Résultats

Nous utilisons les corpora fournis par la campagne, et aucun corpus additionnel. Le premier corpus est un bitexte constitué de phrases source en anglais et leurs traductions de référence en espagnol, issues des corpora EUROPARL et News Commentary (distribués pour la même campagne). Ce corpus est utilisé pour estimer les modèles de traduction et de langage. Pour l'estimation de qualité, nous utilisons un second corpus de 1832 phrases en espagnol fournies pour la campagne avec leurs traductions automatiques par le système de référence, les estimations de qualité de référence et les valeurs des paramètres prédictifs de référence. Ce corpus est scindé en deux : 1000 paires de phrases pour le corpus d'entraînement des SVM, et 832 pour le corpus de développement (utilisé par l'algorithme de sélection de paramètre, section 5.3). Un corpus de test a été distribué par la suite.

Les résultats en termes d'EM et *Erreur Quadratique Moyenne* (EQM) sont présentés dans le tableau 5.1. L'EM a été décrite dans la formule 5.1, et l'EQM est définie par :

$$EQM(\mathbf{s}, \mathbf{r}) = \sqrt{\frac{\sum_{i=1}^{n}(s_i - r_i)^2}{n}} \tag{5.2}$$

Chaque ligne du tableau 5.1 correspond aux performances d'un ensemble de paramètres prédictifs. REF+LORIA est l'union des paramètres de référence (section 2.3.2) et de ceux que nous proposons (section 3.2). La colonne 'sélection param.' indique si l'algorithme de sélection de paramètres a été appliqué ou non. Nous avons testé deux noyaux pour la SVM : linéaire (indiqué par LIN dans le tableau) et *Radial Basis Function*, comme décrit dans la section 1.4.2, indiqué par *RBF*. Les paramètres γ et C du noyau RBF sont soit les paramètres par défaut de LibSVM (DEF dans le tableau), soit optimisés par grid search sur le corpus de développement (OPT). Les valeurs de l'EM et de l'EQM sont données pour le corpus de développement et le corpus de test.

Ces résultats montrent que les améliorations observées sur le corpus de développement sont toujours confirmées sur le corpus de test. Les paramètres de REF seuls permettent déjà d'obtenir

			Dev		Test	
Paramètres	**sélection param.**	**noyau**	**EM**	**EQM**	**EM**	**EQM**
REF	oui	RBF DEF	0.63	0.79	0.69	0.83
LORIA	non	RBF DEF	0.66	0.82	0.73	0.87
REF+LORIA	non	RBF DEF	0.62	0.78	0.69	0.82
REF+LORIA	oui	RBF DEF	0.61	0.77	0.69	0.83
REF+LORIA	non	RBF OPT	0.62	0.77	0.68	0.82
REF+LORIA	non	LIN	0.62	0.78	0.69	0.83
REF+LORIA	oui	LIN	0.61	0.77	0.68	0.82

TABLE 5.1 – Performances des différentes configurations en terme d'EM et EQM

	$\geqslant 0.9$	$\geqslant 0.8$	$\geqslant 0.7$
+/-	64	103	127
+	56/49/3/4	94/87/3/4	117/105/6/6
-	8/0/4/4	9/0/4/5	10/0/4/6

TABLE 5.2 – Corrélation des paramètres au sein de l'ensemble REF+LORIA.

de bonnes performances, meilleures que celles de LORIA seul. La fusion des deux ensembles (REF+LORIA) permet d'améliorer encore ces résultats, mais le progrès n'est pas statistiquement significatif. L'algorithme de sélection de paramètre permet de progresser d'encore 0.01 point, ainsi que l'optimisation des paramètres du noyau RBF. Cependant, nous avons eu la surprise de constater que le noyau linéaire donnait des résultats aussi bons que le noyau RBF optimisé. Avec ce noyau, le système que nous avons proposé pour la campagne d'annotation de confiance WMT2012 se place au cinquième rang sur dix-neuf systèmes évalués.

Outre les performances de l'estimation de qualité, nous avons voulu étudier la corrélation existant entre les différents paramètres prédictifs. Nous avons donc calculé la corrélation linéaire existant entre chaque paire de paramètres de l'ensemble REF+LORIA, soit 2145 paires. Le tableau 5.2 présente dans la ligne +/- le nombre de paires de paramètres dont la valeur absolue du coefficient de corrélation dépasse les seuils 0.9, 0.8 ou 0.7 respectivement. Dans la ligne +, j'indique combien de ces paires ont un coefficient de corrélation positif et dans la ligne moins, combien ont un coefficient de corrélation négatif. J'indique à chaque fois quatre valeurs : le nombre de paires dans l'ensemble REF+LORIA, le nombre de paires dans LORIA (c'est-à-dire les paires de paramètres prédictifs fortement corrélés au sein de ceux que nous proposons), le nombre de paires dans REF, et le nombre de paires d'un élément de REF et d'un élément de LORIA (la redondance entre nos paramètres et ceux de la référence). On remarque qu'environ 6% des paires de paramètres sont fortement corrélés, essentiellement au sein de LORIA. L'ensemble LORIA comprenant presque trois fois plus de paramètres que REF, et beaucoup d'entre eux étant des variantes les uns des autres, ce n'est pas très étonnant. Il existe très peu de corrélations entre les paramètres de REF et de LORIA (notons que nous avons pris soin de supprimer en amont les paramètres qui étaient manifestement les mêmes).

Plus en détail, on observe une forte corrélation entre les paramètres utilisant les n-grammes (section 3.2.1) *vraisemblance dans n-gramme* et *repli moyen*, ainsi qu'entre les paramètres *vraisemblance n-gramme* et *vraisemblance n-gramme à rebours*. Les paramètres *longueur de la phrase*

source et *longueur de la traduction* sont, sans surprise, très fortement corrélés (avec un coefficient de 0.98), ainsi que les vraisemblances de la phrase source et de la traduction. Il y a très peu de paires dont le coefficient de corrélation est négatif. On retiendra celles-ci, peu surprenantes, dont le coefficient de corrélation se situe entre -1 et -0.7 :

- longueur de la phrase source, et sa vraisemblance dans un modèle n-gramme ;
- longueur de la traduction, et sa vraisemblance dans un modèle n-gramme ;
- rapport du nombre de mots hors vocabulaire dans la phrase source sur sa longueur, et le proportion d'unigrammes de la phrase source effectivement observés dans le corpus d'apprentissage ;
- nombre de mots hors vocabulaire dans la phrase source, et la proportion d'unigrammes de la phrase source effectivement observés dans le corpus d'apprentissage.

Cette information semble utile pour la sélection de paramètres, afin de réduire leur nombre en éliminant ceux qui sont redondants. Mais cela n'est pas si simple : prenons l'exemple des longueurs de la phrase source et de la traduction. Leur corrélation est de 0.98, mais si on a une paire de phrases où les longueurs de la phrase source et de la phrase cible sont très différentes, cela indique probablement une erreur de traduction. Les deux paramètres sont donc utiles, et doivent être conservés, car l'un est inutile sans l'autre. Ce n'est donc pas ce critère qui est utilisé pour la sélection de paramètres, mais le critère de réduction d'erreur décrit plus haut.

Enfin, intéressons-nous aux paramètres qui sont écartés par l'algorithme de sélection. Je ne présenterai ce résultat que pour l'ensemble REF+LORIA avec le noyau linéaire. L'algorithme écarte 18 paramètres prédictifs de l'ensemble LORIA (sur 49, soit 37%) et 3 de l'ensemble REF (sur 17, soit 18%). La plupart des paramètres écartés issus de l'ensemble LORIA sont les probabilités n-gramme avec prise en compte du contexte (section 3.1.3). En effet, on pouvait s'attendre à ce que la prise en compte du contexte des mots soit inutile au niveau de la phrase, puisqu'on ne s'intéresse pas au score d'un mot mais de la traduction dans son ensemble.

5.5 Conclusion

La participation à cette campagne a pour nous, bien sûr, été une opportunité de confronter notre système d'estimation de confiance à ceux proposés par d'autres équipes de recherche, mais elle a aussi et surtout été l'occasion d'effectuer une analyse fine des propriétés des paramètres prédictifs que nous utilisons et de ceux proposés par d'autres centres de recherche, notamment leurs corrélations, et de tester la robustesse de notre système à l'ajout de nouvelles sources d'informations (les paramètres prédictifs proposés par la campagne). Nous avons pu vérifier que notre système, comme on l'attend des méthodes d'apprentissage statistique, bénificie effectivement de ces informations supplémentaires (amélioration d'un point de l'EQM par rapport au système de référence, et de 5 points par rapport à notre système sans les paramètres de référence), et que l'ajout de ces nouveaux paramètres se faisait sans difficulté particulière. Cela nous a également poussé à implémenter un algorithme de sélection de paramètre dans notre système, qui contribue marginalement à l'amélioration des performances et à la réduction du temps de calcul.

6

Conclusion

Ici s'achève la partie de ce manuscrit consacrée à l'étude des mesures de confiance. Rappelons en les points principaux : après une formalisation mathématique du problème de l'estimation de confiance, j'ai présenté une méthode permettant la génération à faible coût d'une grande quantité de données d'entraînement pour l'estimation des paramètres de classification et de regression, donnant de bons résultats. Nous avons ensuite développé un système d'estimation de confiance utilisant les paramètres prédictifs que nous avons estimés les plus représentatifs de l'état de l'art et des paramètres originaux utilisant l'information mutuelle et d'autres issus de la reconnaissance vocale, ainsi que des techniques de classification et de régression utilisant les réseaux de neurones, la régression logistique ; la régression des moindres carrés partiels et les machines à support de vecteurs. Nous avons déterminé que de façon générale, les réseaux de neurones et la régression des moindres carrés partiels appliqués aux paramètres prédictifs basés sur les n-grammes et le repli étaient les techniques à la fois les plus efficaces et les moins coûteuses en terme de calcul. La combinaison de plusieurs paramètres prédictifs à l'aide de ces techniques de régression et classification permet d'améliorer les performances de la classification d'1,3 points au niveau des mots et 6 points au niveau des phrases en termes de taux d'égale erreur, par rapport au meilleur paramètre prédictif utilisé seul.

Nous avons conduit une expérience destinée à évaluer l'intérêt et l'utilisabilité des mesures de confiance pour une tâche de post-édition. Cette expérience a montré que notre système d'estimation de confiance n'était pas prêt à être utilisé pour une tâche réelle, mais nous a permis de dégager des axes de recherches prometteurs et des points à améliorer en priorité. Il nous est notamment apparu qu'il était important de distinguer les notions de performance d'un système d'annotation de confiance (erreur d'une regression, précision et rappel d'une classification) de l'utilité perçue par ses utilisateurs, dont le ressenti dépend du temps gagné, au confort d'utilisation, du nombre de faux positifs, etc. Nous avons également constaté que les utilisateurs dusystème de post-édition ne souhaitent ni une annotation au niveau des mots, ni au niveau des phrases, mais à un niveau intermdiaire, des segments ayant un sens pour un humain.

Enfin, nous avons confronté notre approche à un système de référence lors de la campagne WMT2012. Les résultats montrent une amélioration modeste par rapport à la référence. Le développement d'un algorithme de sélection de paramètres nous a permis d'obtenir des informations sur ceux qui sont les moins efficaces. Les résultats suggèrent également qu'une SVM utilisant un noyau plus simple que celui que nous utilisions dans nos expériences précédentes (section 3), à savoir le noyau linéaire, donne d'aussi bon résultats.

Dans la partie suivante, je vais présenter un système de traduction automatique que j'ai développé, destiné à la traduction automatique de la parole spontané. Ce système est développé

à partir des outils bien connus Sphinx et Moses, et implémente des techniques originales de segmentation de la parole continue. Il a également la particularité de baser le processus de traduction sur une transcription phonétique de la parole, et non une transcription en mots et phrases, afin de le rendre plus robustes aux erreurs de reconnaissance. Enfin, je propose d'intégrer des mesures de confiance au processus de traduction, là encore en vue d'améliorer la robustesse du système.

Troisième partie

S2TT : un système de traduction automatique de la parole spontanée à grand vocabulaire.

Nous allons dans cette partie présenter un système de traduction automatique de la parole spontanée tirant parti des caractéristiques propres à cette forme de communication, par opposition à l'approche naïve consistant à enchaîner une étape de transcription automatique et une étape de traduction de texte. Il faudra donc relever les défis évoqués dans la section 1 de cette partie (page 85) et développer des algorithmes spécifiques pour y répondre.

Je m'attacherai donc à développer un système de traduction de la parole spontanée à grand vocabulaire. Il sera testé sur le corpus de la campagne ESTER2 [Gall 06], des enregistrements d'informations radiodiffusées en français (France Inter, Radio France International, TVME). Il ne s'agit pas à proprement parler de parole spontanée (on parle de parole *planifiée — planed speech*), mais cela en est assez proche : les émissions sont en direct et comprennent des interviews, notamment par téléphone, ce qui introduit la difficulté supplémentaire de la qualité d'enregistrement. L'avantage d'utiliser ce corpus est qu'il est bien connu de la communauté, que des modèles acoustiques et de langage efficaces ont été développés pour la transcription automatique, et que des transcriptions de référence sont disponibles. Les modèles de traduction seront entraînés sur la partie *news commentary* (émissions d'information) du corpus EuroParl (voir partie I, section 1.5.6), et le cas échéant adaptés pour rendre possible la traduction de transcriptions phonétiques (section 2.4).

Je proposerai une évaluation des performances du système dans différentes configurations (section 3) en le comparant à différentes références. Même si certains résultats sont qualitatifs (la complexité du système implique des efforts de développement qu'il n'a pas été possible de mener à terme pendant la durée de cette thèse), nous verrons qu'il présente des performances honorables pour un prototype.

1

La traduction automatique de la parole : introduction et état de l'art

La traduction automatique de la parole spontanée présente des difficultés particulières. Un style souvent plus décontracté qu'à l'écrit, une grammaire et une syntaxe parfois approximatives, des hésitations, des répétitions, des disfluences, respirations, pauses et bruits divers rendent la tâche ardue aux systèmes de traduction automatique et même aux interprètes professionnels. On ne peut donc pas se contenter d'utiliser les systèmes développés pour la traduction de texte pour traduire les hypothèses de transcriptions générées par un système de reconnaissance : il faut soit développer des modèles et des systèmes spécifiques, soit traiter les hypothèses de transcription avant de les traduire afin qu'elles soient le plus proche possible de données textuelles. Les auteurs de [Dech 07] montrent par exemple que la suppression des disfluences et d'autres étapes de normalisations permettent d'obtenir une amélioration de 0.4 du score BLEU par rapport à la traduction directe de la meilleure hypothèse de reconnaissance. De plus, la ponctuation n'est pas explicite dans le discours, mais est en général présente dans les corpora utilisés pour entraîner les modèles de traduction : Matsoukas *et al.* montrent dans [Mats 07] qu'on améliore la qualité de la traduction en réintroduisant de la ponctuation dans les hypothèses de reconnaissance, même si cette amélioration est modeste. Beaucoup plus fondamental est le problème de la segmentation [Gale 07, Dech 07, Mats 07, Fuge 08]. En effet, un texte écrit est naturellement segmenté en phrases. Mais cette segmentation est beaucoup moins évidente à l'oral, que ce soit à cause d'une syntaxe relâchée ou parce qu'un discours est parfois improvisé (on peut commencer une phrase sans bien savoir quand on va la finir). Un discours est donc plutôt un flux de mots avec des pauses qu'une succession de phrases [Fuge 08]. Or, non seulement les données sur lesquelles sont estimés les paramètres des modèles de traduction sont en général segmentées en phrases, mais les modèles IBM reposent intrinsèquement sur cette segmentation, car ils modélisent la probabilité d'une phrase. Le flux de mots généré par le module de reconnaissance d'un système de traduction de la parole doit donc être segmenté avant d'être traduit. Les auteurs de [Mats 07] montrent que leur meilleure méthode de segmentation (utilisant un modèle de langue et les silences du signal acoustique, par opposition à la méthode naïve consistant à segmenter uniquement lors des silences) donne une réduction du TER trois fois plus importante (-2.7 points) que celle obtenue grâce à une réduction de trois points du WER du module de reconnaissance (-0.8 points de TER). Mieux encore, Fügen *et al.* montrent dans [Fuge 08] que la meilleure segmentation pour la traduction n'est pas celle qui imite le découpage d'un texte en phrase, mais qu'il vaut mieux générer des segments courts (neuf mots en moyenne) correspondant à des parties cohérentes d'une phrase. Le taux d'erreur de la reconnaissance, bien sûr, a une influence importante : les auteurs

de [Fuge 08] et [Dech 07] montrent qu'il y a une corrélation négative quasiment linéaire avec le score BLEU. Mais le vecteur oral n'est pas seulement un obstacle à la traduction automatique : il est également porteur d'informations absentes du médium écrit, notamment la prosodie, qui est porteuse de sens. Marie Ostendorf montre dans [Oste 09] que les transcriptions riches (c'est-à-dire contenant plus d'informations que la seule transcription des paroles prononcées) bénéficie aux applications de traitement de la langue naturelle et notamment la traduction automatique, car elles permettent d'utiliser des informations inexistantes à l'écrit, comme la prosodie. Par exemple, le système développé dans le cadre du projet VerbMobil (traduction de conversations typiques de voyage, réservation d'hôtel et prise de rendez-vous — [Bub 97, Wahl 00]) utilise des informations sur la prosodie pour ses modèles sémantiques utilisés pour la traduction.

Toute la difficulté de la traduction de la parole est donc de développer des stratégies pour prendre en compte les spécificités de l'oral. On distingue trois grandes tendances dans la littérature. L'approche classique, que j'appellerai ici *linéaire* (section 1.2), consiste à enchaîner une étape de transcription (section 1.1) et une étape de traduction. Afin d'être robuste aux disfluences et aux erreurs de reconnaissance, et pour prendre en compte les spécificités de l'oral, les systèmes utilisant ces approches doivent traiter l'hypothèse de transcription avant de la traduire (suppression des disfluences, segmentation, ...), pour obtenir des performances raisonnables.

Afin de diminuer la sensibilité aux erreurs de reconnaissance et de mieux modéliser le processus de traduction de la parole, on peut étudier des approches différentes, et notamment des méthodes permettant de se passer d'une transcription complète du signal acoustique. Par exemple, on peut appliquer les algorithmes de traduction automatique statistique aux treillis de mots fournis par un système de reconnaissance (que j'appellerai *couplage faible* — section 1.4). On peut également proposer une approche totalement différente, ne distinguant plus transcription et traduction, en utilisant des transducteurs finis (je parlerai de *couplage fort*, section 1.3). Ces différentes approches sont présentées et comparées dans [Seli 00]. Dans ce chapitre, après une brève présentation de la reconnaissance automatique de la parole, nécessaire pour la compréhension des systèmes de traduction de la parole, je vais passer en revue ces différentes méthodes, en expliquant rapidement l'approche linéaire et l'approche par transducteurs, avant de m'attarder plus longuement sur l'approche intégrée, qui est celle utilisée par le prototype que je propose (le logiciel S2TT — chapitre 2).

1.1 La reconnaissance automatique statistique de la parole

Je vais dans cette section présenter brièvement les principes de la reconnaissance automatique de la parole (RAP) par des méthodes statistiques, prélude indispensable à la traduction automatique de la parole. La RAP statistique est conceptuellement très proche de la traduction automatique statistique, dont elle est l'ancêtre, et dont le lecteur a eu une présentation dans la partie I, section 1.5. Je passerai donc assez rapidement sur la RAP, le but étant seulement de présenter les principes nécessaires à la compréhension du système S2TT (section 2). Pour plus de détails, on pourra se référer par exemple à [Smai 91] et [Hato 03].

Nous allons reprendre les notations déjà utilisées dans les partie I, section 1.5 et partie II, section 1. Le problème de la reconnaissance automatique de la parole est le suivant : on suppose qu'une phrase $\mathbf{s} \in \mathcal{V}_S^*$ a été prononcée, et on observe un signal acoustique $\mathbf{a}$. On veut, à partir de $\mathbf{a}$, retrouver $\mathbf{s}$ selon un modèle probabiliste Θ :

$$\mathbf{s} = \arg\max_{\mathbf{w} \in \mathcal{V}_S^*} P_\Theta(\mathbf{w}|\mathbf{a})$$

En vertu de la règle de Bayes, cette équation se réécrit (comme dans le cas de la traduction automatique statistique, partie I, section 1.5) :

$$\begin{aligned}\mathbf{s} &= \arg\max_{\mathbf{w}\in\mathcal{V}_S^*} P_\Theta(\mathbf{w}|\mathbf{a}) \\ &= \arg\max_{\mathbf{w}\in\mathcal{V}_S^*} P_\Theta(\mathbf{a}|\mathbf{w}) \times \frac{P_\Theta(\mathbf{w})}{P_\Theta(\mathbf{a})} \\ &= \arg\max_{\mathbf{w}\in\mathcal{V}_S^*} P_\Theta(\mathbf{a}|\mathbf{w}) \times P_\Theta(\mathbf{w})\end{aligned}$$

$P_\Theta(\mathbf{w})$ est donné par un modèle de la langue source, souvent un modèle n-gramme (partie I, section 1.5.4). $P_\Theta(\mathbf{a}|\mathbf{w})$ est donné par un modèle acoustique utilisant historiquement des *modèles de Markov cachés* (*Hidden Markov Model, HMM*).

Afin de décrire le signal acoustique, continu, on commence par le discrétiser en segments recouvrants de quelques dixièmes de millisecondes. Chaque segment est ensuite représenté par un vecteur de coefficients réels représentant, pour simplifier, les harmoniques et la force du signal. Ce vecteur est appelé une *observation* et sera noté $\mathbf{o}_i$. Le signal se trouve donc représenté par la séquence d'observations $\mathbf{o}_1, .., \mathbf{o}_T$. On suppose que cette séquence d'observations a été émise par une succession d'états *cachés* $q_1, .., q_T$. Un modèle de Markov est défini par :

- L'ensemble des états $\mathcal{Q}$,
- L'ensemble des observations possibles $\mathcal{O}$,
- Les probabilités initiales des états $\pi(q) = P(q_1 = q)$ pour $q \in \mathcal{Q}$,
- Les probabilités de transitions entre états $\{P(Q_{i+1} = q|Q_i = q')\}_{(q,q')\in\mathcal{Q}^2}$,
- Les distributions des observations, c'est-à-dire les probabilités qu'un état q émette une observation $\mathbf{o}$: $b_i(\mathbf{o}) = P(\mathbf{O}_i = \mathbf{o}|Q_i = q)$.

À chaque mot du vocabulaire correspond une succession d'états. Le but du HMM est de retrouver la succession d'états la plus vraisemblable, et donc la suite de phonèmes prononcés (dont on déduit la suite de mots), étant donnée une suite d'observations. En pratique, l'entraînement du modèle acoustique consiste à déterminer un modèle de Markov pour chaque mot et ses variantes de prononciations. Chaque modèle de Markov associera une vraisemblance à une suite d'observations données. Le décodage consiste ensuite à déterminer la succession des modèles (donc des mots) qui maximise la vraisemblance de la séquence d'observations, en tenant compte également du modèle de langue, ce qui revient à déterminer le chemin de moindre coût dans un graphe. Cela peut être réalisé au moyen de l'algorithme de Viterbi. Comme dans le cas de la traduction, on peut extraire du graphe la meilleure hypothèse de transcription, ou les n meilleures (n-best list), ou bien conserver le graphe lui-même pour représenter un ensemble d'hypothèses.

1.2 Approche linéaire de la traduction automatique de la parole

Cette approche de la traduction automatique de la parole est conceptuellement la plus simple, et a donc été la première explorée [Take 98, Ney 00, Ney 04, Zens 05, Maus 06]. On suppose que le module de reconnaissance de la parole fournit la meilleure hypothèse, et on utilise un système de traduction automatique pour traduire cette hypothèse, après éventuellement une étape parfois très complexe de transformation et d'analyse (figure 1.1) : segmentation, suppression des disfluences, analyse sémantique, etc.

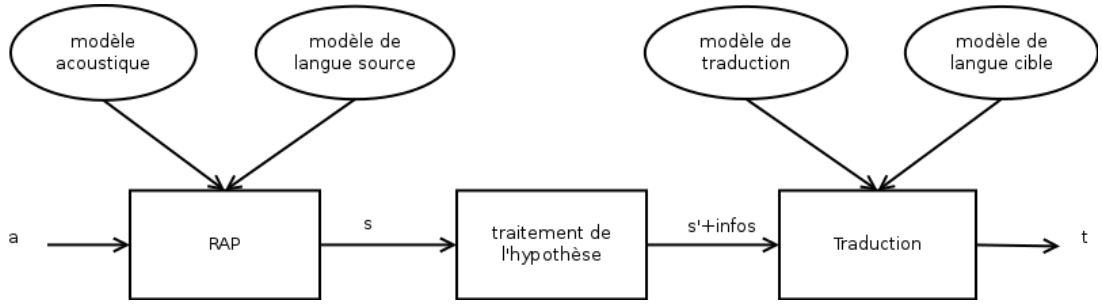


FIGURE 1.1 – Système linéaire (non couplé) de traduction automatique de la parole.

Comme évoqué précédemment, les étapes de traitement du flux des mots reconnus sont d'une importance cruciale pour la qualité de la traduction. Une bonne segmentation, par exemple, peut largement améliorer la qualité et la vitesse de la traduction [Mats 07, Fuge 08]. Paradoxalement, les auteurs montrent qu'une bonne segmentation n'est pas forcément un découpage en phrases ressemblant à celui d'un texte écrit, car il est mal défini à l'oral. Au contraire, préférer des segments courts (neufs mots en moyenne), formant un ensemble cohérent mais pas forcément une phrase, permet de nettement améliorer la traduction. L'analyse de la phrase peut également être très poussée, comme dans le système proposé dans le cadre du projet VerbMobil [Bub 97, Wahl 00], qui effectue une analyse syntaxique et sémantique de la phrase et utilise les informations ainsi extraites pour la traduction. Cela peut aller encore plus loin : les systèmes Janus [Wosz 93], Nespole ! [Metz 02] et Mastor [Liu 03] n'effectuent pas à proprement parler une traduction de la phrase reconnue ; à partir de l'hypothèse de transcription, ils génèrent une représentation symbolique du sens de la phrase (une interlangue — partie I, section 1.2) et à partir de cette représentation génèrent une phrase dans la langue cible.

Malgré son apparente simplicité, cette approche est utilisée en pratique et peut donner de très bons résultats. On citera notamment le système ATR-MATRIX (traduction entre l'anglais, le chinois et le japonais — [Naka 06, Take 98]), Diplomat [Blac 02], développé pour les besoins de l'armée américaine, ou le système développé par Fügen *et al.* [Fuge 08] notamment dans le cadre du projet TC-STAR [19] (voir une description du projet dans [Goll 05]).

1.3 Les transducteurs

Les systèmes linéaires ont l'avantage de la simplicité de développement, car il sont souvent construits à partir de composants préexistant pour la reconnaissance et la traduction qui, en outre, peuvent avoir bénéficié de longues phases de développement et d'optimisation. Cette approche a néanmoins des limitations intrinsèques :

- Puisque seule la meilleure hypothèse du module de reconnaissance est traduite, les erreurs de transcription sont mécaniquement propagées dans la traduction.
- Déterminer séparément la transcription la plus probable et sa traduction la plus probable peut conduire à une traduction qui n'est pas la plus probable étant donnée le signal acoustique (section 1.4).

Des modèles spécifiques *intégrés* ou *fortement couplés* ont donc été développés. Ils reposent sur l'idée que l'étape de reconnaissance et l'étape de traduction ne seront plus distinguées, mais qu'on déterminera en une seule étape l'hypothèse de traduction la plus probable étant donnée

19. http ://www.tc-star.org

le signal acoustique :

$$\mathbf{t}^* = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{a})$$

La méthode la plus répandue est d'utiliser des transducteurs finis (*Finite State Transducers, FST*). Les FST sont des automates finis (au sens de la théorie des langages formels) ayant la fonction supplémentaire d'émettre un symbole (éventuellement le symbole vide ϵ à chaque transition). On imagine donc que si le langage reconnu est la langue source et les symboles émis des mots de la langue cible, on peut essayer de construire un transducteur effectuant la traduction de la langue source vers la langue cible. Pour illustrer, même si ce n'est pas très rigoureux, on peut dire que les transducteurs développés ressemblent aux treillis de mots décrits dans la partie I, section 1.5.5, sauf qu'ils modélisent le processus de traduction entre les langues source et cible en toute généralité (dans les limites du domaine couvert) et pas seulement pour une phrase source particulière. On voit donc que pour des raisons combinatoires, l'utilisation de FST est donc réservée aux applications où le domaine est restreint, comme dans le cadre du projet EuTrans [Amen 00, Casa 01] (conversation de voyage et réservation d'hôtel), du système Atros [Llor 99] (traduction de requêtes dans une base de données géographiques) ou Transnizer [Bang 01] (traduction de requêtes dans un centre d'appels).

Les transducteurs sont donc construits en énumérant les suites de mots ou de segments possibles dans la langue source selon le domaine couvert (langage d'entrée). Les symboles émis sont les traductions possibles. Un mot pouvant avoir plusieurs traductions, les transitions sont probabilistes et déterminées selon un modèle de traduction (un modèle IBM, par exemple). Le réordonnancement est implicitement géré par le transducteur, en « retardant » l'émission de certaines traduction (voir l'exemple de la figure 1.2 ci-dessous). L'état acceptant correspond à la fin de phrase (`</s>`). Si une phrase source peut-être acceptée en suivant plusieurs chemins, c'est celui de plus grande probabilité qui est choisi (en utilisant l'algorithme de Viterbi par exemple). Supposons par exemple qu'on veuille traduire des phrases couramment utilisées en voyage. Notre guide de conversation contient les quatre phrases suivantes :

phrase française	traductions anglaises
Je cherche la plage	*I am looking for the beach* *I look for the beach*
Je cherche un hôtel	*I am looking for an hostel* *I look for an hostel*
Je cherche un hôtel bon marché	*I am looking for a budget hostel* *I am looking for a cheap hostel*
Je veux un hôtel bon marché	*I want an budget hostel* *I want a cheap hostel*

Alors le processus de traduction est modélisé par le transducteur donné dans la figure 1.2.

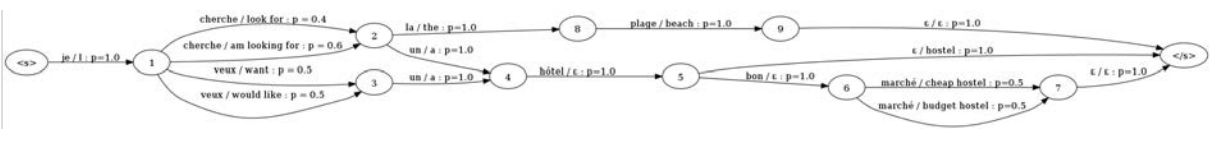


FIGURE 1.2 – Exemple de transducteur fini pour la traduction de phrases de voyage.

Afin d'intégrer le processus de reconnaissance au processus de traduction et qu'ils ne fassent plus qu'un, au lieu d'accepter un langage fait de mots, les transducteurs vont accepter un langage fait de sons. Les arcs ne porteront plus comme étiquette des graphèmes, mais des HMM (section 1.1). La probabilité d'une transition ne sera donc plus seulement la probabilité de traduction, mais les probabilités de reconnaissance (fournie par le HMM) et de traduction combinées. Le lecteur trouvera plus de détails dans [Casa 09]. Cette approche a été proposée dès les années 90 [Jime 95] et développée par la suite dans, entre autres, [Gisp 02, Casa 04a, Casa 04b]. Les auteurs de [Bang 02] proposent de séparer les processus de traduction (effectuée à l'aide d'un FST) et de réordonnancement (effectuée après l'étape de traduction) afin de diminuer la complexité du transducteur. Les modèles qu'ils utilisent alignent non pas les graphèmes, mais les arbres de dépendance, selon une méthode proposée dans [Alsh 98]. Cela permet de mieux prendre en compte les règles spécifiques à chaque langue concernant l'ordre des mots (par exemple, position des adjectifs par rapport au nom). [Matu 05] propose d'utiliser les FST pour traduire les treillis générés par le module de reconnaissance. Cette méthode se rapproche du « couplage faible » décrit dans la section suivante. Il montre à cette occasion que plus un treillis est dense (c'est-à-dire plus il représente un vaste ensemble d'hypothèses), meilleure est la traduction : cela confirme le bien fondé de l'idée d'utiliser les treillis pour améliorer la robustesse aux erreurs de reconnaissance (sections 1.4 et 2.1). [Math 06] décrit également une méthode pour traduire les treillis fournis par le système de reconnaissance, mais au lieu de traduire en une fois tout le treillis, il propose de traduire des segments afin de diminuer la complexité, qui seront ensuite réarrangés selon un modèle spécifique.

1.4 Approche statistique faiblement couplée de la traduction automatique de la parole

Comme on l'a vu dans la section précédente, les FST représentent une approche nouvelle de la traduction automatique de la parole, qui ne distingue pas la reconnaissance de la traduction. Cela permet d'optimiser directement la probabilité de la traduction étant donné le signal acoustique, et donc d'éviter l'étape de transcription proprement dite : après tout, ce n'est pas elle qui nous intéresse, et les erreurs commises à ce niveau se propagent à la traduction. Malheureusement, la taille des transducteurs impose de limiter le domaine couvert par le système : un vocabulaire riche (de l'ordre de cent mille mots pour le système de reconnaissance ANTS utilisé lors des campagnes Ester, Ester2 et Étape) donnerait un transducteur bien trop gros pour être utilisé en pratique. Par exemple, le système développé dans le cadre du projet EuTrans a un vocabulaire d'environ 600 mots [Amen 00]. Afin de passer à l'échelle, nous avons opté pour une approche *faiblement couplée* : la transcription et la traduction ne forment plus un seul processus, mais on n'effectue cependant pas une transcription « totale » : le résultat du module de reconnaissance n'est pas dans ce cas une hypothèse, mais un ensemble d'hypothèses, qui peut être représenté par une n-best list, un treillis ou, comme dans le système que nous proposons, un réseau de confusion (section 2.2).

Dans ce qui suit, je vais donc présenter les différents systèmes à couplage faible proposés dans la littérature, avant de présenter celui que j'ai développé. Je commencerai par une formalisation selon l'approche du canal bruité similaire à celle présentée dans la partie I, section 1.5 pour la traduction de texte. Si $\mathbf{a}$ est un signal acoustique, $\mathbf{t}$ une phrase dans la langue cible et $\mathbf{s}$ une phrase dans la langue source (la langue du signal), on cherche une phrase $\mathbf{t}^*$ telle que :

$$\mathbf{t}^* = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{a}) \tag{1.1}$$

Afin de rendre le problème accessible à l'estimation numérique, on doit cependant le décomposer, et introduire la notion de transcription intermédiaire. Cependant, en l'état, cette formulation n'impose pas de devoir déterminer une transcription correcte :

$$\begin{aligned}
\arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{a}) &= \arg\max_{\mathbf{t}} P(\mathbf{a}|\mathbf{t})P(\mathbf{t}) \\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{a},\mathbf{s}|\mathbf{t})P(\mathbf{t}) \qquad (1.2)\\
&= \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{a}|\mathbf{s},\mathbf{t})P(\mathbf{s}|\mathbf{t})P(\mathbf{t}) \qquad (1.3)
\end{aligned}$$

Le passage de 1.2 à 1.3 se fait en vertu de la règle de Bayes. Pour simplifier encore l'expression, on va faire l'hypothèse très naturelle que la probabilité du signal acoustique ne dépend que de la phrase source $\mathbf{s}$, et non de la phrase cible $\mathbf{t}$. Intuitivement cela signifie que les prononciations possibles d'une phrase sont entièrement déterminées par sa transcription. Cette hypothèse n'est prise en défaut que dans les cas où une phrase peut avoir plusieurs sens qui ne se distinguent que par leur prononciation, ce qui peut arriver, mais reste tout de même fort rare. Ainsi, on peut simplifier 1.3 et on obtient :

$$\arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{a}) = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{a}|\mathbf{s})P(\mathbf{s}|\mathbf{t})P(\mathbf{t}) \tag{1.4}$$

On reconnaît donc :

$P(\mathbf{a}\|\mathbf{s})$:	modèle acoustique
$P(\mathbf{s}\|\mathbf{t})$:	modèle de traduction
$P(\mathbf{t})$:	modèle de la langue cible

Et on observe deux propriétés remarquables : d'une part, **le modèle de langue source ne joue aucun rôle**, et d'autre part, **la meilleure traduction au sens probabiliste ne dépend pas d'une transcription particulière**. Cependant, la somme sur $\mathbf{s}$ (c'est-à-dire sur l'ensemble des phrases possibles de la langue source) rend cette formulation inutilisable en pratique. On va donc faire une hypothèse plus forte : $P(\mathbf{a}|\mathbf{s})$ **est presque toujours nul, sauf pour un ensemble réduit $\mathcal{S}$ de phrases $\mathbf{s}$**. C'est cela qui distingue le couplage faible du couplage fort, puisque cela revient à réintroduire une forme de transcription. Intuitivement, cela signifie qu'on va ignorer, dans l'estimation de probabilité, les phrases de la langue source qui n'ont « aucun rapport » avec le signal acoustique. En pratique, cette approximation rend la traduction dépendante de la transcription, mais d'un ensemble de transcriptions que nous qualifierons de « plausibles », et non plus d'une transcription particulière comme dans l'approche linéaire :

$$\arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{a}) \simeq \arg\max_{\mathbf{t}} \sum_{\mathbf{s}\in\mathcal{S}} P(\mathbf{a}|\mathbf{s})P(\mathbf{s}|\mathbf{t})P(\mathbf{t}) \tag{1.5}$$

Déterminer l'ensemble $\mathcal{S}$ des transcriptions plausibles et leurs probabilités sera le rôle du module de reconnaissance automatique et des mesures de confiance. Si le modèle acoustique et le

système de transcription automatique étaient parfaits, $\mathcal{S}$ ne contiendrait qu'une seule hypothèse (plus les éventuels homophones) : la transcription correcte. En pratique, il peut en contenir plusieurs milliers. Le module de traduction, quant à lui, se chargera des modèles de traduction et de langue cible, ainsi que du calcul de la somme et l'argmax. Une des difficultés du problème sera de trouver un ensemble $\mathcal{S}$ contenant bien l'ensemble des transcriptions plausibles, et une représentation d'icelui permettant un calcul efficace de la somme.

Les auteurs de [Ney 99, Ney 04] suggèrent, mais sans les implémenter, des méthodes faiblement couplées de traduction de la parole à grand vocabulaire. Ils proposent des stratégies novatrices pour calculer $\sum_{\mathbf{s}} P(\mathbf{a}|\mathbf{s})P(\mathbf{s}|\mathbf{t})P(\mathbf{t})$ (équation 1.4). Une des pistes évoquées est de représenter l'ensemble $\mathcal{S}$ par une n-best list. [Sale 04] présente une expérience de traduction automatique utilisant pour représenter l'ensemble $\mathcal{S}$ les treillis fournis par un système de reconnaissance, mais seulement avec un vocabulaire limité. Il montre cependant l'intérêt de l'approche en prouvant que la présence des probabilités de transcription dans le treillis permet d'améliorer la qualité de la traduction : ainsi, le couplage même faible de la transcription et de la traduction a un intérêt pratique. Afin de diminuer la complexité computationnelle de la traduction d'un treillis, combinatoirement très lourde, [Bert 07] et [Shen 06] proposent d'utiliser non pas le graphe de mots fourni par le module de reconnaissance, mais le *réseau de confusion* qui en est une simplification. [Bert 05] montre que l'algorithme de décodage d'un réseau de confusion est une généralisation de l'algorithme de traduction d'une phrase déjà utilisé dans Pharaoh [Koeh 04]. Cet algorithme a depuis été implémenté dans Moses [Koeh 07a]. C'est cette méthode que nous utiliserons (voir section 2.2).

2

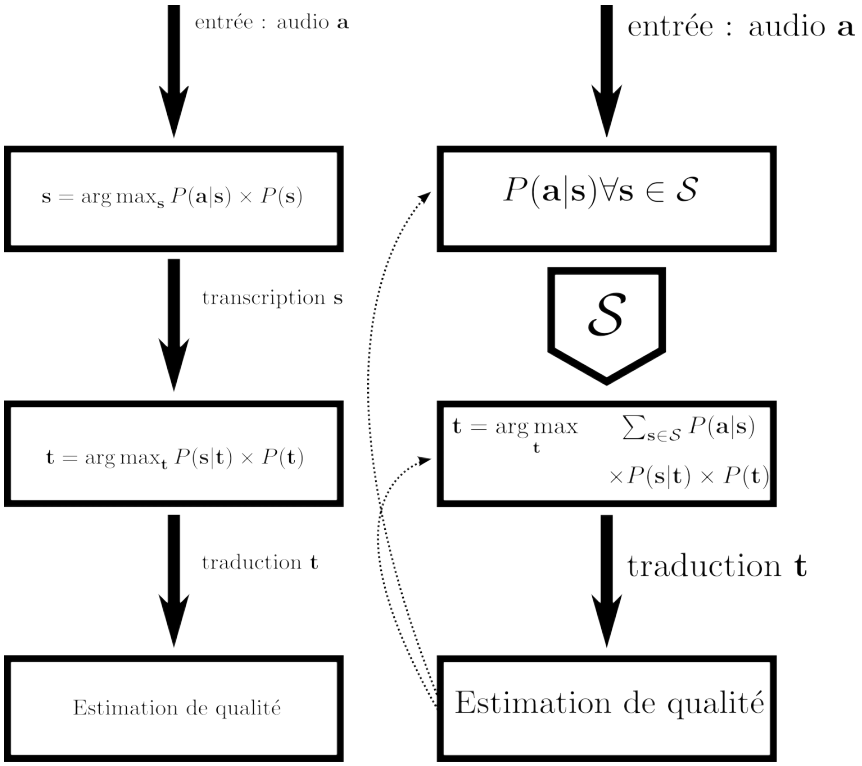
Le système s2tt

2.1 Principe de fonctionnement

Le système S2TT (**S**peech **to** **T**ext **T**ranslation system) utilise des composants standards et implémente en plus des algorithmes propres à la traduction de la parole. Le signal audio est d'abord transcrit en utilisant les bibliothèques de CMU SphinxBase et PocketSphinx. On obtient ainsi la meilleure hypothèse de reconnaissance et la séquence de phonèmes associée. Un modèle d'erreur et des mesures de confiance sont ensuite utilisés pour créer des hypothèses alternatives sous forme d'un réseau de confusion, afin d'améliorer la robustesse aux erreurs : c'est l'ensemble $\mathcal{S}$ défini dans la section 1.4. Ce réseau de confusion est ensuite traduit en utilisant la bibliothèque de Moses [Koeh 07a] pour générer une hypothèse de traduction. Afin d'évaluer ses performances, il sera comparé à un système utilisant uniquement des composants état-de-l'art, ANTS [Illi 04] et Moses (section 3.1). Nous comparerons également les résultats obtenus par ces différents systèmes avec des traductions obtenues par Google et par des interprètes, et avec les traductions de transcriptions de référence du signal audio. Cela permet d'estimer la marge de progression relative aux différents composants du système (section 3).

Le signal acoustique est transformé en hypothèses de traductions textuelles en passant par plusieurs étapes. Le système est conçu pour qu'un maximum d'informations puisse être transmis d'une étape à l'autre. Un schéma du système S2TT et du système de référence « linéaire » est donné dans la figure 2.1. Voici les différentes étapes de la transcription automatique :

1. *Segmentation du signal acoustique :* le signal lu depuis un fichier RAW ou depuis un microphone est découpé en segments se chevauchant, suffisamment courts pour être efficacement traités par le module de reconnaissance.
2. *Sélection d'un modèle adapté :* un échantillon de chaque segment est décodé à l'aide de différents modèles (femme/homme, parole téléphonique ou non). Celui qui donne le meilleur résultat en terme de vraisemblance acoustique sera utilisé pour décoder tout le segment.
3. *Transcription enrichie :* à partir de chaque segment, on calcule la meilleure hypothèse de transcription en graphèmes, les phonèmes correspondants, les évènements non verbaux (silences, bruits, ...), des informations temporelles et des mesures de confiance.
4. *Fusion des segments :* les transcriptions des segments chevauchant sont fusionnées en se basant sur les mots communs aux extrémités et les informations temporelles.
5. *Segmentation pour la traduction :* le flux ainsi obtenu est resegmenté d'une façon appropriée pour la traduction statistique (idéalement des phrases ou des propositions pas trop longues)
6. *Traduction des segments* selon un modèle adapté à la traduction de la parole.



(a) Un système de référence « linéaire »

(b) Le système S2TT

FIGURE 2.1 – Schéma d'un système de traduction linéaire (a) et de S2TT (b)

Afin de tirer le meilleur parti des microprocesseurs modernes, comptant généralement plusieurs cœurs, et d'améliorer ainsi les performances du système, les étapes de lecture et segmentation du signal, de transcription, de segmentation de la transcription et de traduction sont exécutées par des threads distincts.

2.2 Étape de reconnaissance

Le but de cette étape est de déterminer l'ensemble $\mathcal{S}$ et les probabilités $P(\mathbf{a}|\mathbf{s})$ évoqués dans l'équation 1.5. Le système de transcription automatique que nous utilisons, Pocketsphinx [20], fournit un ensemble d'hypothèses, sous la forme d'une liste n-best ou d'un graphe de mots. Cet ensemble pourrait être utilisé comme un ensemble des transcriptions plausibles. Mais ce n'est pas l'approche que nous allons suivre. Nous allons partir de la meilleure hypothèse fournie par le système, et l'ensemble $\mathcal{S}$ sera un voisinage (au sens de la distance d'édition) de cette hypothèse. Ce voisinage sera déterminé grâce à des connaissances a priori sur les erreurs commises par le système, représentées par une matrice de confusion phonétique. Ce voisinage n'étant pas déterminé par le système de reconnaissance, on peut espérer qu'il contienne certaines hypothèses qui auraient été complètement ignorées par ce dernier. Cette meilleure couverture a un prix : le voisinage étant déterminé à l'aide d'une matrice de confusion estimée a priori, le signal acoustique n'est pas pris en compte et les probabilités $P(\mathbf{a}|\mathbf{s})$ sont moins précisément calculées. Nous

20. `http://cmusphinx.sourceforge.net/`

pourrons cependant compenser partiellement cette perte de précision en utilisant des mesures de confiance.

Je vais maintenant décrire le déroulement de l'étape de reconnaissance. Le signal d'entrée est d'abord segmenté arbitrairement en segments recouvrants $\mathbf{a}_{t_k}^{t'_k}$ (dans nos expériences, des segments de dix secondes, avec trois secondes de recouvrement, c'est-à-dire $t'_k - t_k = 10s$ et $t'_k - t_{k+1} = 3s$). Un échantillon est extrait de chaque segment et décodé par quatre modèles correspondant aux différents types d'enregistrement possibles (parole téléphonique ou non, locuteur femme ou homme). Le modèle donnant le meilleur score acoustique est retenu pour décoder le segment entier. Cette méthode de sélection de modèle est très naïve, mais le but de cette implémentation n'est pas de travailler sur la sélection de modèle. On obtient ainsi, pour chaque segment, une séquence de mots et les phonèmes associés :

$$(\mathbf{w}^i = (w^i, (p_1^i, .., p_{n_{w^i}}^i)))_{i=1..l(\mathbf{a}_{t_k}^{t'_k})} \tag{2.1}$$

Notons qu'ici, « mot » est entendu au sens large : il peut aussi s'agir de séquences de mots, si le vocabulaire du système de reconnaissance en contient, mais aussi de silence, de bruit, de respiration, etc. ; dans ce cas, la liste de phonèmes associée est vide. La segmentation initiale étant arbitraire, les différentes séquences $(\mathbf{w}^i)_{i=1..l(\mathbf{a}_{t_k}^{t'_k})}$, c'est-à-dire les transcriptions des segments recouvrants successifs, sont fusionnées : le mot le plus long de la partie commune entre un segment et le suivant, qui soit le même dans les deux transcriptions, est choisi comme nouvelle frontière entre les deux segments.

On obtient ainsi une chaîne de mots et de phonèmes qui correspond à la meilleure hypothèse de transcription selon les modèles utilisés. Comme évoqué dans la section 1.4, ce n'est pas cette hypothèse qui va être « traduite », mais son voisinage. La méthode présentée dans ce qui suit est inspirée de l'article [Jian 11]. Supposons que nous disposons de l'ensemble $\mathcal{P}$ des phonèmes du langage source (un vocabulaire, en somme) et d'une matrice de confusion phonétique $\mathbf{c}$, telle que si p et q sont deux phonèmes de $\mathcal{P}$, $\mathbf{c}(p, q)$ est la probabilité observée qu'un phonème p dans la transcription automatique soit aligné au sens de Levenshtein avec un phonème q dans la transcription de référence (c'est-à-dire la probabilité que le système ait reconnu un p au lieu d'un q). Alors, chaque phonème p_k^i de la meilleure hypothèse de transcription (Équation 2.1) est remplacé par la liste :

$$((q, \mathbf{c}(p_k^i, q)))_{q \in \mathcal{P}} \tag{2.2}$$

Le procédé est illustré dans la figure 2.2. Nous avons choisi de modéliser la confusion au niveau des phonèmes et non des graphèmes (afin de simplifier le schéma, je présente ici une décomposition en syllabe, ce qui ne nuit pas à la compréhension du procédé). La raison en est que le nombre de graphèmes et la diversité des erreurs possibles sont tels qu'ils rendent impossible l'estimation d'une matrice de confusion précise. Modéliser les erreurs au niveau des phonèmes plutôt que des graphèmes permet de réduire drastiquement cette combinatoire (donc la largeur du réseau de confusion) tout en conservant la même expressivité : en effet, le nombre de phonèmes est de l'ordre de la cinquantaine, alors que la taille du vocabulaire utilisé par le système de reconnaissance est de plusieurs dizaines de milliers de mots. De plus, si un mot de la transcription automatique est erroné, il y a fort à parier qu'il partage tout de même quelques phonèmes avec le mot correct : les erreurs n'arrivent pas tout à fait au hasard. En pratique, afin que la vitesse de traduction ne soit pas trop dégradée, on réduira la largeur du réseau de confusion en

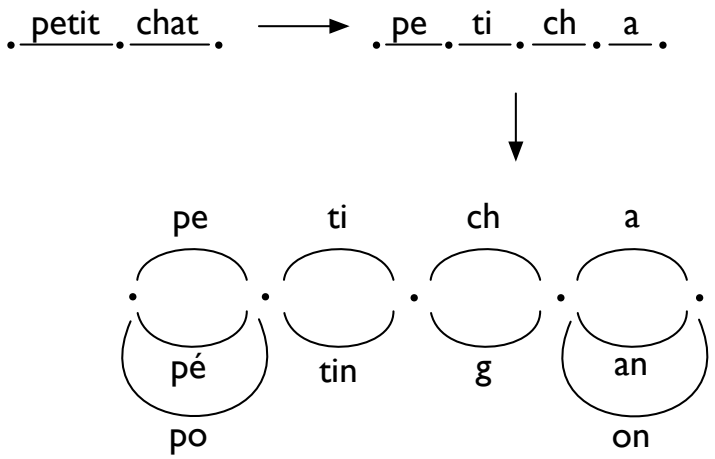


FIGURE 2.2 – Transformation d'une hypothèse de transcription en réseau de confusion phonétique (ici syllabique pour la simplicité de l'exemple).

ne considérant que les erreurs les plus probables, c'est-à-dire les phonèmes q tels que $\mathbf{c}(p_k^i, q)$ est plus grand qu'un certain seuil.

On obtient ainsi un réseau de confusion prenant en compte toutes les erreurs qu'a pu faire le système et les probabilités associées. Par commodité nous identifierons le réseau de confusion et l'ensemble $\mathcal{S}$ des hypothèses qu'il représente. Ainsi nous pourrons identifier une hypothèse $\mathbf{s} \in \mathcal{S}$ et un chemin dans le réseau de confusion. Dans notre exemple, le locuteur a donc pu prononcer `petit chat` (hypothèse $\mathbf{s}^1$), qui est la meilleure hypothèse selon le module de reconnaissance, ou peut-être le nom de l'île indonésienne `pétiga` (« phonèmes » `pé ti g a`, hypothèse $\mathbf{s}^2$), ou peut-être parlait-il du menhir des `pétigons` (`pé ti g on`, hypothèse $\mathbf{s}^3$) dans l'arrière pays cannois. Si la construction d'un réseau de confusion phonétique permet de représenter toutes les hypothèses plausibles de reconnaissance, il n'en reste pas moins que la meilleure hypothèse reste celle fournie par le module de reconnaissance et qu'il n'existe à ce stade aucune raison de lui en préférer une autre, à supposer que tout phonème a plus de chance d'être correctement reconnu que d'être confondu avec n'importe lequel des autres phonèmes du « vocabulaire » (dit autrement, nous ne disposons pas à ce stade d'information permettant de détecter une erreur et on devrait en théorie toujours choisir l'hypothèse initiale). En pratique, les matrices de confusion que nous avons estimées vérifiaient toujours cette propriété. Formellement cette condition s'écrit :

$$\forall p, q \in \mathcal{P}_S, p \neq q \Rightarrow \mathbf{c}(p, p) > \mathbf{c}(p, q)$$

Dans notre exemple, cela signifie que :

$$\begin{aligned} P(\mathbf{a}|\text{pe ti ch a}) &\geqslant P(\mathbf{a}|\text{pé ti g a}) \\ &\geqslant P(\mathbf{a}|\text{pé ti g on}) \end{aligned}$$

En cas d'erreur de reconnaissance, c'est en fait le contexte qui permettra peut-être, au moment de la traduction, de préférer un autre chemin dans le réseau de confusion. Supposons par exemple que le début des hypothèses de transcription $\mathbf{s}^1$, $\mathbf{s}^2$ et $\mathbf{s}^3$ soit :

`l i l d e` (*l'île de*)

Si nous traduisions directement la meilleure hypothèse, ces trois transcriptions auraient sans doute des traductions différentes, respectivement $\mathbf{t}^1$, $\mathbf{t}^2$ et $\mathbf{t}^3$. `l i l d e pe ti ch a`, transcription « phonétique » de *L'île de petit chat*, n'existant pas, ce segment n'est probablement pas présent dans la table de traduction (rappelons que la traduction se fait de la représentation phonétique de la langue source vers la représentation graphémique de la langue cible). Le score $P(\text{l i l d e pe ti ch a}|\mathbf{t}^1) \times P(\mathbf{t}^1)$ de l'hypothèse de traduction est donc vraisemblablement assez faible. Il en va de même de $P(\mathbf{s}^3|\mathbf{t}^3) \times P(\mathbf{t}^3)$. En revanche le score $P(\text{l i l d e pé ti g a}|\mathbf{t}^2) \times P(\mathbf{t}^2)$ de la traduction de *l'île de Pétiga* est sans doute plus élevé. Il est donc possible d'avoir :

$$\begin{aligned}\sum_{\mathbf{s}\in\mathcal{S}} P(\mathbf{a}|\mathbf{s}) \times P(\mathbf{s}|\mathbf{t}^2) \times P(\mathbf{t}^2) &> \sum_{\mathbf{s}\in\mathcal{S}} P(\mathbf{a}|\mathbf{s}) \times P(\mathbf{s}|\mathbf{t}^1) \times P(\mathbf{t}^1) \\ &> \sum_{\mathbf{s}\in\mathcal{S}} P(\mathbf{a}|\mathbf{s}) \times P(\mathbf{s}|\mathbf{t}^3) \times P(\mathbf{t}^3)\end{aligned}$$

Ainsi, la traduction $\mathbf{t}^2$ peut être retenue. Tout se passe comme si la transcription avait été corrigée grâce au contexte, au lieu de simplement choisir la meilleure hypothèse proposée par le système de reconnaissance.

2.2.1 Estimation de la matrice de confusion

La matrice de confusion utilisée est simplement estimée en comptant les erreurs commises par le système de reconnaissance sur un corpus audio pour lequel on dispose d'une transcription de référence. Les erreurs devant être comptées au niveau des phonèmes, il faut d'abord phonétiser la transcription de référence (si PocketSphinx fournit la phonétisation correspondant à la transcription calculée, les transcriptions de référence sont données au niveau des graphèmes et non des phonèmes).

J'ai donc développé un programme de phonétisation utilisant BDLEX [Calm 98] et les règles régissant les liaisons en français [21]. Les règles de prononciations étant ambigües, ce programme génère l'ensemble des prononciations possibles. Si un des phonèmes possibles est aligné (au sens de Levenshtein) avec le phonème effectivement fourni par PocketSphinx, celui-ci est considéré correct. Si le phonème p fourni par PocketSphinx est aligné avec le phonème q de la transcription de référence (une substitution, qu sens de Levenshtein), on compte une erreur (p, q). Si p est aligné avec plusieurs alternatives de prononciation $(q_1, .., q_n)$, on compte $\frac{1}{n}$ erreur pour chaque alternative possible. Les phonèmes manquants (suppression) ne sont pas pris en compte. Si p n'est aligné avec aucun phonème de la phonétisation de référence (insertion), on considère qu'il est aligné avec la *transition vide* ϵ et on compte une erreur. On obtient donc la probabilité $P(\epsilon|p)$ que le phonème p de la transcription automatique soit un artefact, un bruit, et qu'il faille donc le supprimer. On peut donc ajouter dans le réseau de confusion une ϵ-transition, c'est-à-dire un arc porteur du symbole spécial ϵ, qui a la propriété suivante : $\forall x, y \in \mathcal{P}_S, x\epsilon y = xy$. En pratique cependant, on obtient souvent des réseaux de confusion contenant un grand nombre d'ϵ-transition. Dans ce cas, lors de l'étape de traduction (section 2.4, page 102), le décodeur (Moses, dans nos expérience) avait tendance à favoriser fortement ces arcs, produisant des traductions très courtes et lacunaires. En effet, le coût de ces arcs est fortement sous-estimé car il ne dépend que de $P(\epsilon|p)$, et pas des modèles de langue et de traduction. Nous avons donc préféré les

21. `http://www.etudes-litteraires.com/regles-de-liaison.php`, `http://fr.wikipedia.org/wiki/Liaison_en_français`

supprimer, mais il conviendra dans le futur de déterminer une façon d'estimer correctement le coût de ces arcs.

2.2.2 Utilisation de mesures de confiance

Afin d'affiner les estimations de probabilité *a priori* d'erreur fournies par la matrice de confusion, nous allons utiliser des mesures de confiance sur la phrase *source* (et non sur l'hypothèse de traduction comme dans la partie II). Ce développement étant tout à fait expérimental, nous n'avons utilisé, afin de limiter la complexité de l'implémentation et des tests, qu'un seul paramètre prédictif. Nous avons choisi celui qui donne isolément les meilleurs résultats, le comportement de repli d'un modèle de langage n-gramme (partie II, section 3.1.1, page 54).

Rappelons d'abord que la matrice de confusion phonétique fournit une probabilité de confusion $P_c(q|p)$ *a priori*, c'est-à-dire ne dépendant pas de l'instance considérée, et que $P_c(p|p)$ peut être considéré comme une mesure de confiance *a priori*, car elle ne tient pas compte du reste de l'hypothèse ni du signal acoustique. Afin d'améliorer la précision, nous allons combiner cette estimation *a priori* avec une estimation *a posteriori* fournie par une mesure de confiance au niveau mot de la phrase *source* $P(C(\mathbf{S},i)=1)$ (en utilisant des notations similaires à celles de l'équation 1.4 page 38. Notamment, $C(\mathbf{S},i)$ est la variable aléatoire représentant la correction du i-ème mot de $\mathbf{S}$). Si p est un phonème de la prononciation du mot s_i de l'hypothèse de transcription $\mathbf{s}$, la probabilité de confusion devient alors :

$$P(q|p) \quad = \quad P(C(\mathbf{s},i)=1) \times P(q|p,C(\mathbf{s},i)=1) + P(C(\mathbf{s},i)=0) \times P(q|p,C(\mathbf{s},i)=0)$$

Avec :

$$P(q|p,C(\mathbf{s},i)=1) \quad = \quad \begin{cases} 0 & \text{si} \quad q \neq p \\ 1 & \text{sinon} \end{cases}$$

$$P(q|p,C(\mathbf{s},i)=0) \quad = \quad \begin{cases} P(p \text{ est faux}|p,C(\mathbf{s},i)=0) \times \frac{P_c(q|p)}{1-P_c(p|p)} & \text{si} \quad q \neq p \\ P(p \text{ est juste}|p,C(\mathbf{s},i)=0) & \text{sinon} \end{cases}$$

En utilisant ces formules, on peut donc produire un réseau de confusion contenant des probabilités plus précises que les simples P_c. Cependant, nous n'avons pas de méthode pour estimer les paramètres $P(p \text{ est faux}|p,C(\mathbf{s},i)=0)$ et $P(p \text{ est juste}|p,C(\mathbf{s},i)=0)$. Pour les besoins des expériences préliminaires, nous avons supposé qu'un phonème d'un mot erroné avait une chance sur deux d'être correct, mais cette valeur est arbitraire. Une analyse plus précise des résultats de la transcription permettrait d'affiner cette estimation. Enfin, pour des raisons de complexité computationelle, nous n'avons pas pu traduire l'intégralité de notre corpus de test en utilisant cette méthode. Le système S2TT demande manifestement encore des efforts de développement. Nous présenterons donc seulement quelques exemples de transcriptions (graphémique, pour la clarté de l'exemple, même si le modèle de traduction considère une représentation phonétique de la source) et de traductions en matière de preuve de concept :

Texte prononcé	Texte reconnu	Traduction proposée	Traduction de référence
Face à vous Marie-Claude Bonneville secrétaire de l' association « Souriez vous êtes filmés »	`face à vous marie-claude bonne ville secrétaire de l'association souriez vous êtes filmés`	`face you mary good secretary of the association mice and footage you are`	Opposite you, Marie-Claude Bonneville, secretary of the association "Smile - You're On Camera"
levallois une ville surveillée par des caméras depuis une vingtaine d'années maintenant	`levallois une ville surveillé par des caméras depuis une vingtaine d'années minces`	`a city law monitored by cameras since their twenties slim`	Levallois, a town with camera surveillance for the last twenty years

Il est bien sûr délicat de commenter deux exemples de traduction. On observe néanmoins que le système est, bien sûr, sensible aux erreurs de reconnaissance. On observe aussi que la traduction par phonèmes induit certaines erreurs qui n'auraient pas eu lieu avec un système classique. Par exemple, `souriez` se prononce comme `souris et` et est traduit par `mice and`. Cependant, une telle traduction ne devraient pas être retenue par un modèle bien entraîné, car elle ne colle pas du tout avec le contexte. Nous pensons qu'une plus longue phase de développement résoudra ce problème.

2.2.3 Modèles pour la transcription

Les modèles acoustiques utilisés pour l'étape de transcription ont été entraînés sur le corpus distribué pour la campagne ESTER-2 (tableau 2.1) en utilisant le programme SphinxTrain [22]. L'ensemble de transcriptions plausibles $\mathcal{S}$ (équation 1.5) étant déterminé par exploration autour d'une hypothèse optimale, nous utiliserons tout de même un modèle de langage du français pour déterminer cette dernière. Ainsi, nous pouvons, pour nos expériences, suivre la méthode standard et éprouvée de transcription automatique, mais il serait bon de mener ces expériences en implémentant rigoureusement la méthode proposée dans la section 2.2, figure 2.1 (b), c'est-à-dire en utilisant le système de reconnaissance automatique pour déterminer un ensemble de transcriptions plausibles (l'ensemble $\mathcal{S}$). Le modèle de langage du français est un modèle trigramme utilisant la méthode de lissage de Kneser-Ney améliorée [Knes 95], entraîné sur des articles des journaux Le Monde et l'Humanité (580M mots au total), et des transcriptions d'émissions radiophoniques, fournis pour la campagne ESTER-2 [23] [Grav 04] totalisant 130M mots. L'outil utilisé pour l'entraînement du modèle est SRILM [Stol 02].

Reconnaissance	Audio (heures)	Mots	
Entraînement	178	800k	
Développement	5	40k	
Modèle de langue français	-	700M	
Traduction	Paires de phrases	Mots	
		Français	Anglais
Entraînement	1.7M	53M	48M
Modèle de langue anglais	1.6M	-	45M
Développement	3K	86K	77K

Table 2.1 – Tailles des corpora utilisés.

22. http ://cmusphinx.sourceforge.net/
23. http ://www.afcp-parole.org/camp_eval_systemes_transcription/index.html

2.3 Segmentation pour la traduction

La resegmentation des hypothèses du module de reconnaissance est une tâche complexe mais d'une importance capitale pour la traduction [Mats 07]. En effet, les modèles probabilistes que nous utilisons, les modèles IBM [Brow 93], reposent sur la notion de phrase. Il est donc important de segmenter les hypothèses de reconnaissance en séquences de mots aussi semblables que possibles à des phrases bien construites. Or, l'outil que nous utilisons pour entraîner les modèles, Giza++ [Och 03], impose une borne sur la longueur des séquences de mots (cent mots pour les dernières campagnes WMT), et le coût de calcul de la traduction augmente rapidement avec la longueur des phrases, alors que les phrases sont parfois très longues. Mais surtout, la notion même de phrase n'est pas pertinente à l'oral : Christophe Benzitoun affirme par exemple dans [Benz 04] que dans une transcription, « la phrase est au mieux une approximation graphique, un compromis entre structure syntaxique, intonation et mise en page ». L'expression orale est selon lui structurée par ses phénomènes propres, souvent plus complexes qu'à l'écrit. Claire Blanche-Benveniste affirme dans [Blan 87], exemples expérimentaux à l'appui, qu'il n'existe parfois pas de transcription univoque d'un « segment » oral et plaide pour le développement de méthodes de transcription nouvelles conçues pour préserver toute la subtilité de l'expression orale, et de modèles spécifiques la décrivant.

Nous n'allons donc pas tenter d'obtenir une hypothétique segmentation en « phrases ». Nous avons développé une méthode pour extraire du flux de parole reconnu des segments plus petits, aussi cohérents que possibles sémantiquement, afin qu'ils soient traduits rapidement et sans difficulté par le système Moses.

Le système de reconnaissance que nous utilisons utilise les silences et les changements de locuteur pour segmenter la transcription. Si un changement de locuteur indique en général la fin d'un segment sémantiquement cohérent (mais pas toujours), il en va autrement des silences. Le locuteur peut reprendre sa respiration ou chercher ses mots au milieu d'une « phrase » (je continuerai à parfois employer ce terme pour la commodité de la description), ou au contraire en enchaîner plusieurs sans pause. Une segment cohérent peut également être laissé en suspens, et parfois, la distinction et la frontière entre deux segments et ambigüe.

L'algorithme implémenté pour segmenter la transcription automatique d'une façon appropriée pour la traduction se déroule comme suit (algorithme 2 page 101) : on choisit d'abord une longueur maximale pour les segments L_{max} qui sera fixée à 40 dans nos expériences (c'est-à-dire la valeur utilisée pour entraîner le système de référence de la campagne WMT09). Ensuite, on cherche une position i minimale parmi les L_{max} premiers éléments de la transcription tel que le modèle de langage (un modèle n-gramme d'ordre trois) prédise une fin de phrase ou un signe de ponctuation comme « , », « : », « ; » etc. à la position $i+1$ (pour la simplicité de l'explication, on dira plus loin qu'il génère le symbole `</s>`). Si une telle position existe, alors on extrait le segment $w_1, .., w_i$. Sinon, on cherche un silence. Si un silence existe en position i, le segment $w_1, .., w_i$ est extrait. À défaut, on détermine la position $i \in \{1, .., L_{max}\}$ maximisant la probabilité d'une fin de phrase selon le même modèle de langage et on extrait le segment correspondant. On recommence ensuite avec le reste du flux, tant qu'il n'est pas vide. L'algorithme est formalisé dans le pseudo-code 2. P est une distribution de probabilité donnée par un modèle trigramme du langage source, et $\mathcal{V}^S$ le vocabulaire de la langue source.

Algorithme 2: Segmentation des hypothèses de transcription.

begin
- ÉTAT INITIAL : $\mathbf{w} = w_1, .., w_N$; segment = $\emptyset$
- $L \leftarrow min(N, L_{max})$
- $i \leftarrow generates\_eos(w_1, .., w_L)$
- **if** $i \geqslant 1$ **then**
  - $\mathbf{w} \leftarrow w_{i+1}, .., w_N$
  - $segment \leftarrow w_1, .., w_i$
  - **return** *segment*
- $i \leftarrow is\_sil(w_1, .., w_L)$
- **if** $i \geqslant 1$ **then**
  - $\mathbf{w} \leftarrow w_{i+1}, .., w_N$
  - $segment \leftarrow w_1, .., w_{i-1}$
  - **return** *segment*
- **if** $L = L_{max}$ **then**
  - $i \leftarrow \arg\max_{1 \leqslant j \leqslant L} P(< /s > | w_{j-3}, .., w_j)$
  - $\mathbf{w} \leftarrow w_{i+1}, .., w_N$
  - $segment \leftarrow w_1, .., w_i$
  - **return** *segment*
- **else**
  - attendre jusqu'à ce que $N \geqslant L_{max}$ ou jusqu'à ce que tout le signal ait été transcrit
  - aller à **begin**
- **return**

`generates_eos` :

begin
- Entrée : $w_1, .., w_n$
- **for** *i=1..n* **do**
  - **if** $\arg\max_{w \in \mathcal{V}^S} P(w | w_{i-3}, .., w_i) =< /s >$ **then**
    - **return** *i*
- **return** *0*

`is_sil` :

begin
- Entrée : $w_1, .., w_n$
- **for** *i=1..n* **do**
  - **if** w_i=<*SILENCE*> **then**
    - **return** *i*
- **return** *0*

2.4 Étape de traduction

Les modèles de traduction, quant à eux, sont des modèles IBM-5 entraînés à l'aide du logiciel Giza++ [Och 03] sur un bitexte composé de 1.6M paires de phrases (50M mots) extraites du corpus EUROPARL [Koeh 05]. Des détails sur la méthode d'entraînement sont disponibles sur `http://statmt.org/wmt09/baseline.html`. Afin d'adapter le modèle pour la traduction de programmes d'information radiophonique, l'étape de développement (*tuning*) a été réalisée sur 3000 paires de phrases (environ 80k mots) du corpus d'information distribué pour la campagne WMT 2009. Le tableau 2.1 (page 99) récapitule les caractéristiques des corpora.

Une fois les modèles entraînés, les tables de traduction graphémiques ont été adaptées pour permettre la traduction d'une transcription phonétique comme décrite dans la section précédente. Pour cela, on utilise la méthode déjà décrite utilisant un dictionnaire de prononciation et des règles afin de déterminer les variantes. Le côté source des tables (les séquences de mots en français) est phonétisé, le côté cible, quant à lui, demeurant sous forme de graphèmes afin d'obtenir une traduction immédiatement lisible. La distorsion maximale est multipliée par le nombre moyen de phonèmes dans un mot.

Les mesures de confiance décrites dans la partie II ont été implémentées dans le système, et l'utilisateur a la possibilité d'en utiliser une ou plusieurs (via un réseau de neurones également implémenté dans le système grâce à la bibliothèque FANN) pour vérifier la correction des traductions proposées. Le système fournit ainsi *in fine* un fichier xml contenant la liste des segments reconnus dans la phrase source, leur transcription phonétique, les traductions proposées et des indications de correction aux niveaux des mots et des phrases.

3

Évaluation

3.1 Système de référence

S2TT est comparé avec un système linéaire basé sur ANTS, le système de traduction automatique développé par l'équipe Parole du LORIA [Illi 04, Brun 05], basé sur JULIUS [Lee 01], et Moses. Je vais décrire très brièvement ici le fonctionnement de ANTS. Pour plus de détails, le lecteur se rapportera aux articles cités précédemment.

ANTS segmente les données audio selon la source de la parole (enregistrement en studio ou par téléphone), le genre du locuteur, l'apparition de passages musicaux et les silences. Les segments sont ensuite redécoupés en segments chevauchant de durée maximale six secondes. Ces segments sont ensuite transcrits en utilisant JULIUS [Lee 01], puis les transcriptions des segments chevauchant sont fusionnées en utilisant les mots communs entre les segments.

ANTS utilise trois modèles acoustiques pour la reconnaissance (et des modèles additionnels pour la classification du signal) : deux pour les passages enregistrés à un taux de 16kHz (en studio), selon le sexe du locuteur (femme ou homme) et un pour les passages enregistrés à 8kHz (enregistrements par téléphone). On détermine le modèle le plus adapté en mettant en compétition les différents modèles acoustiques, et on retient celui qui donne les meilleurs résultats en termes de vraisemblance acoustique, avec une contrainte de régularisation imposant que l'alternance des modèles ne soit pas trop rapide et donc irréaliste. Les paramètres de ces modèles sont estimés sur 90 heures d'enregistrements d'émissions radiophoniques de France-Inter, France-Info, RFI et RTM, distribués dans le cadre de la campagne ESTER [Gall 06]. Le corpus d'entraînement des modèles de langage est composé des archives des journaux *Le Monde* et *L'Humanité*, soit 580M mots au total, et des transcriptions d'informations radiodiffusées (corpora *ESTER* et *TNS*) totalisant 130M mots.

La transcription automatique se fait en deux passes : un treillis est construit en utilisant un modèle bigramme lors de la première passe, et un modèle quadrigramme à rebours est utilisé pour déterminer la meilleure hypothèse lors de la seconde passe. Le lexique comprend 127000 prononciations pour 63000 mots. ANTS a été classé en quatrième position lors de la campagne d'évaluation ESTER2 [Illi 04, Gall 06].

Les meilleures hypothèses sont ensuite traduites avec Moses, en utilisant les mêmes modèles pour le système S2TT (section 2.4). On obtient ainsi un système linéaire (section 1.2) entièrement basé sur des composants standards de l'état de l'art (hormis le système de segmentation, pour lequel aucune méthode standard de référence ne semble se dégager).

Système	WER	B-1	B-2	B-1&2
S2TT	25.1	17.9	24.3	29.5
S2TT-1best	25.1	18.4	25.7	30.6
S2TT-Google	25.1	19.4	34.4	38.0
ANTS-Moses	22.3	19.3	27.3	32.3
Ref-Moses	0	23.6	34.1	40.5
Ref-Google	0	24.3	48.2	51.7
Expert-1	0	-	34.8	-
Expert-2	0	30.6	-	-

TABLE 3.1 – Évaluation de la qualité de traduction en fonction des systèmes.

3.2 Corpus de test

Une fraction de ce corpus (252 phrases, soit 35'39" d'enregistrement) a été traduite en anglais par un interprète professionnel, dont la langue maternelle est l'anglais (*Expert-1*), et par une professeur d'anglais, dont la langue maternelle est le français (*Expert-2*), qui a utilisé *The Corpus of Contemporary American English*[24] comme outil d'aide à la traduction. Cette partie servira de corpus de test. La version des traductions réalisée par *Expert-1* sera appelée R-1, celle réalisée par *Expert-2* sera appelée R-2. Il y a de grandes différences de style et de vocabulaire entre R-1 (littéraire) et R-2 (proche de la source), avec des conséquences notables sur l'évaluation. Les transcriptions de référence et les traductions automatiques et de référence sont présentées en annexe C page 129.

3.3 Résultats

Le tableau 3.1 présente les résultats obtenus avec différents systèmes. Outre l'évaluation des performances de S2TT, ce tableau nous permet d'estimer l'influence des différents composants sur les performances, et les marges de progression liées aux uns et aux autres :

- **S2TT :** le système S2TT tel que présenté plus haut.
- **S2TT-1best :** le système S2TT tel que présenté plus haut, mais ce n'est pas le réseau de confusion phonétique qui est traduit mais la meilleure hypothèse de reconnaissance.
- **ANTS-Moses :** un système de référence linéaire réalisé à partir de composants état-de-l'art. L'utilisation de ANTS permet de mesurer l'influence d'une légère amélioration de la reconnaissance.
- **Ref :** les transcriptions de références, donc l'équivalent d'une étape de reconnaissance sans erreur. Utilisé pour estimer la progression atteignable en améliorant l'étape de transcription.
- **Google :** considéré comme un excellent système de traduction automatique. Utilisé pour estimer la marge de progression disponible en améliorant l'étape de traduction.

Les scores BLEU sont calculés par le script `multi-bleu.perl` fournit avec Moses. B-1 est le score bleu calculé en prenant pour référence R-1, B-2 en prenant pour référence R-2, et B-1&2 en prenant pour référence R-1 et R-2 en même temps. Ces résultats appellent certains commentaires : le système S2TT utilise pour la phase de reconnaissance les mêmes modèles que ANTS, mais n'a pas bénéficié d'une aussi longue phase d'optimisation. Le taux d'erreur

24. http ://corpus.byu.edu/coca/

de reconnaissance est donc légèrement moins bon. S2TT-1best obtient des résultats légèrement meilleurs que S2TT, ce qui suggère que les probabilités de confusion ne sont pas bien estimées par la méthode exposée dans la partie 2.2. Sans doute notre méthode de phonétisation n'est-elle pas assez précise. De plus, les probabilités sont estimées *a priori*, c'est-à-dire sans tenir compte des caractéristiques de l'instance (qualité du signal, complexité de la phrase reconnue, estimation de confiance sur l'hypothèse de transcription...), ce qui est une lacune évidente.

La ligne *S2TT-Google* présente les résultats obtenus en traduisant avec l'outil Google Translate les hypothèses de reconnaissances générées par S2TT, après resegmentation automatique (Section 2.3). Cela donne une idée de la progression réalisable en améliorant le système de traduction (environ 8 points).

Les lignes *Ref-Moses* et *Ref-Google* présentent l'évaluation de la traduction obtenue utilisant non pas les hypothèses de transcription automatique mais les transcriptions de référence (donc un taux d'erreur de 0%), traduites respectivement avec Moses et Google. Cela permet d'estimer la marge de progression propre au système de reconnaissance (environ 10 points de BLEU entre *S2TT* — qui utilise Moses — et *Ref-Moses*).

Enfin, il est intéressant de noter que les scores B-1 sont bien plus bas que les scores B-2. En effet, les traductions R-1 sont plus longues d'environ 10%, et les tournures plus idiomatiques, que dans R-2, qui respecte plus la structure de la phrase originale et utilise un vocabulaire plus standard. On observe d'ailleurs que le score bleu de R-1 évalué à l'aune de R-2, et celui de R-2 évalué à l'aune de R-1, sont très bas, sachant qu'ils sont tous les deux des traductions réalisées par des humains.

4

Conclusion

J'ai présenté dans cette partie le système S2TT (*Speech To Text Translation*), un prototype de système intégré de traduction de la parole spontanée à grand vocabulaire. La traduction de la parole pose des défis particuliers (rappelés dans le chapitre 1) : en effet, l'oral n'est pas de l'écrit lu, et est régi par ses règles propres, souvent plus complexes que celles de l'écrit. La notion de phrase notamment, essentielle aux modèles de traduction statistiques de la langue écrite, n'y est pas pertinente. Les constructions syntaxiques sont plus complexes (on retrouve souvent des constructions qu'on pourrait appeler « à tiroirs », c'est-à-dire en enchaînement de références très long), et le discours est structuré par des évènements particuliers (prosodie, pauses, reprises, interpellations, disfluences, raclements de gorge, gestes, etc.). Basé sur des composants standards (Sphinx et Moses) initialement développé pour la traduction de l'écrit, S2TT implémente des techniques spécifiques à la traduction de la parole : prise en compte de l'incertitude au niveau de la reconnaissance grâce aux réseaux de confusion phonétiques (section 2.2) et aux mesures de confiance (section 2.2.2), et segmentation du flux de parole reconnu de façon adaptée à la traduction (section 2.3). Évalué sur une tâche de traduction d'émissions radiophoniques, il obtient *in fine*, avec un score BLEU d'environ 30, des résultats comparables à ceux d'un système de référence linéaire (c'est-à-dire composé d'un système de reconnaissance, qui fournit une hypothèse unique, brute, de transcription, et d'un système de traduction, qui traduit cette hypothèse), mais légèrement moins bons (celui-ci obtient un score BLEU de 32 — tableau 3.1). J'attribue cette différence à la qualité très élevée du système de reconnaissance de référence, qui est développé au sein de notre équipe depuis des années et a donc bénéficié d'une mise au point particulièrement approfondie qui fait défaut à celui du système S2TT. Néanmoins l'écart est faible et ces résultats nous encouragent à poursuivre dans cette voie. L'algorithme de segmentation, notamment, pourrait être amélioré, en utilisant par exemple un *Hidden Events Language Model* modélisant explicitement les phénomènes de l'oral tels que les pauses, et l'estimation des probabilités de confusion phonétique bénéficierait sans doute d'une amélioration des mesures de confiance. Ces deux axes sont, à mon sens, des pistes prioritaires de recherche.

Conclusion

Je vais dans cette conclusion reprendre le fil conducteur de la thèse, les contributions apportées, les résultats obtenus et les pistes qui restent à suivre. J'ai présenté dans ce manuscrit des recherches visant à mettre au point un système de traduction automatique de la parole spontanée à grand vocabulaire utilisable par le plus grand nombre. Il est nécessaire pour cela de relever des défis techniques considérables (reconnaissance automatique de la parole spontanée — partie III, sections 1.1 page 86 et 2.2 page 94, traduction automatique de la parole — partie III, sections 2.3 page 100 et 2.4 page 102, robustesse aux erreurs — partie III, section 2.2 page 94), mais aussi d'apporter un soin particulier à l'ergonomie du système, notamment afin d'avertir l'utilisateur, qui ne maîtrise pas forcément la langue source, d'erreurs potentielles de traduction, en utilisant des mesures de confiance (partie II page 35).

C'est à ce dernier point, les mesures de confiance, que je me suis attelé en premier lieu. J'ai opté pour le paradigme de l'apprentissage automatique. Dans cette approche, un certain nombre de mesures, appelées « paramètres prédictifs », sont effectuées pour obtenir autant d'informations que possible sur la traduction proposée, sous forme numérique donc aisément accessible à la représentation statistique. Des techniques de classification et de régression sont ensuite utilisées pour obtenir une estimation de la qualité de la traduction ou de la probabilité de sa correction, permettant une diminution de 1,3 points du taux d'erreur de classification des mots et de 6 points pour la classification des phrases, par rapport au meilleur paramètre prédictif utilisé seul (partie II, section 5.4, page 75). Les expériences menées dans ce cadre, notamment lors de la campagne WMT 2012 (partie II, chapitre 5, page 73) ont pouvé la robustesse de cette approche et la capacité du système ainsi développé à tirer partie de nouveaux paramètres prédictifs pour améliorer les estimations de confiance. En revanche, l'expérience de post-édition que j'ai menée (partie II, chapitre 4), a montré que les estimations de confiance n'étaient pas assez précises pour être utiles à des traducteurs. Certaines erreurs commises par le système sont particulièrement gênantes pour eux, et provoquent frustration et perte de temps plutôt qu'un gain d'efficacité. Je pense également qu'il est important de développer des mesures de confiance à un niveau intermédiaire entre les mots et les phrases, qui corresponde à une construction linguistique plus pertinente. En effet, les mots ne sont pas traduits isolément les uns des autres, et on peut souvent distinguer au sein d'une phrase différentes séquences qui ne sont pas forcément toutes correctes ou toutes incorrectes. Une première approche serait d'utiliser la segmentation proposée par Moses. En effet, les modèles de traduction de l'état de l'art sont basés sur des séquences de mots et non des mots isolés (partie I, section 1.5.3, page 11), formant une partition de la phrase source et de l'hypothèse. De nombreuses partitions différentes sont possibles, mais lors de la traduction, il est possible de connaître celle qui a été retenue par le décodeur. À plus long terme, un algorithme de segmentation, tel que celui proposé dans la partie III, section 2.3, page 101, mais plus précis, pourrait être utilisé pour proposer des segments ayant du sens pour un utilisateur humain. Au cours de plusieurs discussions avec des membres de la communauté des chercheurs et utilisateurs de traduction automatique, cet axe de recherche m'est apparu comme prioritaire.

Ces mesures de confiance ont ensuite été appliquées à la traduction de la parole. Un prototype de traduction automatique de la parole spontanée à grand vocabulaire, le système *Speech-To-Text Translation* (S2TT), a été développé (partie III). Basé sur des composants standards (Sphinx, pour la transcription automatique de la parole, et Moses, pour la traduction de texte), il intègre des techniques originales développées pour le problème particulier qu'est la traduction de la parole spontanée. Notamment, une méthode de segmentation automatique est développée, qui vise à produire des segments adaptés à la traduction automatique sans forcément chercher à imiter la structure de phrases qui est propre à l'écrit (partie III, section 2.3, page 100). De plus, S2TT effectue la traduction d'un réseau de confusion phonétique, et non pas de la

meilleure hypothèse de transcription, afin de prendre en compte l'ambiguité et l'incertitude de la transcription automatique (partie III, section 2.2, page 94). Il intègre également des mesures de confiance au niveau de la reconnaissance pour améliorer encore la robustesse, et au niveau de la traduction pour fournir des probabilités de correction à l'utilisateur. Les expériences effectuées sur la traduction d'émissions radiophoniques montrent que S2TT obtient de bonnes performances (un score BLEU d'environ 30), quoique légèrement moins bonnes qu'un système plus classique et non spécifique à l'oral (score BLEU d'environ 32), mais dont le module de reconnaissance a bénéficié d'un développement plus long. S2TT reste un prototype, et certaines fonctionnalités, notamment l'utilisation de mesures de confiance sur l'hypothèse de transcription dans le réseau phonétique, demandent un effort de développement supplémentaire avant d'être réellement utilisables. L'étape de reconnaissance est également particulièrement gourmande en ressources, les modèles demandent à être améliorés et leurs paramètres affinés, et le besoin de modèles de langue spcifiques à l'oral et ses phénomènes propres se fait pressant ; l'estimation de la matrice de confusion phonétique (partie III, section 2.2.1, page 97) pourrait être plus précise. Le coût d'une ϵ-transition doit être estimé plus rigoureusement, et il faut mettre au point une méthode pour tenir compte des phonèmes manquants. Le phonétiseur pourrait également être amélioré : à place d'un outil *ad hoc*, un système état de l'art pourrait être utilisé, qui permette de désambigüiser les variantes de prononciations, améliorant ainsi l'estimation des probabilités de confusion. À moyen terme, le modèle proposé dans partie III, section 2.2, page 94 devra être implémenté rigoureusement, c'est-à-dire en utilisant le système de reconnaissance automatique pour proposer un ensemble d'hypothèses de transcription plausibles sans s'appuyer sur une hypothèse particulière. Une première piste consiste à utiliser le treilli de mots ou la n-best list fournie par le module de reconnaissance, quitte à l'élaguer ou à la simplifier. Cette approche se heurte pour l'instant à des limites computationnelles, dont il ne fait nul doute qu'un développement soigneux viendra à bout. Des efforts peuvent égalemement consacrés à l'algorithme de segmentation, en utilisant par exemple un *Hidden Events Language Model* modélisant explicitement les phénomènes de l'oral tels que les pauses.

S2TT a été développé dans l'optique d'être pérenne, réutilisable, et de proposer en un seul logiciel un système complet implémentant des techniques spécifiques à la traduction de la parole. Il a prouvé son potentiel, mais un effort d'ingénierie reste nécessaire pour obtenir une base d'expérimentation fiable et efficace et, qui sait, un outil utilisable par les professionels de la traduction et le grand public.

À plus long terme, quand le système sera stabilisé, certaines pistes de recherche me paraissent particulièrement riches et prometteuses. Sur les mesures de confiance tout d'abord : si les techniques de classification et régression multivariées permettent d'obtenir des systèmes efficaces, ils restent des boîtes noires. L'absence de caractère *explicatif* des estimations est perturbant pour les utilisateurs du système, et notamment les post-éditeurs, qui ne savent pas s'ils peuvent ou non faire confiance au système, et perdent donc du temps à analyser eux-même la phrase. On peut imaginer introduire dans le système d'annotation de confiance des connaissances expertes grammaticales et sémantique permettant, outre de détecter l'erreur, d'en expliquer la raison, voire de proposer une correction. Par exemple, si un système d'analyse grammaticale détecte une erreur d'accord, le système pourrait proposer une nouvelle traduction en augmentant le poids du modèle de langage pour améliorer les chances d'obtenir une phrase grammaticalement correcte ; si un système d'analyse sémantique détecte une différence de sens entre la phrase source et l'hyopthèse de traduction, augmenter le poids, pour cette instance, du modèle de traduction pourrait permettre d'obtenir une meilleure hypothèse. Ce n'est qu'un exemple, et bien d'autres façons de faire sont envisageables. J'y vois là une façon de tirer le meilleur des deux approches : les avantages de la couverture étendue des systèmes de traduction statistiques, et de la précision

des systèmes à base de règles. Idéalement, on aimerait même proposer des systèmes entièrement automatiques, ne reposant pas sur une base de connaissances expertes, capables d'une analyse syntaxique et sémantique assez précise pour détecter des erreurs et proposer, dans la mesure du possible, des corrections. Il faudra donc, dans les années qui viennent, être attentif aux avancées dans ces branches du traitement automatique des langues naturelles.

Enfin, sur la modélisation du phénomène très particulier et très insuffisamment connu (et reconnu) qu'est l'expression orale : comme je le rappelle dans la partie III, sections 1 (page 85) et 2.3 (page 100), l'oral n'est pas de l'écrit lu. C'est un medium à part entière, presque une autre langue, régie par ses phénomènes propres et, contraitement à ce qu'on pourrait croire, souvent plus sophistiqués qu'à l'écrit. Pauses, hésitations, disfluences, répétitions, « faux départs », prosodie, jeux de mots, sous-entendus, et même regards, gestuelle, attitude... sont des phénomènes essentiels de la communication orale et de la bonne comprehension entre deux interlocuteurs. Chacun a fait l'expérience au moins une fois d'une remarque mal interprétée par son interlocuteur au téléphone, parce que le sourire qui l'accompagnait n'a pas été transmis par le medium. Pour espérer réaliser une traduction automatique de bonne qualité, la nécessité de modéliser la langue orale dans toute sa complexité et sa subtilité me paraît donc criante. Les débuts seront modestes bien sûr, et avant de songer s'attaquer aux regards, il faudra commencer par prendre en compte les phénomènes d'élocution pure (pauses, hésitations, répétitions, etc.). Les HELM me semblent être la première piste à explorer pour cela, mais il faut aussi envisager de revenir totalement sur les modèles de langage les plus couramment utilisés, les modèles n-grammes, car l'hypothèse d'un phénomène markovien me paraît trop simple pour rendre justice à l'expression orale. C'est là un programme ambitieux mais, j'en suis convaincu, qui permettra de développer des systèmes de traduction de la parole d'une nature nouvelle et d'une qualité inégalée.

Bibliographie

[Abdi 03] H. Abdi, W. Dowling, D. Valentin, and B. Edelman. "Partial Least Square Regression". *Encyclopedia of Social Sciences Research Methods*, 2003.

[Alp 08] N. Alp and C. Turhan. "English to Turkish Example-Based Machine Translation with Synchronous SSTC". In : *Information Technology : New Generations, 2008. ITNG 2008. Fifth International Conference on*, pp. 674–679, Ieee, 2008.

[Alsh 98] H. Alshawi, S. Bangalore, and S. Douglas. "Learning phrase-based head transduction models for translation of spoken utterances". In : *Fifth International Conference on Spoken Language Processing*, 1998.

[Amen 00] J. Amengual, A. Castaño, A. Castellanos, V. Jiménez, D. Llorens, A. Marzal, F. Prat, J. Vilar, J. Benedi, F. Casacuberta, *et al.* "The EuTrans Spoken Language Translation System". *Machine Translation*, Vol. 15, No. 1, pp. 75–103, 2000.

[Baby 04] B. Babych and A. Hartley. "Extending the BLEU MT evaluation method with frequency weightings". In : *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 621, Association for Computational Linguistics, 2004.

[Bach 11] N. Bach, F. Huang, and Y. Al-Onaizan. "Goodness : A method for measuring machine translation confidence". In : *49th Annual Meeting of the Association for Computational Linguistics*, pp. 211–219, 2011.

[Bane 05] S. Banerjee and A. Lavie. "METEOR : An automatic metric for MT evaluation with improved correlation with human judgments". In : *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

[Bang 01] S. Bangalore and G. Riccardi. "A finite-state approach to machine translation". In : *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8, Association for Computational Linguistics, 2001.

[Bang 02] S. Bangalore and G. Riccardi. "Stochastic finite-state models for spoken language machine translation". *Machine Translation*, Vol. 17, No. 3, pp. 165–184, 2002.

[Benz 04] C. Benzitoun, E. Campione, J. Deulofeu, S. Henry, F. Sabio, S. Teston, A. Valli, J. Véronis, *et al.* "L'analyse syntaxique de l'oral : problèmes et méthodes". 2004.

[Bert 05] N. Bertoldi and M. Federico. "A new decoder for spoken language translation based on confusion networks". In : *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pp. 86–91, IEEE, 2005.

[Bert 07] N. Bertoldi, R. Zens, and M. Federico. "Speech translation by confusion network decoding". In : *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, pp. IV–1297, IEEE, 2007.

[Blac 02] A. Black, R. Brown, R. Frederking, K. Lenzo, J. Moody, A. Rudnicky, R. Singh, and E. Steinbrecher. "Rapid Development of Speech-to-Speech Translation Systems". In : *Seventh International Conference on Spoken Language Processing*, 2002.

[Blan 87] C. Blanche-Benveniste and C. Jeanjean. *Le français parlé. Edition et transcription.* 1987.

[Blat 03] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. "Confidence estimation for machine translation.". In : *Final report, JHU/CLSP Summer Workshop*, 2003.

[Brow 93] P. F. Brown and al. "The mathematics of statistical machine translation : parameter estimation". *Computational Linguistics*, Vol. 19, pp. 263–311, 1993.

[Brow 96] R. Brown. "Example-based machine translation in the pangloss system". In : *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pp. 169–174, Association for Computational Linguistics, 1996.

[Brun 05] A. Brun, C. Cerisara, D. Fohr, D. Fohr, I. Illina, D. Langlois, and O. Mella. "ANTS le système de transcription automatique du LORIA". In : *Worshop Ester*, Avignon, France, 2005.

[Bub 97] T. Bub, W. Wahlster, and A. Waibel. "Verbmobil : The combination of deep and shallow processing for spontaneous speech translation". In : *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, pp. 71–74, IEEE, 1997.

[Call 11] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. "Findings of the 2011 Workshop on Statistical Machine Translation". In : *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22–64, Association for Computational Linguistics, Edinburgh, Scotland, July 2011.

[Call 12] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. "Findings of the 2012 Workshop on Statistical Machine Translation". In : *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 10–51, Association for Computational Linguistics, Montréal, Canada, June 2012.

[Calm 98] M. de Calmès and G. Pérennou. "BDLEX : a Lexicon for Spoken and Written French". In : *Proceedings of 1st International Conference on Langage Resources & Evaluation*, pp. 1129–1136, Grenade, Espagne, May 1998.

[Casa 01] F. Casacuberta. "Finite-state transducers for speech-input translation". In : *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 375–380, IEEE, 2001.

[Casa 04a] F. Casacuberta, H. Ney, F. Och, E. Vidal, J. Vilar, S. Barrachina, I. Garcıa-Varea, D. Llorens, C. Martınez, S. Molau, *et al.* "Some approaches to statistical and finite-state speech-to-speech translation". *Computer Speech & Language*, Vol. 18, No. 1, pp. 25–47, 2004.

[Casa 04b] F. Casacuberta and E. Vidal. "Machine translation with inferred stochastic finite-state transducers". *Computational Linguistics*, Vol. 30, No. 2, pp. 205–225, 2004.

[Casa 09] F. Casacuberta and E. Vidal. "Speech-to-Speech Translation". 2009.

[Chan 11] C.-C. Chang and C.-J. Lin. "LIBSVM : A library for support vector machines". *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27 :1–27 :27, 2011. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Cort 95] C. Cortes and V. Vapnik. “Support-vector networks”. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[Dech 07] D. Déchelotte, H. Schwenk, G. Adda, and J. Gauvain. “Improved machine translation of speech-to-text outputs”. *Antwerp. Belgium*, pp. 2441–2444, 2007.

[Demp 77] A. Dempster, N. Laird, and D. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. *Journal of the Royal Statistical Society B*, Vol. 39, pp. 1–38, 1977.

[Dodd 02] G. Doddington. “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”. In : *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145, Morgan Kaufmann Publishers Inc., 2002.

[Duch 02] J. Duchateau, K. Demuynck, and P. Wambacq. “Confidence scoring based on backward language models”. In : *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 221–224, Orlando, Florida, 2002.

[Eise 10] A. Eisele and Y. Chen. “MultiUN : A Multilingual Corpus from United Nation Documents”. In : N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pp. 2868–2872, European Language Resources Association (ELRA), Valletta, Malta, may 2010.

[Eise 11] A. Eisele and C. Lavecchia. “Using Statistical Machine Translation for Computer-Aided Translation at the European Commission”. In : *Proceedings of the Third Joint EM+/CNGL Workshop ”Bringing MT to the User : Research Meets Translators” (JEC’11)*, October 2011.

[Faus 94] L. V. Fausett. *Fundamentals of neural networks*. Prentice-Hall Englewood Cliffs, NJ, 1994.

[Fuge 08] C. Fügen, A. Waibel, and M. Kolss. “Simultaneous translation of lectures and speeches”. *Machine translation*, Vol. 21, No. 4, pp. 209–252, 2008.

[Gale 07] M. Gales, X. Liu, R. Sinha, P. Woodland, K. Yu, S. Matsoukas, T. Ng, K. Nguyen, L. Nguyen, J. Gauvain, *et al.* “Speech recognition system combination for machine translation”. In : *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, pp. IV–1277, IEEE, 2007.

[Gall 06] S. Galliano, E. Geoffrois, G. Gravier, J. Bonastre, D. Mostefa, and K. Choukri. “Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news”. In : *Proceedings of LREC*, pp. 315–320, Genoa, Italy, 2006.

[Gand 03] S. Gandrabur and G. Foster. “Confidence estimation for translation prediction”. In : *Proceedings of the seventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 95–102, Edmonton, Canada, 2003.

[Gisp 02] A. Gispert and J. Mariño. “Using X-grams for speech-to-speech translation”. In : *Seventh International Conference on Spoken Language Processing*, 2002.

[Goll 05] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney. “Cross domain automatic transcription on the tc-star epps corpus”. In : *International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USE*, pp. 825–828, 2005.

[Good 53] I. Good. “The population frequencies of species and the estimation of population parameters”. *Biometrika*, Vol. 40, No. 3-4, pp. 237–264, 1953.

[Grav 04] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. M. Tait, and K. Choukri. "ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques". In : *Proc. JEP*, Fez, 2004.

[Guo 04] G. Guo, C. Huang, H. Jiang, and R. Wang. "A comparative study on various confidence measures in large vocabulary speech recognition". In : *Proceedings of the International Symposium on Chinese Spoken Language Processing*, pp. pp. 9–12, Hong Kong, China, 2004.

[Guyo 03] I. Guyon and A. Elisseeff. "An Introduction to Variable and Feature Selection". *Journal of Machine Learning Research (Special Issue on Variable and Feature Selection)*, pp. 1157–1182, 2003.

[Hart 68] P. Hart, N. Nilsson, and B. Raphael. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". *Systems Science and Cybernetics, IEEE Transactions on*, Vol. 4, No. 2, pp. 100–107, july 1968.

[Hato 03] J. Haton, C. Cerisara, D. Fohr, Y. Laprie, and K. Smaïli. *Reconnaissance de la parole : Du signal à son interprétation.* Dunod, 2003.

[He 10] Y. He, Y. Ma, J. van Genabith, and A. Way. "Bridging SMT and TM with translation recommendation". In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 622–630, Association for Computational Linguistics, 2010.

[Hsu 03] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. "A Practical Guide to Support Vector Classification". Tech. Rep., Department of Computer Science, National Taiwan University, 2003.

[Illi 04] I. Illina, D. Fohr, O. Mella, and C. Cerisara. "The Automatic News Transcription System : ANTS some Real Time experiments". In : *8th International Conference on Spoken Language Processing - ICSLP' 2004*, p. 4, Jeju, Corée du Sud, 2004.

[Jaya 05] S. Jayaraman and A. Lavie. "Multi-engine machine translation guided by explicit word matching". In : *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pp. 101–104, Association for Computational Linguistics, 2005.

[Jeli 80] F. Jelinek. "Interpolated estimation of Markov source parameters from sparse data". *Pattern recognition in practice*, pp. 381–397, 1980.

[Jian 11] J. Jiang, Z. Ahmed, J. Carson-Berndsen, P. Cahill, and A. Way. "Phonetic Representation-Based Speech Translation". In : *Proceedings of the 13th MT Summit Conference*, pp. 81–88, Xiamen, China, 2011.

[Jime 95] V. Jiménez, A. Castellanos, and E. Vidal. "Some results with a trainable speech translation and understanding system". In : *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, pp. 113–116, IEEE, 1995.

[Jura 00] D. Jurafsky, J. Martin, A. Kehler, K. Vander Linden, and N. Ward. *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*, Chap. Machine translation. Vol. 2, Prentice Hall New Jersey, 2000.

[Kemp 97] T. Kemp and T. Schaaf. "Estimating confidence using word lattices". In : *Proc. EUROSPEECH*, pp. 827–830, Rhodes, 1997.

[Kim 03] W. Kim and S. Khudanpur. "Cross-lingual lexical triggers in statistical language modeling". In : *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 17–24, Association for Computational Linguistics, 2003.

[Knes 95] R. Kneser and H. Ney. “Improved backing-off for m-gram language modeling”. In : *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, pp. 181–184, IEEE, 1995.

[Koeh 03] P. Koehn, F. Och, and D. Marcu. “Statistical phrase-based translation”. In : *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 48–54, Association for Computational Linguistics, 2003.

[Koeh 04] P. Koehn. “Pharaoh : a beam search decoder for phrase-based statistical machine translation models”. *Machine translation : From real users to research*, pp. 115–124, 2004.

[Koeh 05] P. Koehn. “Europarl : A Parallel Corpus for Statistical Machine Translation”. In : *Conference Proceedings : the tenth Machine Translation Summit*, pp. 79–86, AAMT, Phuket, Thailand, 2005.

[Koeh 06] P. Koehn and C. Monz. “Manual and Automatic Evaluation of Machine Translation between European Languages”. In : *Proceedings on the Workshop on Statistical Machine Translation*, pp. 102–121, Association for Computational Linguistics, New York City, June 2006.

[Koeh 07a] P. Koehn and al. “Moses : open source toolkit for statistical machine translation”. In : *Proceedings of the 45th Annual Meeting of the ACL*, pp. 177–180, Association for Computational Linguistics, Stroudsburg, PA, USA, 2007.

[Koeh 07b] P. Koehn and H. Hoang. “Factored translation models”. In : *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, p. 876, 2007.

[Kono 91] I. Kononenko and I. Bratko. “Information-based evaluation criterion for classifier’s performance”. *Machine Learning*, Vol. 6, No. 1, pp. 67–80, 1991.

[Kuhn 90] R. Kuhn and R. De Mori. “A cache-based natural language model for speech recognition”. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 12, No. 6, pp. 570–583, 1990.

[Lang 12] D. Langlois, S. Raybaud, K. Smaïli, *et al.* “LORIA System for the WMT12 Quality Estimation Shared Task”. In : *the Seventh Workshop on Statistical Machine Translation*, pp. 114–119, 2012.

[Lave 07] C. Lavecchia, K. Smaïli, D. Langlois, and J.-P. Haton. “Using inter-lingual triggers for Machine translation”. In : *Proceedings of the Eighth International Conference on Speech Communication and Technology (INTERSPEECH)*, pp. 2829–2832, Antwerp, Belgium, 2007.

[Lave 08] C. Lavecchia, D. Langlois, K. Smaïli, *et al.* “Phrase-based machine translation based on simulated annealing”. In : *Sixth international conference on Language Resources and Evaluation*, 2008.

[Lave 10] C. Lavecchia. *Les Triggers Inter-langues pour la Traduction Automatique Statistique*. PhD thesis, Université Nancy II, June 2010.

[Lee 01] A. Lee, T. Kawahara, and K. Shikano. “Julius — an open source real-time large vocabulary recognition engine”. In : *Proceedings of the EUROSPEECH conference*, pp. 1691–1694, 2001.

[Lepa 05] Y. Lepage and E. Denoual. "Purest ever example-based machine translation : Detailed presentation and assessment". *Machine Translation*, Vol. 19, No. 3, pp. 251–282, 2005.

[Leve 66] V. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In : *Soviet physics doklady*, pp. 707–710, 1966.

[Liu 03] F. Liu, L. Gu, Y. Gao, and M. Picheny. "Use of statistical N-gram models in natural language generation for machine translation". In : *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, pp. I 636 – I 639, IEEE, 2003.

[Llor 99] D. Llorens, F. Casacuberta, E. Segarra, J. Sánchez, P. Aibar, and M. Castro. "Acoustic and syntactical modeling in the ATROS system". In : *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, pp. 641–644, IEEE, 1999.

[Math 06] L. Mathias and W. Byrne. "Statistical phrase-based speech translation". In : *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 2006.

[Mats 07] S. Matsoukas, I. Bulyko, B. Xiang, K. Nguyen, R. Schwartz, and J. Makhoul. "Integrating speech recognition and machine translation". In : *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1281–1284, 2007.

[Matu 05] E. Matusov, S. Kanthak, and H. Ney. "On the integration of speech recognition and statistical machine translation". In : *Ninth European Conference on Speech Communication and Technology*, Citeseer, 2005.

[Maus 06] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. "The RWTH statistical machine translation system for the IWSLT 2006 evaluation". In : *Proc. of the International Workshop on Spoken Language Translation*, pp. 103–110, 2006.

[Maus 09] A. Mauser, S. Hasan, and H. Ney. "Extending statistical machine translation with discriminative and trigger-based lexicon models". In : *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1*, pp. 210–218, Association for Computational Linguistics, 2009.

[Mena 02] S. W. Menard. *Applied logistic regression analysis. Sage university papers, Quantitative applications in the social sciences*, Sage, Thousand Oaks, Calif. [u.a.], 2002.

[Mene 01] A. Menezes and S. Richardson. "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora". In : *Proceedings of the workshop on Data-driven methods in machine translation-Volume 14*, pp. 1–8, Association for Computational Linguistics, 2001.

[Metz 02] F. Metze, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, K. Laskowski, L. Levin, T. Schultz, *et al.* "The NESPOLE! speech-to-speech translation system". In : *Proceedings of the second international conference on Human Language Technology Research*, pp. 378–383, Morgan Kaufmann Publishers Inc., 2002.

[Mill 95] G. Miller. "WordNet : a lexical database for English". *Communications of the ACM*, Vol. 38, No. 11, pp. pp. 39–41, 1995.

[Naga 84] A. Nagao Makoto. *Artificial and human intelligence*, Chap. Framework of a Mechanical Translation between Japanese and English by Analogy Principle in Artificial˜

Human Intelligence, Alick Elithorn and Ranan Banerji eds. Elsevier Science Publishers, NATO, 1984.

[Naka 06] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. "The ATR multilingual speech-to-speech translation system". *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14, No. 2, pp. 365–376, March 2006.

[Ney 00] H. Ney, S. Nießen, F. J. Och, C. Tillmann, H. Sawaf, and S. Vogel. "Algorithms for Statistical Translation of Spoken Language". *IEEE Trans. on Speech and Audio Processing, Special Issue on Language Modeling and Dialogue Systems*, Vol. 8, pp. 24–36, January 2000.

[Ney 04] H. Ney. "The statistical approach to spoken language translation". In : *proceedings of the International Workshop on Spoken Language Translation*, 2004.

[Ney 94] H. Ney, U. Essen, and R. Kneser. "On structuring probabilistic dependences in stochastic language modelling". *Computer Speech and Language*, Vol. 8, No. 1, pp. 1–38, 1994.

[Ney 99] H. Ney. "Speech translation : coupling of recognition and translation". In : *Proc. ICASSP*, pp. 1149–1152, Phoenix, 1999.

[Niss 03] S. Nissen. "Implementation of a Fast Artificial Neural Network Library (fann)". Tech. Rep., Department of Computer Science University of Copenhagen (DIKU), 2003. http ://fann.sf.net.

[Och 00] F. J. Och and H. Ney. "Improved statistical alignment models". In : *ACL '00 : Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 440–447, Association for Computational Linguistics, Morristown, NJ, USA, 2000.

[Och 02] F. Och and H. Ney. "Discriminative training and maximum entropy models for statistical machine translation". In : *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 295–302, Association for Computational Linguistics, 2002.

[Och 03] F. Och and H. Ney. "Giza++ : Training of statistical translation models". 2003.

[Och 99] F. Och, C. Tillmann, H. Ney, *et al.* "Improved alignment models for statistical machine translation". In : *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, 1999.

[Oste 09] M. Ostendorf. "Transcribing Human-Directed Speech for Spoken Language Processing". In : *Proceedings of Interspeech 2009*, pp. 21–27, 2009.

[Papi 01] K. Papineni, S. Roukos, T. Ward, and W. Zhu. "Bleu : a method for automatic evaluation of machine translation". In : *Proceedings of the 40th Annual of the Association for Computational linguistics*, pp. 311–318, Philadelphia, USA, 2001.

[Plit 10] M. Plitt and F. Masselot. "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context". *The Prague Bulletin of Mathematical Linguistics*, Vol. 93, No. -1, pp. 7–16, 2010.

[Quir 04] C. Quirk. "Training a sentence-level machine translation confidence measure". In : *Proceedings of LREC 2004*, pp. 825–828, Citeseer, Lisbon, Portugal, 2004.

[Rayb 09a] S. Raybaud, C. Lavecchia, D. Langlois, and K. Smaïli. "New Confidence Measures for Statistical Machine Translation". In : *Proceedings of the International Conference on Agents and Artificial Intelligence*, pp. 61–68, Porto, Portugal, 2009.

[Rayb 09b] S. Raybaud, C. Lavecchia, D. Langlois, and K. Smaïli. "Word- and sentence-level confidence measures for machine translation". In : *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pp. 104–111, Barcelona, Spain, 2009.

[Rayb 11] S. Raybaud, D. Langlois, and K. Smaïli. ""This sentence is wrong." Detecting errors in machine-translated sentences.". *Machine Translation*, Vol. 25, No. 1, pp. 1–34, 2011.

[Razi 07] J. Razik. *Mesures de confiance trame-synchrones et locales en reconnaissance automatique de la parole.* PhD thesis, Université Nancy I, Oct. 2007.

[Razi 08] J. Razik, O. Mella, D. Fohr, J. Haton, *et al.* "Transcription automatique pour malentendants : amélioration à l'aide de mesures de confiance locales". In : *Journées d'Étude de la Parole*, 2008.

[Sale 04] S. Saleem, S. Jou, S. Vogel, and T. Schultz. "Using word lattice information for a tighter coupling in speech translation systems". In : *Proc. Int. Conf. on Spoken Language Processing*, pp. 41–44, 2004.

[Schm 94] H. Schmid. "Probabilistic part-of-speech tagging using decision trees". In : *Proceedings of International Conference on New Methods in Language Processing*, pp. 44–49, 1994.

[Schm 95] H. Schmid. "Improvements In Part-of-Speech Tagging With an Application To German". In : *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50, 1995.

[Seli 00] M. Seligman. "Nine Issues in Speech Translation.". *Machine Translation (Springer)*, Vol. 15, No. 1-2, pp. 149–186, 2000.

[Sene 01] J. Senellart, P. Dienes, and T. Varadi. "New generation systran translation system". In : *In Proceedings of MT Summit IIX Senellart J., Yang J., Rebollo A. 2003. SYSTRAN Intuitive Coding Technology. In Proceedings of MT Summit IX*, Citeseer, 2001.

[Shen 04] L. Shen, A. Sarkar, and F. Och. "Discriminative reranking for machine translation". In : *Proceedings of the Joint HLT and NAACL Conference (HLT 04)*, pp. 177–184, 2004.

[Shen 06] W. Shen, R. Zens, N. Bertoldi, M. Federico, C. per la Ricerca, and S. e Tecnologica. "The JHU Workshop 2006 IWSLT System". *eps*, Vol. 101, p. 14, 2006.

[Sima 07] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn. "Rule-based translation with statistical phrase-based post-editing". In : *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, pp. 203–206, Prague, 2007.

[Siu 99] M. Siu and H. Gish. "Evaluation of word confidence for speech recognition systems". *Computer Speech and Language*, 1999.

[Smai 91] K. Smaïli. *Conception et réalisation d'une machine à dicter à entrée vocale destinée aux grands vocabulaires : le système MAUD.* PhD thesis, Université Nancy1, September 1991.

[Smol 04] A. Smola and B. Schölkopf. "A tutorial on support vector regression". *Statistics and Computing*, Vol. 14, No. 3, pp. 199–222, 2004.

[Snov 06] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. "A study of translation edit rate with targeted human annotation". In : *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, Citeseer, Cambridge, 2006.

[Snov 09] M. Snover, N. Madnani, B. Dorr, and R. Schwartz. “Fluency, adequacy, or HTER ? : exploring different human judgments with a tunable MT metric”. In : *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 259–268, Association for Computational Linguistics, 2009.

[Spec 09] L. Specia, N. Cancedda, M. Dymetman, M. Turchi, and C. N. “Estimating the sentence-level quality of Machine Translation Systems”. In : *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pp. 28–35, Barcelona, Spain, 2009.

[Spec 10] L. Specia and A. Farzindar. “Estimating machine translation post-editing effort with HTER”. In : *Proceedings of JEC 2010 : Second joint EM+/CNGL Workshop “Bringing MT to the user : research on integrating MT in the translation industry”, AMTA 2010*, pp. 33–41, 2010.

[Spec 11] L. Specia. “Exploiting Objective Annotations for Measuring Translation Post-editing Effort”. In : *Proceedings of the 15th Conference of the European Association for Machine Translation*, pp. 73–80, 2011.

[Stol 02] A. Stolcke. “SRILM – An extensible language modeling toolkit”. In : *ICSLP*, pp. 901–904, Denver, USA, 2002.

[Take 98] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto. “A Japanese-to-English speech translation system : ATR-MATRIX”. In : *Fifth International Conference on Spoken Language Processing*, 1998.

[Tant 07] C. Tantug, K. Oflazer, and I. El-Kahlout. “BLEU+ : a tool for fine-grained BLEU computation”. 2007.

[Till 97a] C. Tillmann and H. Ney. “Word Trigger and the EM Algorithm”. In : *Proceedings of the Conference on Computational Natural Language Learning*, pp. 117–124, Madrid, Spain, 1997.

[Till 97b] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. “Accelerated DP based search for statistical translation”. In : *Fifth European Conference on Speech Communication and Technology*, 1997.

[Tobi 95] R. Tobias. “An introduction to partial least squares regression”. In : *Proceedings of the Twentieth Annual SAS Users Group International Conference, Cary, NC : SAS Institute Inc*, pp. 1250–1257, Citeseer, Orlando, Florida, 1995.

[Ueff 03] N. Ueffing, K. Macherey, and H. Ney. “Confidence measures for statistical machine translation”. In : *In Proc. MT Summit IX*, Citeseer, 2003.

[Ueff 04] N. Ueffing and H. Ney. “Bayes decision rule and confidence measures for statistical machine translation”. In : *Proceedings of EsTAL Espana for Natural Language Processing*, pp. 70–81, Springer, 2004.

[Ueff 05] N. Ueffing and H. Ney. “Word-level confidence estimation for machine translation using phrase-based translation models”. In : *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT ’05)*, pp. pp. 763–770, Association for Computational Linguistics, Morristown, NJ, USA, 2005.

[Uhri 97] C. Uhrik and W. Ward. “Confidence Metrics Based on N-Gram Language Model Backoff Behaviors”. In : *Fifth European Conference on Speech Communication and Technology*, pp. 2771–2774, Rhodes, Greece, 1997.

[Veal 97] T. Veale and A. Way. "Gaijin : A bootstrapping, template-driven approach to example-based MT". In : *Proc. of the NeMNLP97*, 1997.

[Voge 96] S. Vogel, H. Ney, and C. Tillmann. "HMM-based word alignment in statistical translation". In : *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pp. 836–841, Association for Computational Linguistics, 1996.

[Wahl 00] W. Wahlster. *Verbmobil : foundations of speech-to-speech translation*. Springer verlag, 2000.

[Weav 55] W. Weaver. *Translation*, pp. 15–23. MIT Press, 1955.

[Wold 84] S. Wold, A. Ruhe, H. Wold, and W. Dunn III. "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses". *SIAM Journal on Scientific and Statistical Computing*, Vol. 5, No. 3, pp. pp. 735–743, 1984.

[Wosz 93] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Sloboda, M. Tomita, *et al.* "Recent advances in JANUS : a speech translation system". In : *Proceedings of the workshop on Human Language Technology*, pp. 211–216, Association for Computational Linguistics, 1993.

[Xion 10] D. Xiong, M. Zhang, and H. Li. "Error detection for statistical machine translation using linguistic features". In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 604–611, Association for Computational Linguistics, 2010.

[Zens 05] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. "The RWTH phrase-based statistical machine translation system". In : *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 155–162, 2005.

[Zhan 01] R. Zhang and A. Rudnicky. "Word level confidence annotation using combinations of features". In : *Seventh European Conference on Speech Communication and Technology*, pp. 2105–2108, Aalgorg, Denmark, 2001.

A

Récapitulatifs des performances des différentes mesures de confiance

Mesures de confiance sur les mots :

Paramètre	Référence	Taux d'égale erreur	NMI
trigramme	3.1.1, page 54	42.1	4.86×10^{-3}
trigramme à rebours	3.1.1, page 54	42.9	-3.93×10^{-3}
repli	3.1.1, page 54	37.0	6.11×10^{-2}
repli à rebours	3.1.1, page 54	38.1	1.09×10^{-2}
trigrammes et contexte	3.1.3, page 56	36.3	4.57×10^{-3}
IM intra-langue	3.1.4, page 57	45.8	9.46×10^{-4}
IM inter-langues	3.1.5, page 58	45.7	-2.21×10^{-1}
IBM-1	3.1.6, page 59	45.0	-1.84×10^{-3}
Régression Logistique	3.1.8, page 60	36.8	-2.61×10^{-2}
PLSR	3.1.8, page 60	37.5	-5.84×10^{-2}
SVM	3.1.8, page 60	36.7	-1.87×10^{-1}
Réseau de Neurones	3.1.8, page 60	35.0	6.06×10^{-2}

Mesures de confiance sur les phrases :

Paramètre	Référence	Taux d'égale erreur	NMI
ppl trigramme	3.2.1, page 62	41.7	4.02×10^{-3}
ppl trigramme à rebours	3.2.1, page 62	41.3	3.97×10^{-3}
repli moyen	3.2.1, page 62	34.2	4.15×10^{-3}
IMI	3.2.2, page 63	39.0	9.46×10^{-4}
CMI	3.2.2, page 63	48.1	-0.30×10^{-4}
PLS	3.2.6, page 65	29.0	8.14×10^{-2}
SVM	3.2.6, page 65	38.0	-2.56×10^{-1}
Réseau de Neurones	3.2.6, page 65	41.3	-2.44×10^{-2}

B

Questionnaire distribué aux volontaire après l'expérience de post-édition

Pré-Questionnaire

Quel est votre niveau d'anglais (niveau estimé, diplômes éventuels, profession si pertinent, immersion dans un pays anglophone...) ?

De manière générale vous sentez-vous à l'aise avec les outils informatiques ?
❑ OUI
❑ NON

Travaillez-vous ou avez vous travaillé dans le domaine de la traduction ?
❑ OUI
❑ NON

Travaillez-vous ou avez vous travaillé dans le domaine des langues étrangères ?
❑ OUI
❑ NON

Utilisez-vous vos compétences linguistiques dans un cadre professionnel ?
❑ OUI
❑ NON

Utilisez-vous des outils d'aide à la traduction ?
❑ OUI
❑ NON

Si oui, lesquels ?

Utilisez-vous des traducteurs en ligne ?
❑ OUI
❑ NON

Si oui, lesquels ?

(a) Page 1

Questionnaire

De manière générale, qu'avez-vous pensé de l'interface (convivialité, efficacité...) ?

Les phrases à traduire étaient-elles difficiles ?

Les traductions proposées initialement par le système étaient-elles plutôt :
❑ totalement correctes (rien à corriger sauf exception)
❑ plutôt correctes (une, deux erreurs légères par phrase)
❑ plutôt incorrectes (un travail sérieux de correction pour chaque phrase)
❑ totalement incorrectes (toutes les phrases, sauf exception, devaient être lourdement réécrites)

Vous a-t-il semblé difficile de décider quelle corrections apporter (choix cornéliens...) ?
Avec la détection d'erreur : ❑ OUI ❑ NON Sans la détection d'erreur : ❑ OUI ❑ NON

Est-ce que le système vous a aidé à corriger les erreurs (lors de la phase où cette aide était active) ?
❑ pas du tout
❑ un peu
❑ beaucoup
❑ absolument

L'aide vous-a-t'elle été utile pour accélérer votre travail ?
❑ OUI
❑ NON

L'aide vous-a-t'elle été utile pour améliorer votre qualité de correction ?
❑ OUI
❑ NON

Est-ce que l'aide vous a aidé à détecter les erreurs ?
❑ OUI
❑ NON
Pourquoi ?

(b) Page 2

Avez-vous utilisé le curseur placé à gauche de l'interface ?
❑ OUI
❑ NON

Si oui, avez-vous constaté une différence dans l'aide apportée ?
❑ OUI
❑ NON

Commentez : quelle différence ? Positive ? Négative ?

Lors de l'utilisation de ce curseur, avez-vous trouvé le niveau de décision qui vous convenait ?
❑ OUI
❑ NON

Finalement, qu'avez-vous pensé de cette aide ?

Pour améliorer l'aide, que faudrait-il proposer ?

Si cette fonctionnalté « surlignement des mots supposés incorrects » était améliorée (si nécessaire), puis intégrée à un outil professionnel, ou à un traducteur en ligne déjà existant, pensez-vous que vous l'utiliseriez ?

(c) Page 3

Voyez-vous un intérêt pédagogique à cette fonctionnalité ?

Avez-vous d'autres remarques ?

Encore merci de votre participation à cette expérimentation

(d) Page 4

C

Transcriptions et traductions de référence du corpus de test de S2TT

Transcription de référence	Traduction automatique par S2TT	Traduction de référence 1	Traduction de référence 2
vous écoutez rfi il est cinq heures en temps universel	you listen rfi it is five hours in time universal	You're listening to RFI, it's five o'clock Universal time	you are listening to rfi it is five o'clock universal time
sept heures à paris nous sommes le samedi sept juillet deux mille sept sept heures paris le journal sylvie berruet bonjour	seven hours in paris we are the saturday seven july 2007 seven hours the paris newspaper to release the good morning	It's seven o'clock in Paris on Saturday July the 7th two-thousand and seven, seven o'clock Paris, news, Sylvie Berruet, hello	seven o'clock in paris this is saturday 7 july two thousand seven paris the news sylvie berruet good morning
bonjour	hello	Hello	hello
première inculpation dans les attentats manqués de londres et de glasgow un médecin iraquien qui figurait parmi les huit suspects a été mis en examen hier soir	first charge in the unsuccessful attempts london glasgow twenty doctors iraqis who was one of the eight suspects was charged yesterday is	A first person has been charged for the failed terrorist attacks in London and Glasgow, an Iranian doctor who was among the eight suspects indicted yesterday	first charge in the failed attacks of London and Glasgow an iraqi doctor who was among the eight suspects was indicted yesterday evening
en france deux perquisitions en deux jours les juges de l ' affaire clearstream ont perquisitionné hier le bureau de dominique de villepin l ' ancien premier ministre	in france two searches in two days judges of the case clearstream searched the bureau yesterday by dominique de villepin the former prime minister	In France two searches in two day. The judges from the Clearstream case searched the office of Dominique de Villepin the former prime minister yesterday	in France two searches in two days the judges of the clearstream affair carried out a search yesterday in the office of dominique de villepin former prime minister
les sports avec le départ du quatre vingt quatorzième tour de france le prologue s ' élance à londres cet après midi pour huit kilomètres de course dans les quartiers historiques de la ville	sports with the departure of 81 fourteenth tour de france the prologue has launched in london this afternoon for him kilometres of race in the cards historic city	Sport now with the start of the ninety-fourth Tour de France. The prologue starts in London this afternoon with an 8 kilometre race through the historic parts of the city	Sports : departure of the ninety-fourth tour de France, the prologue takes a run up in London this afternoon for an eight kilometer race in the historical districts of the city
un tour de france entaché par des affaires de dopage	a tour de france tainted by cases of doping	A Tour de France tainted by doping scandals	a tour de france sullied by doping scandals
et puis wimbledon sourit aux français la finale dames oppose ce samedi marion bartoli à l ' américaine venus williams la triple vainqueur du tournoi	and then we like the gives smiled the french the final women opposed this saturday marion bartoli the american venus williams the triple winner of the tournament	And Wimbledon is going well for French players. The Ladies' Final this Saturday will be between Marion Bartoli and the American Venus Williams, triple winner of the tournament	in addition wimbledon is favourable to the french, women's single final on Saturday brings together marion bartoli and the american venus williams triple winner of the tournament
marion bartoli a réalisé une véritable prouesse hier en sortant la numéro un mondiale	bartoli marion has done a feat there is leaving the number one world	Marion Bartoli pulled off a real feat yesterday by knocking out the world number one	marion bartoli performed a real feat yesterday by knocking out world number one
la belge justine hénin	the belgian justine hénin	the Belgian Justine Hénin	the belgian justine henin
je vous le disais c ' est sydney qui a ouvert le bal avec une cérémonie traditionnelle aborigène de bienvenue et un discours de peter garrett l ' ancien chanteur du groupe australien midnight oil	As I was saying, Sydney kicked the proceedings off with a traditional Aborigine welcome ceremony and a speech by Peter Garrett, the former singer of the Australian group Midnight Oil	it is sydney which opened the ball with a ceremony traditional aboriginal welcome there is a speech of peter garrett this former singer of the australian midnight all	I told you that Sydney started the dance with a traditional aboriginal welcome ceremony and a speech by peter garrett former lead singer of the australian band midnight oil
suivront après sydney tokyo chang hai hambourg londres johannesbourg new york et rio	will follow after sydney tokyo shanghai hamburg london johannesburg new york	Sydney, Tokyo, Shanghai, Hamburg, London, Johannesburg, New York and Rio will follow	after sydney tokyo changhai hamburg new york johannesburg london and rio will follow

je vous le disais plus d ' une centaine d ' artistes mobilisés et parmi eux la béninoise angélique kidjo qui se produit à johannesbourg	i said more of a hundred of artists mobilised among them the béninoise angelic which joe happens age honnis	As I was saying, over a hundred artists are involved including the Beninese singer Angélique Kidjo who will be playing in Johannesburg	I told you more than one hundred artists mobilized and among them the beninese angelique kidjo who is performing in johannesburg
nous avons tous cette terre que nous partageons ensemble et que il y a danger	we have got this earth that we share together and that there is danger	We have this Earth which we all share and there is a danger	we all share this earth and that there is a danger
ce qui m ' épate dans les questions et dans la réaction des gens c ' est le cynisme le fait de nous prendre nous les artistes comme boucs émissaires d ' un problème qui se pose	what amazes me in questions and reactions is cynicism the fact we take we artists as scapegoats of a problem that settles	What strikes me in the people's questions and reactions is the cynicism, the way we the artists are made scapegoats for an existing problem	what amazes me in people's questions and reactions is the cynicism the fact that we the artists are used as scapegoats for a problem which arises
et que on peut aider à à résoudre et que on n ' a pas le droit d ' aider à résoudre	and that we can help to solve but that we are not allwed to help to solve	and which we can help to solve but which we are not allowed to help to solve	and that we can help to solve but we are not allowed to help to solve
on a banalisé publics mais on ne peut pas être des artistes avec de la substance	we are commonplace audience but we cannot be artists with substance	we have trivialised audiences but we can't really be artistes of substance	we trivialized the audience but we can not be artists with substance
depuis les années soixante dix c ' est vrai qu ' il y a beaucoup de différents concerts parce qu ' il y a une urgence pour toucher les gens autrement que les hommes politiques ne touchent plus les gens forcément	since the years 70 it is true there are many different concerts because there is an emergency to touch the people otherwise the policies humans touch people	Since the seventies it's true that there have been a lot of different concerts, because it is urgent to get across to people differently from the way politicians do so, they no longer necessarily touch people	since the seventies it is true that there are many different concerts because there is an urgency to reach people in other ways than politicians who do not reach people any more necessarily
les années soixante pendant lesquelles on avait cette ce ce sentiment de pouvoir changer les choses de pouvoir créer un monde meilleur a disparu	the years 60 during we had this feeling of power change the things power create a better world has disappeared	The sixties when we had the impression that we could change things and create a better world have now disappeared	the sixties during which we had this feeling that we could change things that we could create a better world are gone
au profit d ' un monde de course à à l ' argent à l ' armement à tout sauf à la paix et à l ' harmonie entre les hommes et surtout à la préservation de cette planète	to the profit of a world of race to money weapons to everything except the peace and the harmony between the people and mostly to the preservation of this planet	in favour of a world with a race for money and arms, for everything except peace and harmony between people and above all the preservation of this planet	to the benefits of a world race to money to weapons to anything but peace and harmony between human beings and especially to the preservation of this planet
la béninoise angélique kidjo qui répondait aux questions de josé marino et puis parmi	the béninoise angelic a dj euro responding to questions of josé marie not	The Beninese Angélique Kidjo was interviewed by José Marino and then among	the beninese angelique kidjo who was answering josé marino's questions and then among
l ' environnement on ne parle plus que de cela un peu partout en france	the environment we speak more than this a little everywhere in france	the environment, everyone is talking about that pretty much everywhere in France	the environment one speaks only of that everywhere in France
un plan de bataille présenté par le ministre de l ' écologie du développement et de l ' aménagement durable jean louis borloo	a battle plan presented by the minister of ecology development and the planning sustainable jean-louis borloo	a battle plan presented by the Ecology, Sustainable Development and Territorial Development Minister, Jean-Louis Borloo	a battle plan presented by jean louis borloo the minister of ecology and sustainable development
pendant ce temps le président de la république nicolas sarkozy recevait en fin de journée une délégation d ' experts mondiaux du climat	during this time the president of the republic nicolas sarkozy was finally day a delegation of experts global climate	At the same time, the President Nicolas Sarkozy was receiving a delegation of world climate experts late in the day	meanwhile the president of the republic nicolas sarkozy was receiving late in the day a delegation of world climate experts

une délégation conduite par laurence tubiana la directrice de l ' institut du développement durable et des relations internationales	a delegation led by laurence tubiana the director of the institute of sustainable development international relations	A delegation led by Laurence Tubiana the director of the Institute of Sustainable Development and International Relations	a delegation led by laurence tubiana director of the institute for sustainable development and international relations
je voulais qu ' on réunisse des experts vraiment très très connus pour après une période où en france on a parlé essentiellement des affaires intérieures	i wanted to that we meet experts very , very known for after a period of great france was discussed mainly home affairs	I wanted a meeting of very, very well-known experts for after a period in which only internal affairs were discussed in France	I wished well known experts to be gathered for after a period in which one mainly spoke in france about internal affairs
qu ' on commence à regarder dehors et dehors c ' est des pays émergents c ' est les grands sujets du développement durable le changement climatique	the emerging countries the major issues of sustainable development climate change	that we look outside our country and outside means emerging countries, these are the major subjects of sustainable development, climate change	we could begin to look outside and outside means emerging countries this is the major issue of sustainable development climate change
toutes les questions de de de transformation du modèle industriel qui demandent des formes de négociation complètement différentes	all the issues of processing of industrial model that calls for negotiation efforts completely different	all the questions of how to transform the industrial model requiring completely different forms of negotiations	all issues of the industrial model transformations that require completely different forms of negotiation
merci d ' écouter rfi	thank you to listen to rfi	Thank you for listening to RFI	thank you for listening to rfi
il est six heures trente à paris quatre heures trente en temps universel	it is six hours thirty to paris four hours thirty in time universal	It's half past six in Paris, half past four Universal time	it is half past six in paris half past four universal time
rfi afrique	rfi africa	RFI Africa	rfi africa
olivier four	the ovens noon	Olivier Four	olivier four
bonjour et bienvenue à tous dans cette nouvelle édition d ' afrique matin	Hello and welcome to today's edition of Morning in Africa	good morning and also welcome all in this new edition of africa morning	hello and welcome to another edition of afrique matin
à la une de ce journal la construction de l ' union méditerranéenne l ' immigration la coopération dans le secteur de l ' énergie au programme de la visite de nicolas sarkozy à alger le président français rencontre ce matin abdelaziz bouteflika	a that goes to the construction of the mediterranean union immigration cooperation in the energy sector in the programme of the visit of the french president nicolas sarkozy algiers ronces against this morning abdel aziz bouteflika	Today's headlines feature the construction of the Mediterranean Union, immigration, cooperation in the energy sector on the schedule of Nicolas Sarkozy's visit to Algiers. The French president is meeting Abdelaziz Bouteflika this morning	at the front page of this news the construction of the mediterranean union, immigration, cooperation in the energy sector on the agenda of nicolas sarkozy's visit in algiers the French president meets Abdelaziz Bouteflika this morning
avant d ' aller en tunisie c ' est le tout premier déplacement de nicolas sarkozy au maghreb depuis son arrivée à l ' élysée	before going in tunisia this is the first displacement of nicolas sarkozy the maghreb since its arrival at the elysée	before leaving for Tunisia it's Nicolas Sarkozy's first visit to the Maghreb since his arrival at the Elysée	before going to tunisia it is nicolas sarkozy's first trip to the Maghreb since his arrival at the elysee
en rdc le secrétaire exécutif de la jeunesse du rcd a certainement été assassiné floribert bwana chui avait disparu samedi dernier son corps a été retrouvé près de goma	in democratic republic of the congo the executive secretary youth the rcd has certainly been murdered of horrific grass acquittal yes had disappeared last saturday her body was found near goma	In the Democratic Republic of the Congo, the country's executive secretary for youth has very probably been assassinated. Floribert Bwana Chui disappeared last Saturday and his body has been found near Goma	rdc/ DRC the executive secretary of rdc/drc youth has certainly been murdered floribert bwana chui had disappeared last saturday his body was found near goma

au zimbabwe le président mugabé veut faire plier les commerçants qui ne respectent pas la baisse des prix imposée par les autorités mille trois cents personnes ont été interpellées ces derniers jours	in zimbabwe president mugabe wants to persuade traders who do not respect the price reductions imposed by the authorities thousand three hundred people were detained in recent days	In Zimbabwe, President Mugabé wants to force shop owners to respect price reductions imposed by the authorities. One thousand three hundred people have been questioned over the last few days	in zimbabwe president mugabe wants to bend the tradesmen who do not respect the lower prices imposed by the authorities 1300 people were arrested in recent days
à suivre afrique matin reportage en mauritanie où l ' accès à l ' électricité est un véritable parcours du combattant dans les quartiers défavorisés de nouakchott manon rivière nous racontera comment ça marche ou plutôt comment ça ne marche pas	to follow africa morning report in mauritania where access to electricity is a real obstacle course in disadvantaged areas of nouakchott manon river he tells us how it works or rather , how it works	Morning in Africa to follow. A report from Mauritania where obtaining access to electricity is a true obstacle course in the poor areas of Nouakchott. Manon Riviere tells us how this works or rather how it doesn't work.	next afrique matin report in mauritania where access to electricity is a real obstacle course in the inner city of Nouakchott manon rivière will tell us how it works or rather how it does not work
et puis l ' invité d ' afrique matin le journaliste algérien zine cherfaoui il écrit dans el watan il nous donnera son point de vue sur la visite de nicolas sarkozy à alger	and then it was not invited to after ouattara the algerian journalist algiers do sherpas yes it writing in el watan we will give its view on the visit to algiers nicolas sarkozy	And Morning In Africa's guest, the Algerian journalist Zine Cherfaoui. He writes for El Watan. He will give us his views on Nicolas Sarkozy's visit to Algiers	and then the guest of afrique matin the algerian journalist zine cherfaoui who writes in el watan will give us his views on nicolas sarkozy's visit in algiers
c ' est donc ça toute première	is so this is his first visit	It is therefore his first	this is thus his first
visite hors d ' europe depuis son arrivée à l ' élysée aujourd 'hui le président français nicolas sarkozy se rend en afrique du nord il est accompagné du ministre des affaires étrangères bernard kouchner et de la secrétaire d ' état aux droits de l ' homme rama yade	outside of europe since its arrival at the elysée today french president nicolas sarkozy is in north africa it is accompanied by the foreign minister bernard kouchner and secretary of state for human rights rama yade	trip outside Europe since his arrival at the Elysée. Today the French President Nicolas Sarkozy is going to North Africa. He will be accompanied by the Foreign Affairs Minister Bernard Kouchner and the Secretary of State for Human Rights Rama Yade	visit outside europe since his arrival at the elysee today french president nicolas sarkozy travels to north africa he is accompanied by foreign minister bernard kouchner and secretary of state for human rights rama yade
nicolas sarkozy va rester quelques heures seulement en algérie hein le temps de rencontrer le président bouteflika il reprendra ensuite l ' avion pour se rendre en tunisie	nicolas sarkozy will remain only a few hours in algeria has time to meet president bouteflika resume afterwards it the plane to travel in tunisia	Nicolas Sarkozy will only stay a few hours in Algeria right, the time required to meet President Bouteflika and then he'll take a plane to Tunisia	nicolas sarkozy is going to stay for a few hours in algeria eh the time to meet president bouteflika then he will resume his trip to tunisia
dans ces deux pays le président français vient promouvoir son idée d ' union méditerranéenne mais les dossiers de l ' immigration de la sécurité et de la coopération seront également à l ' ordre du jour	in these two countries the french president has just promote its idea of mediterranean union but the dossiers immigration security and cooperation will also be on the agenda	In these two countries, the French President will be promoting his idea of a Mediterranean Union but immigration, law and order and cooperation are also on the meeting schedule	in these two countries the french president wants to promote his idea of mediterranean union but the immigration security and cooperation files will also be on the agenda
en ce qui concerne plus particulièrement l ' algérie la france veut nouer un nouveau partenariat avec son voisin de la rive sud de la méditerranée	in this applies particularly to france can we a new partnership with its neighbour of the southern shore of the mediterranean	More specifically concerning Algeria, France wishes to establish a new partnership with its neighbour on the south shore of the Mediterranean	concerning particulary algeria france wants to establish a new partnership with its neighbor on the south shore of the mediterranean
le président français souhaite aussi un rapprochement entre la sonatrach la société algérienne des hydrocarbures et des sociétés françaises comme gaz de france suez ou total envoyé spécial de rfi à alger christophe boisbouvier	the french president also wish a rapprochement between the sonatrach algerian society hydrocarbons and of the french companies as gaz-de-france suez of august the talent à yes special rfi in algiers christophe wood bouvier	The French President also hopes for closer links between Sonotrach, the Algerian hydrocarbons company, and French companies like Gaz de France, Suez or Total. RFI's correspondent in Algiers, Christophe Boisbouvier	french president also wants closer links between the algerian company sonatrach petroleum and french companies like gaz de france suez or total rfi special correspondant in algiers christophe boisbouvier

moins de parole et plus d ' actes ce pourrait être la devise de nicolas sarkozy pour essayer de dépassionner la relation entre paris et alger	less a relative to the island more act this could be the motto of nicolas sarkozy im to try to take the heat out of the relationship between paris and algiers	Less talk and more action could be Nicolas Sarkozy's byword for his attempts to create calmer relations between Paris and Algiers	less word and more actions that could be the motto of Nicolas Sarkozy to try to defuse relationships between paris and algiers
depuis trois ans les rapports sont tendus sans doute parce que jacques chirac a voulu placer la barre trop haut avec sa proposition de traité d ' amitié	for three years the reports are nervous without doubt because jacques=chirac wanted to set the bar too high with its treaty proposal friendship	For the last three years, relations have been strained probably because Jacques Chirac set the bar too high with his proposed friendship treaty	the relationships have been strained for three years probably because jacques chirac wanted to place the bar too high with his proposed friendship treaty
aujourd 'hui le projet de traité est enterré et la querelle sur la mémoire et la guerre d ' algérie est mise de côté comme dit joliment le journal el watan on se dirige vers un pacte de non agression	today the draft treaty is buried it has dispute on the memory and war algeria the left as so nicely puts the newspaper el watan we are moving towards a pact not aggression	The planned treaty has now been forgotten and the argument about remembering the Algerian war has been put aside as the newspaper El Watan so beautifully puts it. The most likely now is a non-aggression pact	today the draft treaty is buried and the dispute over memory and the algerian war is set aside as newspaper el watan nicely said one is heading towards a non-aggression pact
place désormais du point de vue de paris à un partenariat d ' exception entre la france et l ' algérie sur des dossiers concrets comme l ' échange du gaz algérien contre le nucléaire civil français	place now from the point of view of paris a partnership of exception between france and algeria on the practical issues as the exchange of algerian gas against the civil nuclear french	This leaves more room for Paris' views on a exceptional partnership between France and Algeria covering matters like exchanges of Algerian gas for French civil nuclear energy	make way now from Paris point of view for a partnership of exception between france and algeria on concrete issues like the exchange of algerian gas for the french civil nuclear power
ou sur le contrôle des flux migratoires il y aurait selon paris trois cent cinquante mille algériens sans papiers en france	or on controlling migratory flows , there would be in paris three hundred and fifty thousand algerians without papers in france	or controlling immigration levels. According to Paris, there are three hundred and fifty thousand illegal Algerian immigrants in France	or the control of migration flows according to paris there would be three hundred and fifty thousand algerians without papers in france
entre nicolas sarkozy et abdelaziz bouteflika le courant passe la dernière fois qu ' ils se sont vus en novembre dernier à alger les deux hommes se sont parlés pendant cinq heures d ' affilée ce n ' est pas banal	between nicolas sarkozy abdel aziz bouteflika the course is the last time that they have seen in november in algiers the two men have spoken for five hours at a time that is not trivial	Nicolas Sarkozy and Abdelaziz Bouteflika get on well. At their last meeting in November in Algiers, they spoke for five hours which is not usual at all	nicolas sarkozy and bouteflika get on well last time they met last november in algiers the two men spoke for five hours at a stretch which is unusual
mais à l ' époque nicolas sarkozy n ' était que ministre de l ' intérieur avec ses interlocuteurs il pouvait parler de tout et de rien sans engager la france	but at the time nicolas sarkozy was that minister of the interior with his interlocutors we could talk about everything and nothing without engage france	But at that time Nicolas Sarkozy was only Home Secretary. He could talk to people about everything then without committing France to anything	but at the time Nicolas Sarkozy was only the interior minister with his interlocutors he could talk about everything and anything without involving France
aujourd 'hui il sait qu ' il risque d ' être interrogé sur le sujet qui fâche les algériens à savoir le soutien de la france au maroc dans le conflit du sahara christophe boisbouvier alger rfi	today it knows that it may be asked about the inflammatory subjects the algerians namely the support of france to morocco in conflict sahara christopher wood alger rfi	Today he knows he is likely to be asked about the subject which annoys the Algerians namely France's support for Morocco in the Sahara conflict. Christophe Boisbouvier. Algiers, RFI	today he knows that he risks being interviewed on the sore subject which makes Algerians angry namely the support of france in morocco in the sahara conflict christophe boisbouvier algiers rfi
ah et puis justement sur la question du sahara le président français nicolas sarkozy affirme dans une interview publiée par el watan et el khabar qu ' il est grand temps de trouver une solution durable au conflit du sahara occidental	and then precisely on the question of the sahara french president nicolas sarkozy says in an interview published by the drink year in the arab harka of large ports to find a lasting solution to the conflict in western sahara	Oh and as we were talking about the Sahara, the French president Nicolas Sarkozy has said in an interview published by El Watan and El Khabar that it is high time to find a lasting solution to the conflict in the Western Sahara region	ah and then just on the sahara issue french president nicolas sarkozy said in an interview published by el watan and el khabar that it is high time to find a lasting solution to the conflict in western sahara

une solution raisonnable et acceptable par chacune des parties souligne le président français	a solution in the next reasonable acceptable by each of the parties underlines the french president	A reasonable solution which is acceptable for all parties as the French president stressed	a reasonable solution acceptable to both parties stressed the french president
le calme est maintenant revenu dans la principale prison de nouakchott en mauritanie hier des détenus se sont révoltés ils ont commencé à tout casser le mobilier et les portes la garde nationale a été appelée en renfort	calm is now income in the main prison nouakchott in mauritania yesterday of detainees have rebelled and they have begun to destroy everything the furniture the doors windows the national guard was called in reinforcement	Calm has now returned to the main prison of Nouakchott in Mauritania. Prisoners revolted yesterday and began smashing furniture and doors. The country's national guard were called	the calm has now returned to the main prison in nouakchott mauritania yesterday inmates rioted they started breaking the furniture and doors the national guard was called as reinforcements
merci à vous d ' écouter france inter il est dix neuf heures	thank you to listen to france-inter it is 7.15 p.m.	Thanks for listening to France Inter, it's 7pm	thank you to you for listening to france inter it is 7 o'clock p.m.
le dix huit vingt continu avec le journal de mickaël thébault bonsoir mickaël	the eighteen will continue with the newspaper michael thébault goodbye michael	The 6-8 programme continues with Mickael Thibaut's news feature. Good evening Mickael	the 6-8 p.m. news bulletin continues with mickaël thébault's report good evening mickaël
bonsoir	good evening	Good evening	good evening
et si c ' était un français socialiste de surcroît qui prenait la tête du fmi	the six were in france and the socialists furthermore that it was the head of the imf	And what if a French socialist of all people was named as head of the IMF ?	what if it was a french and moreover a socialist who would be at the head of the imf
dominique strauss kahn semble en tous cas très bien parti pour décrocher la direction du fonds monétaire international	dominique strauss-kahn have any case very well party to obtain the direction of the international monetary fund	Dominique Strauss Kahn does seem in a very good position to be appointed head of the international monetary fund	in all cases dominique strauss-kahn appears all set to win the leadership of the international monetary fund
il y avait déjà le soutien du chef de l ' état l ' union européenne lui dit aussi oui concert de louanges avec de très rares voix discordantes	which has already had the support of the head of the state of the european union has also said yes the chorus of praise with very few dissenting voice	He already had the head of state's support. The European Union say yes as well. A chorus of praise with very few dissenting voices	he already had the support of the head of state the european union also agrees it is a chorus of praise with very few dissenting voices
dominique de villepin lui devant les juges c ' est pour la fin du mois affaire clearstream et cette fois la mise en examen semble proche	dominique de villepin him before the judges for the end of the month of case clearstream and this time the indictment seems near	As for Dominique de Villepin he is in court at the end of the month for the Clearstream affair and this time charges seem likely	As for Dominique de Villepin before the judges that is for the end of the month, the clearstream affair and this time the indictment seems close
nicolas sarkozy lui vient d ' arriver à tunis après alger cet après midi visite éclair au maghreb où ce n ' est ni l ' heure ni le lieu de la repentance	nicolas sarkozy him has just reached tunis algiers after this afternoon lightning visit the maghreb where there was a city in city instead of repentance	As for Nicolas Sarkozy, he just arrived in Tunis from Algiers this afternoon. A lightning visit to the Maghreb countries where it is neither the time nor the place for repentence	As for nicolas sarkozy he has just arrived in tunis after algiers this afternoon after a flying visit to the maghreb where this is neither the time nor the place for repentance
il l ' a redit et vous l ' entendrez dans un instant	it has reiterated wanting to make in a moment	He has repeated it and you will hear him in a moment	he has repeated and you will hear in a moment
c ' est parti pour le tepa le fameux paquet fiscal soumis depuis cet après midi aux députés la ministre des finances lyrique le ps mordant	it is party for the outset the so-called tax package submitted since this afternoon to members of the g8 finance ministers ps deaths which	The "Tepa" is off and running, the famous tax package was submitted to members of parliament this afternoon. The Finance Minister waxed lyrical, the PS was biting	here we go for the tepa the tax package submitted this afternoon to members of parliament the finance minister was lyrical and the socialist party caustic

deux chiffres également à la une ce soir un million deux cents mille français consomment régulièrement du cannabis cinq cents mille chaque jour dernier rapport de l ' observatoire français des drogues et des toxicomanies son directeur dans un instant	two figures also a tonight a million two hundred thousand french consume regularly cannabis five hundred thousand every day the latest report by the french monitoring centre for drugs and drug addiction spend director in a moment	Two figures are also in the headlines this evening. One million two hundred thousand French people regularly consume cannabis. Five thousand a day. Latest report from the French Drugs and Addiction Observatory, we'll hear from its director in a moment	two figures also in focus this evening one million two hundred thousand french people consume cannabis regularly five hundred thousand every day according to the french observatory for drugs and drug addiction last report its director in a moment
le tour de france troisième étape remportée il y a quelques minutes par le suisse cancellara qui reste donc maillot jaune c ' était l ' étape la plus longue de cette grande boucle deux mille sept et peut être aussi celle qui restera comme la plus lente	the tour de france three ème step won a few minutes ago by the swiss when it is lara which remain therefore yellow jersey there is the stage the longer this great loop 2007 but perhaps also the one which will remain as the more slowly than	The third stage of the Tour de France was won a few minutes ago by the Swiss Cancellera who therefore retains the yellow jersey. It was the longest stage of the 2007 "Big Loop" and perhaps also will turn out to have been the slowest	the tour de france third stage won a few minutes ago by the swiss cancellara who therefore is still wearing yellow jersey it was the longest stage of this large loop two thousand seven and perhaps one that will remain as the slowest
journal du tour à suivre avec jean françois rhein	the journal of all reassured jean-françois some	The Tour news feature will follow with Jean François Rhein	diary of the tour to follow with jean francois rhein
quel temps pour demain légère légère amélioration nous promet météo france mais ce ne sera toujours pas l ' été gardez le parapluie à portée de main	time for tomorrow slight slight improvement promises us weather france two still not been kept of takes support tomorrow	What will the weather be like tomorrow ? The French Met Office promises us a slight improvement but it still won't be summery. Keep an umbrella close to hand	the weather tomorrow slight slight improvement promises météo france but it will still not be summer keep the umbrella handy
météo complète de joël collado dès la fin de ce journal et juste après bien sûr le téléphone sonne d ' alain bédouet où l ' on reparlera du paquet fiscal réforme nécessaire ou cadeau électoral	to a weather full joe nicolas doze from the end of this newspaper of course the telephone rings alain=bedouet where we will return of the tax package has necessary reforms or gift electoral	A full weather forecast by Joel Collado at the end of this news programme and just after, of course, "Le Téléphone Sonne" with Alain Bédouet where the subject of discussion will be the tax package - a necessary reform or an electoral gift ?	complete weather forecast by joël collado at the end of this news bulletin and of course right after alain bédouet's le téléphone sonne where we will talk about the tax package necessary reform or election gift
le socialiste michel sapin et l ' ump hervé mariton vont en débattre	the socialist michel tree and the ump hervé mariton hear you to discuss	The socialist Michel Sapin and the UMP party's Hervé Mariton will debate this	socialist michel sapin and ump herve mariton will discuss about that
vos questions vos commentaires dès maintenant au zéro un quarante cinq vingt quatre sept mille ou sur le site trois w point france inter point com	your question your comments now at a zero forty-five twenty-four seven thousand on the site www point france-inter .com	Your questions and comments from now on by phone on "oh" one forty-five twenty-four seven thousand or on the site www dot France Inter dot com	your questions your comments now at zero one four five two four seven zero zero zero or on the site three w dot france inter dot com
france inter	france-inter	France inter	france inter
c ' est un premier pas très important je vais maintenant m ' attacher à convaincre les autres parties prenantes voilà donc ce soir mot pour mot la toute première déclaration de celui qui a désormais de sérieuses chances de prendre la tête du fonds monétaire international le fmi	this is a very important first step i shall now to focus to convince the other parties advocating so tonight word for word the first statement two that which has now serious opportunities to take the lead in the international monetary fund the imf	It is a very important first step. I will now strive to convince the other parties involved. Word for word this was the very first public statement by the person who henceforth has a serious chance of leading the IMF or International Monetary Fund	this is a very important first step i am now going to attempt to convince the other parties here tonight these are word for word the very first statement of the person who now has a serious chance to lead the international monetary fund the imf

dominique strauss kahn a en effet obtenu ce midi le soutien officiel de toute l ' union européenne les vingt sept ministres des finances l ' ont choisi lui et nul autre	dominique strauss-kahn indeed achieved this lunchtime the official support throughout the european union the 27 finance ministers that it chooses him and no other	Today at midday Dominique Strauss Kahn indeed obtained the official support of the whole European Union. The twenty-seven finance ministers have chosen him alone	dominique strauss kahn indeed obtained this afternoon the official support of all the european union the 27 finance ministers chose him and no other
la pologne a bien essayé de placer son ancien premier ministre mais pas le moindre soutien	poland has tried to put its former prime minister but by any support	Poland had tried to place its former prime minster but there was absolutely no support for this	poland tried to place its former prime minister but not the least support
il faut dire françoise degois que l ' ancien ministre socialiste des finances fait l ' unanimité sur ses compétences comme économiste et comme négociateur et qu ' il a laissé un aussi bon souvenir à bercy qu ' à bruxelles	it must be said françoise=degois kohl former socialist minister of finance is one aine imitated on its powers as economist and as a negotiator and he has also left a good memories to bercy that brussels	It has to be said, François Dégois, that the socialist finance minister has unanimous support for his talents as an economist and negotiator and that he made just as good an impression at Bercy as in Brussels	it must be said francoise degois that there is a general agreement about the skills of the former socialist minister of finance as a negotiator and as an economist and that he also left a good memory to bercy and brussels
quelle couronne de lauriers il n ' est plus habitué depuis longtemps dominique strauss kahn à ce déluge de louanges les dernières remontent à ses années bercy	some sek of laurier it is more accustomed long dominique strauss-kahn to the flood of praise the latest date back to cézanne and bercy	What a laurel wreath. He has not been used for a long time to this deluge of praise, Dominique Strauss Kahn, the last praise dates from his Bercy years	what a laurel wreath dominique strauss kahn is no longer used to this deluge of praise the last ones go back to his last years in bercy
lorsqu ' il est le brillant ministre de l ' économie et des finances de lionel jospin une ascension brisée par l ' affaire de la mnef qui l ' oblige à quitter le gouvernement en quatre vingt dix neuf	when it the brilliant minister for the economy and finance lionel jospin a rise shattered by the make the mnef it obliges us to leave the government quatre-vingt-dix-neuf	when he was Lionel Jospin's brilliant economy and finance minister, a rise halted by the MNEF affair which forced him to leave the government in ninety-nine	when he was lionel jospin's brilliant economy and finance minister an ascent broken by the mnef case which forced him to leave the government in ninety nine
depuis il a été blanchi certes mais strauss kahn n ' a jamais vraiment retrouvé sa superbe un courant minoritaire dans le parti une désignation ratée à la présidentielle un siège de député à sarcelles un climat irrespirable bref dsk commençait à s ' ennuyer ferme	since it was certainly laundered but strauss-kahn has never really regained his superb to current minority in the party a designation missed a presidential a seat of member to sarcelles by climate irrespirable brief dsk began to annoy firm	since his name has been cleared of course but Strauss Kahn has never really regained his aura. A minority element in the party, a failed nomination for the presidential elections, an MP's seat in Sarcelles, a stifling atmosphere, in sort DSK was starting to get seriously bored	strauss kahn has been cleared since but has never really regained his superb a minority in the party a missed appointment to the presidential elections a parliamentary seat in sarcelles a stifling atmosphere in short dsk was beginning to be bored stiff
aujourd 'hui nicolas sarkozy et l ' europe lui offrent un job taillé à sa mesure ce qu ' il n ' a jamais cessé d ' être un économiste brillantissime europhile polyglotte et créatif	today nicolas sarkozy europe gives him a job taya its measure what it has never ceased to be a brilliant economist here am passes polyglot and creative	Today Nicolas Sarkozy and Europe are offering a job tailored to fit what he has never stopped being - a highly brilliant economist and a multilingual and creative Europhile	today nicolas sarkozy and europe offer him a job tailored to fit what he has never ceased to be a brilliant economist and creative multilingual europhile
à un moment où le fmi veut justement se racheter une conduite placer la barre au centre gauche réguler la mondialisation et donner du poids aux pays émergents	at a time when the imf is precisely to redeem itself set the bar the centre left regulate globalisation and give weight to the emerging countries	just at a time when the IMF wants to redeem itself, steer a centre-left course, regulate globalisation and give weight to emerging countries	at a time when the imf wants to redeem itself to raise the bar in the center left to regulate globalization and give weight to emerging countries

strauss kahn était donc la bonne personne à la bonne place en acceptant cette offre il trouve un rebond inespéré	strauss-kahn was the right person in the right place by accepting this offer it finds a lift unhoped-for	Strauss Kahn was therefore the right person at the right time, by accepting this offer he has found an unhoped-for way to bounce back	strauss kahn was therefore the right person at the right place by accepting this offer he finds an unexpected rebound
le dix huit vingt continue avec le journal de mickaël thébault bonsoir mickaël	the eighteen will continue with the newspaper michael thebault	The 6-8 programme continues with Mickael Thibaut's news feature. Good evening Mickael	the eighteen twenty continues with mickaël thébault's news bulletin mickaël
bonsoir	good evening	Good evening	good evening
quand l ' ouverture sarkozyenne fait des ravages au ps encore un éléphant quitte le troupeau	when the opening sarkozyenne wreaks havoc the ps in an elephant who think it is	When Sarkozy's openness policy wreaks havoc in the PS - another major figure leaves the herd	when sarkozy's political opening causes havocs in ps an elephant still leaves the herd
jack lang démissionne des instances dirigeantes du parti réactions à gauche et à droite dans un instant	lang resign of leadership of the party action birth will in a moment	Jack Lang has resigned from the party's leadership bodies - reactions from the left and the right in a moment	jack lang resigned from the leadership of the party reactions from the left and right in a moment
la peine maximale pour pierrot le fou pierre bodein condamné à la réclusion criminelle à perpétuité dont trente ans incompressibles	the maximum penalty for pierrot the mad stone botton condemned to the criminal locked in perpe which thirty years incompressible	Maximum sentence for Mad Pierrot. Pierre Bodein has been sentenced to life imprisonment with a minimum of thirty years	maximum penalty for pierrot le fou pierre bodein sentenced to life imprisonment with thirty years incompressible
verdict cet après midi de la cour d ' assises du bas rhin	verdict this afternoon the court of assizes bas-rhin	This afternoon's verdict at the Bas Rhin criminal court	verdict this afternoon of the assize court of the bas rhin
la polémique continue chez air france le personnel exige l ' arrêt des expulsions d ' étrangers sur ses lignes	polemic continues in air france staff requires the arrest of expulsions of foreigners to its limits	Controversy continues at Air France. Staff are demanding the end to expulsions of foreigners on its flights	controversy continues for air france the employees require the end of foreigners deportation on its lines
pas question de remettre en cause les décisions de l ' état lui répond la direction	no question to question the decisions of talus meets the direction	It is out of the question to dispute government decisions according to the management	state decisions will not be challenged replied the management
au pakistan cette fois c ' est officiel il n ' y a plus aucun islamiste retranché dans la mosquée rouge d ' islamabad mais le flou ce soir quant au bilan réel	in pakistan this time it is official there is no islamic retranchés in the red mosque in islamabad but the vagueness this evening about the actual record	In Pakistan, this time it's official - no more Islamists are entrenched in the Red Mosque in Islamad but the real situation remains vague this evening	it is now official in Pakistan there is no more entrenched islamist in islamabad red mosque but this evening things are still vague as to the actual record
le tour de france quatrième étape remportée aujourd 'hui par le norvégien thor hushovd	the tour de france fourth stage won today by the norwegian bulls this loft	The fourth stage of the Tour de France won today by the Norwegian Thor Hushovd	the tour de france fourth stage was won today by the norwegian thor hushovd
mais pas de changement au général le suisse cancellara reste en jaune journal du tour à suivre avec jean françois rhein	but not change the general when switzerland will remain yellow journal of turn ensures with jean-françois some	but no change to the overall rankings, the Swiss Cancellara stays in the yellow jersey. Tour magazine to follow with Jean François Rhein	but no change in the general the swiss cancellara remains in yellow diary of the tour to follow with jean francois rhein
quel temps pour demain cette fois météo france nous le promet ce ne sera pas encore l ' été pour tout le monde mais ça commencera enfin à y ressembler	some time for tomorrow this time we weather france the promises it will not yet been for everyone but it begins by also seem	what will the weather be like tomorrow ? This time the French Met Office promises it won't quite be summer for everyone but it will at last start to look like it	what the weather forecast for tomorrow this time météo france promises it will still not be summer for everybody but it will finally start to look like it

aide météo complète de joël collado dès la fin de ce journal	weather full joël gelas towers of the end of this journal	A full weather forecast by Joel Collado at the end of this news programme	joël collado at the end of this news bulletin
et juste après bien sûr le téléphone sonne d ' alain bédouet	of just after of course the telephone rings of alain bédouet	and just after, of course, "Le Téléphone Sonne" with Alain Bédouet	and of course right afterwards alain bédouet's téléphone sonne
question ce soir sur la situation dans les prisons françaises	questions this evening on the situation in the prisons french	tonight's subject is s the situation in French prisons	tonight debate on the situation in french prisons
vos commentaires vos questions dès maintenant au zéro un quarante cinq vingt quatre sept mille ou sur le site trois w point france inter point com	your comments your questions now at a zero forty-five twenty-four seven thousand -vous on the site www point france-inter .com	Your questions and comments as of now by phone on "oh" one forty-five twenty-four seven thousand or on the site www dot France Inter dot com	your questions your comments now at zero one four five two four seven zero zero zero or on the site three w dot france inter dot com
france inter	france-inter	france inter	france inter
était ce la stratégie de nicolas sarkozy une certitude en tous cas ce soir sa politique d ' ouverture fait comme un grand courant d ' air au ps	the strategy dess nicolas sarkozy confederation certainty in all cases this evening its policy of openness as a great course of ayrault ps	was this Nicolas Sarkozy's strategy ? One thing is sure this evening, his policy of openness is a blowing a new wind through the Socialist Party	was it nicolas sarkozy's strategy what is sure in any case tonight is that his political opening is like a big stream of air in the socialist party
et dans ces cas là hé bien les portes claquent	and in these cases is well the doors claquent	and in those cases, well, doors tend to slam	and in these cases well doors slam
jack lang démissionne donc des instances dirigeantes du parti il l ' a annoncé ce matin dans une lettre à françois hollande	every angle resign therefore 's governing bodies of the party he announced this morning in a letter to françois holland	Jack Lang has resigned from the party's leadership committees, he announced this today in a letter to François Hollande	jack lang therefore resigns of the party leadership he announced this morning in a letter to françois hollande
je ne me reconnais plus dans tes méthodes de direction dit il	i am not recognise more tame methods of direction said -	I no longer can identify with your leadership methods, he said	I no longer recognize myself in your methods of leadership he said
réponse directe évidemment à cette décision hier soir du bureau national d ' exclure tout membre qui participerait à une commission mise en place par le gouvernement	direct answer clearly enough decisions yesterday evening the bureau national the excluded return member who participate in a committee set up by the government	Obviously a direct answer to the National Bureau's decision last night to expel any members who take part in a commission set up by the government	obviously a direct response to last night's national office decision to exclude any member who participates in a commission set up by the government
or jack lang vous le savez est invité par nicolas sarkozy à participer à une commission sur la réforme des institutions	however lang you know is invited by nicolas sarkozy to participate in a commission on the reform of institutions	Now as you know Jack Lang was invited by Nicolas Sarkozy to take part in a commission on the institutional reform	now jack lang you know is invited by nicolas sarkozy to attend a commission on institutional reform
mais après six personnalités venues du ps ou de la gauche qui sont maintenant au gouvernement plus dominique strauss kahn	but after six personalities from the ps or of the left which are now the government more dominique strauss-kahn	but after the six other figures from the Socialist Party or the left in general who are now in the government plus Dominique Strauss Kahn	but after six personalities from ps or the left who are now in government plus dominique strauss kahn
qui ne pense plus qu ' à la direction du fmi hé bien cela commence à faire beaucoup françoise degois	who are thinking more that the direction of the imf well it begins to make great françoise=degois	whose only thoughts are for the IMF leadership post, well, that's starting to get a bit much Françoise Degois	who thinks only of the leadership of the imf well this is beginning to be a little too much françoise degois
on ne voit pas pourquoi nicolas sarkozy s ' arrêterait en si bon chemin	we see no reason why nicolas sarkozy has would stop in if good path	It's hard to see why Nicolas Sarkozy would stop when things are going so well	we do not see why nicolas sarkozy would stop after such a good start
il a trouvé la clé flatter les égos faire miroiter du prestige et de la reconnaissance à des acteurs politiques qui ne voient qu ' un ciel bouché depuis le six mai	it has found the key flatter the echoes hold out of the prestigious recognition will political actors who only see sky bouché since the six may	He has found the way to flatter egos and tempt political figures who have only seen empty horizons since May 6th with prestige and recognition	il a trouvé la clé flatter les égos faire miroiter du prestige et de la reconnaissance à des acteurs politiques qui ne voient qu ' un ciel bouché depuis le six mai

dominique strauss kahn à sa manière et jack lang à la sienne sont les deux dernières prises de guerre en date mais d ' autres devraient suivre	dominique strauss-kahn on its way ed lang the chain are the last two taken of war in date but other should follow	Dominique Strauss Kahn in his way and Jack Lang in his are the two latest taken prisoner but others are likely to follow	dominique strauss kahn in his own way and jack lang in his are the last two booty dated but others should follow
le dix huit vingt continue avec le journal de mickaël thébault bonsoir mickaël	the eighteen will continue with the newspaper michael thébault goodbye michael	The 6-8 programme continues with Mickael Thibaut's news feature. Good evening Mickael	the eighteen twenty continues with mickaël thébault's news bulletin good evening mickaël
bonsoir	a little good under	Good evening	good evening
pas facile de faire de l ' ombre à nicolas sarkozy sauf peut être à porter le même nom	it was not easy to make the shadow to nicolas sarkozy except perhaps to carry the same name	Not easy to overshadow Nicolas Sarkozy unless perhaps you have the same name	not easy to overshadow nicolas sarkozy except perhaps by bearing the same name
la preuve ce soir avec cécilia sarkozy qui vole la vedette à son mari de président le soir d ' épinal jour de la présentation de sa grande réforme des institutions	proof tonight with cecilia sarkozy stealing the show her husband president not the evening of final the day the presentation of its major reform of the institutions	Proof tonight with Cécilia Sarkozy who stole the spotlight from her husband the president this evening in Epinal on the big day of the presentation of his major institutional reforms	the evidence tonight with cecilia sarkozy who steals the limelight from her president husband in the evening of epinal day of the presentation of his major reform of the institutions
vous entendrez le chef de l ' état et nous retrouverons sur place notre envoyé spécial jean françois aquili	you will hear the head of state and we are in place in city special jean-françois who say	you will hear the head of state and we will go over to our reporter there Jean François Aquili	you will hear the head of state and we will meet again our correspondent jean francois aquili
et au ps pendant ce temps c ' est la crise crise qui s ' envenime la faute à l ' ouverture sarkozyenne mais aussi à françois hollande qui aurait dû démissionner à temps c ' est ce que nous dira le député socialiste de la nièvre gaëtan gorce	the english during this time the crisis which is escalating crisis the blame to the opening sarkozyenne but also françois holland said we should have resigned expects that is what we tell the socialist member of the nièvre gaétan gorce	And meanwhile in the PS, it's crisis time, a crisis that's worsening because of the Sarkozian policy of openness but also because of François Hollande who should have resigned in time as Gaetan Gorce, the Socialist MP for Nievre will tell us	in ps meanwhile it is the crisis a crisis that is aggravating the blame is on the sarkozy's opening but also on francois hollande who should have resigned in time that is what we will be told by the socialist nievre member of the parliament gaetan gorce
george bush lui est toujours résolument optimiste les états unis peuvent gagner la guerre en irak assure le président américain vous l ' entendrez discours très offensif alors que le congrès rend lui un rapport très critique	george bush him is always resolutely optimistic the united states can win the war in iraq ensures the american president in wanting to andré a speech very offensive while the congress is a very critical report	As for George Bush, he remains resolutely optimistic. The United States can win the war assures the American president as you will hear. A very bullish speech that comes just as the Congress produces a highly critical report	george bush is still resolutely optimistic the united states can win the war in irak u.s. president assures you will hear him very offensive speech while the congress makes a highly critical report
dans l ' actualité également ce soir l ' examen du tepa le paquet fiscal l ' une des mesures emblématiques le crédit d ' impôts sur les intérêts d ' emprunts vient d ' être adopté le commentaire de la ministre des finances christine lagarde dans un instant	in the topical also this evening examination of the starting the fiscal package one measures emblematic credit to tax the interests in three has just been adopted the comment of finance ministers christine lagarde in a moment	Also in the news tonight, the examination of the tax package, one of the emblematic measures, the tax credit for interest on loans has just been adopted. The comments of the Finance Minister Christine Lagarde in a moment	in the news tonight also an examination of the tax package tepa one of the flagship measures the tax credit on loans interests has just been passed finance minister christine lagarde's comments in a moment

le tour de france cinquième étape remportée aujourd 'hui par l ' italien filippo pozzato le suisse cancellara reste en jaune	the tour de france fifth step won today by the italian filippo fine trappings the swiss when it is lara remain yellow	The fifth stage of the Tour de France was won today by the Italian Filippo Pozzato, the Swiss Cancellera retains the yellow jersey	the tour de france fifth stage was won today by the italian filippo pozzato the swiss cancellara remains leader
les sprinters eux ont commencé à souffrir il y avait huit cols à franchir à suivre le journal du tour avec jean françois rhein	the sprinters can you are starting this suffer it big cola cross to follow the newspaper the tour with jean-françois kidney	As for the sprinters, they are starting to suffer. There were eight hill climbs, news in the Tour feature with Jean François Rhein	the sprinters have started to suffer there were eight passes to cross the diary of the tour to follow with jean francois rhein
le ciel de demain et enfin enfin l ' arrivée du soleil ça s ' arrange absolument partout avec des températures qui devraient remonter en flèche	the sky tomorrow finally and finally the arrival of sun it has helped absolutely wherever with temperatures which should go spiralling	Tomorrow's weather and at last the sun appears. Everything is working out fine absolutely everywhere with temperatures which should shoot up	the sky tomorrow and finally finally the arrival of the sun it is getting absolutely better everywhere with temperatures expected to soar
météo complète de joël collado dès la fin de ce journal	weather comprehensive play the back at the end of this newspaper	Full weather forecast by Joel Collado at the end of this news programme	complete weather forecast by joël collado at the end of this news bulletin
et juste après bien sûr le téléphone sonne animé par alain bédouet thème ce soir la république intégration et ascenseur social avec cette question la garde des sceaux rachida dati est elle un cas particulier ou exemplaire	just after the on the telephone rings lively by alain bédouet subject this evening the republic of integration and social lift with this question the custody of seals mrs dati - it is a particular case you exemplary	and just after, of course, "Le Téléphone Sonne" with Alain Bédouet tonight's subject is the Republic, integration and the social elevator with this question. Is the Minister of Justice Rachida Data an isolated case or an exemplary one ?	and of course right afterwards le téléphone sonne run by alain bedouet its theme tonight republic integration and social mobility with this question is the french minister of justice rachida dati a particular or exemplary case
vos commentaires vos questions dès maintenant au zéro un quarante cinq vingt quatre sept mille ou par internet trois w point france inter point com	comments before your questions now at a zero forty-five twenty-four seven thousand or by internet www point france-inter	Your questions and comments as of now by phone "oh" one forty-five twenty-four seven thousand or on the site www dot France Inter dot com	your comments your questions now at zero one four five two four seven zero zero zero or by internet three w dot france inter dot com
france inter	france inter	France Inter	france inter
ce devait donc être un moment quasi historique pour nicolas sarkozy ce soir à épinal soixante et un ans après la déclaration du général de gaulle dans cette même ville	this should therefore be a historic moment in case for nicolas sarkozy tonight to epinal soixante-et-un years after the declaration of general gaule in this city	It was supposed to be a near-historic moment for Nicolas Sarkozy tonight at Epinal, sixty-one years after General de Gaulle's declaration in the same town	it should therefore have been a quasi historical moment for nicolas sarkozy tonight at epinal sixty-one years after general de gaulle's declaration in the same city
le chef de l ' état devait lui présenter les grandes lignes de sa réforme des institutions alors il va le faire bien sûr	the head of state , was to present the broad lines of its reform of the institutions then it will do of course	The head of state was to present the outline of his institutional reforms well he is going to do so of course	the head of state should submit the outline of his institutional reform so he will do it of course
mais depuis son arrivée ce n ' est pas de cela dont tout le monde parle mais bel et bien de cécilia sarkozy la première dame de france qui est ce soir à tripoli	do since its arrival is the start of what everyone talks but of properly cecilia sarkozy the first lady of france who is this evening in tripoli	but since his arrival it is not that which everyone is talking about but rather Cécilia Sarkozy, first lady of France who is in Tripoli tonight	but since his arrival everyone is talking not about that but well and truly about French first lady cecilia sarkozy who is tonight in tripoli
l ' épouse du président est allée tout à l ' heure à la rencontre des infirmières bulgares condamnées à mort	the hens of president went to earlier to the meeting of the bulgarian nurses condemned to death	Earlier the president's wife went to meet the Bulgarian nurses who have been condemned to death	the president's wife went recently to meet the bulgarian nurses sentenced to death
nicolas sarkozy dès son arrivée n ' a pas pu éviter l ' explication écoutez mais tendez l ' oreille	nicolas sarkozy from its arrival has not been able to avoid explanation listen to putting the laurels	Since his arrival Nicolas Sarkozy has not been able to avoid explaining. You will need to listen closely	nicolas sarkozy on his arrival has not been able to avoid the explanation to listen to it prick up your ears

non non elle ne ne revient pas elle est toujours	non non shoe does not come back she is still	No, no, she isn't coming back, she is still...	no no she is not comming back she is always
elle vient de voir les les infirmières	she comes to see the nurses	She has just seen the nurses	she has just seen the nurses
et elle est partie	and she left	and she has left	and she has gone
de tripoli pour aller voir les enfants	from tripoli to go see the children	Tripoli to go and see the children	from tripoli to see the children
qui ont contracté le le sida	who caught aids	who have caught AIDS	who have contracted AIDS
elle est toujours là bas parce qu ' elle aura ce soir une nouvelle rencontre avec le colonel kadhafi	she is still there because she will have a new encounter with colonel kadhafi tonight	she is still there because tonight she will have another meeting with Colonel Gadaffi	she is still there because she will meet again with colonel gaddafi tonight
et c ' est à dire qu ' il y aura donc un espoir quand même une issue une bonne issue	and that is to say there will be hope even an outcome a good outcome	and that's to say there will therefore be a hope nonetheless, an outcome, a good outcome	so that is to say that there will be hope an outcome a good outcome
elle commentera elle même	she will comment herself	She will comment herself	she will comment on that herself
madame monsieur bonsoir comme chaque dimanche et dans le cadre de notre rétrospective hebdomadaire retraçant les activités royales de la semaine le rappel	madam mr good evening as sunday in the framework of our retrospective weekly volume tracing the royal activities of the week the reminder	Madam, sir, good evening . Like each Sunday and as part of our weekly review of royal activities this week a reminder	good evening sir madam like every sunday and as part of our weekly retrospective tracing the royal activities of the week the recall
activités marquées notamment par le lancement et l ' inauguration de nombreux projets socio économiques à nador et oujda le rappel avec hasna boufkiri	activity marked in particular by the countries inauguration launch new projects socio-economic fernando remind you with sam	Activities marked most notably by the launch and inauguration of numerous socio-economic projects in Nador and Oujda. A summary with Hasna Boufkiri	activities marked by the launch and the inauguration of many social and economic projects in nador and oujda the recall with hasna boufkiri
en début de semaine sa majesté le roi mohamed six a procédé à la pose de la première pierre pour la construction	at the beginning of the week his majesty the king mohammed six has made the laying of the foundation stone for the construction	At the start of the week His Majesty King Mohammed the Sixth laid the first stone for the construction	earlier this week his majesty king mohamed six laid the foundation stone for the construction
de la nouvelle aérogare de l ' aéroport oujda angad qui s ' inscrit dans le cadre du projet d ' extension de l ' aéroport	the new aérogare of the airport oujda a quest which comes within the framework of the draft extension of the airport	of the new terminal at Oujda Angad airport which is part of the planned extension of the airport	of the new terminal of oujda angad airport which comes within the framework of the proposed extension of the airport
pour lequel a été débloquée l ' enveloppe budgétaire de six cent cinquante millions de dirhams	what has been allocated a budget six hundred and fifty million dirhams	which has a budget of six hundred and fifty million dirhams	for which a budget of six hundred fifty million dirhams has been released
mardi sa majesté le roi mohamed six a inauguré au quartier ennajd d ' oujda un espace de vie qui vise à améliorer les conditions d ' accès des populations locales aux services de base	On Tuesday, His Majesty King Mohammed the Sixth inaugurated a centre in the area of Ennadj in Oujda aimed at improving conditions of access to basic services for the local population	tuesday his majesty the king mohammed six has inaugurated in district snows becomes a species which aims to improve the conditions of access of local populations basic service	tuesday his majesty king mohamed six inaugurated in the ennajd district of oujda a living space that aims to improve access conditions to basic services for local populations
à promouvoir la politique de proximité et à assurer des qualifications professionnelles aux jeunes pour leur insertion socio économique dans le cadre de l ' indh	to promote the policy of proximity and ensure a professional qualifications of young for their inclusion socio-economic within the framework of the idea p.m.	promoting the policy of proximity and providing professional qualifications for young people to help their socio-economic insertion as part of the INDH	to promote community politics and to provide professional qualifications for young people for their socio economic integration within the nhri

à cette occasion sa majesté le roi a remis des titres d ' affectation aux premiers bénéficiaires de la deuxième tranche de cette opération et des équipements et matériels divers à plusieurs associations actives au sein de la ville	on this occasion his majesty the king securities of spending the prime beneficiary of the second tranche of this operation equipment and material various several associations working in the city	On this occasion, His Majesty the King gave titles of appointment to the first beneficiaries of the second budget stage of this operation and varied equipment and materials to several associations working in the town	on this occasion his majesty the king gave employment titles to the first beneficiaries of the second phase of this operation and various equipment and material to several active associations in the city
par la suite le souverain s ' est enquis de plusieurs projets qui seront réalisés dans le cadre de l ' indh pour un coût de treize millions de dirhams	subsequently the sovereign has asked several projects which will be made in the framework of the island age for one of thirteen million dirhams	Then the sovereign enquired about several projects which will be carried out as part of the INDH at a cost of thirteen million dirhams	then the sovereign inquired about several projects to be implemented as part of the nhri at the cost of thirteen million dirhams
mercredi toujours sa majesté le roi a donné au centre la selma dans la province de nador le coup d ' envoi du programme d ' habitat kebdana il vise la revalorisation du paysage urbanistique de cette localité	always says his majesty the king to go at the centre to the seven éme art in the province of programme of habitat that it vénère if the recovery of urban landscape of this site	Also on Wednesday at the Selma centre, His Majesty the King launched the Kebdana housing programme. This aims to redevelop the urban landscape in this area	still on wednesday his majesty the king gave to selma center of the nador province the kickoff of the housing program kebdana it is designed to redevelop the urban landscape of this town
sa majesté le roi mohamed six a procédé ensuite à l ' inauguration de l ' aire de repos ras el maa réalisée par la fondation mohamed cinq	his majesty the king mohammed six has made then the opening of the this done by the foundation mohamed five	His Majesty King Mohammed the Sixth then went to the inauguration of the Ras El Maa rest area built by the Mohammed the Fifth Foundation	his majesty king mohamed six then proceeded to the inauguration of the rest area ras el maa carried out by mohamed five foundation
pour la solidarité dans le cadre de l ' opération d ' accueil de marocains résidant à l ' étranger marhaba deux mille sept	for solidarity in the framework of the operation is moroccan resident abroad after 2007	For Solidarity as part of the Marhaba two thousand and seven welcome programme for Moroccans living abroad	for solidarity in the context of the operation of reception of moroccans living abroad marhaba two thousand and seven
cette structure a mobilisé des fonds de l ' ordre de quatre millions de dirhams	this structure has mobilised funds of the order of four million dirhams	this structure has received funds of around four million dirhams	this structure has raised funds of around four million dirhams
dans l ' après midi de ce mercredi un accueil des plus chaleureux a été réservé au souverain par les habitants de la ville de nador	in this afternoon of wednesday a most warm welcome was booked at the king by inhabitants of the city of nador	On Wednesday afternoon, the sovereign received the warmest of welcomes from the inhabitants of the town of Nador	in the afternoon of wednesday a warm welcome was reserved for the sovereign by people of the city of nador
qui s ' étaient massés tout au long des artères empruntées par le cortège royal pour souhaiter la bienvenue à sa majesté le roi mohamed six	it is one that is a arteries taken by the royal trail in welcoming his majesty the king mohammed vi	who had gathered along the roads taken by the Royal procession to welcome His Majesty King Mohammed the Sixth	who were massed along the arteries used by the royal procession to welcome his majesty king mohamed six
et exprimer leur joie pour cette visite aux retombées bénéfiques pour la ville et sa population	and express their joy for this visit retombés beneficial for the city its population	and express their joy at this visit with its benefits for the town and its population	and express their joy for this visit whith the benefits for the city and its people
jeudi sa majesté le roi mohamed six s ' est enquis du programme de restructuration et de viabilisation du site touristique marchica	thursday his majesty the king mohammed vi has asked me about the restructuring programme and filialisation site tourist monique sicard	On Thursday His Majesty King Mohammed the Sixth enquired about the restructuring and provision of services programme for the tourist resort, Marchica	thursday his majesty king mohamed six asked about the restructuring program of the tourist site marchica
pour lequel ont été mobilisés des investissements de l ' ordre de onze milliards de dirhams	for which have been mobilised investment in the order of eleven billion dirhams	which has received investments of around eleven billion dirhams	for which investments in the order of eleven billion dirhams were mobilized

cette semaine sa majesté le roi mohamed six amir al mu minin accompagné de son altesse le prince moulay ismaïl	this week his majesty the king mohammed six thousand mines with his highness the prince moulay ismail	this week, His Majesty King Mohammed the Sixth Amir Al Mu Minin accompanied by his Highness Prince Moulay Ismaïl	this week his majesty king mohamed six amir al mu minin accompanied by his highness prince moulay ismail
a accompli la prière du vendredi à la mosquée lalla amina de nador	done prayer friday to the mosque the line dakar	went to Friday prayers at Lalla Amina Mosque in Nador	completed the friday prayer at the mosque lalla amina of nador
à l ' issue de la prière sa majesté le roi mohamed six accompagné de son altesse le prince moulay ismaïl s ' est enquis du bilan et du programme d ' action de l ' initiative nationale pour le développement humain au niveau de la province de nador	following the prayer his majesty the king mohammed six accompanied by his highness the prince moulay ismail has asked me about the assessment of the action programme of the national initiative for human development at the level of the province of nador	after prayers, His Majesty King Mohammed the Sixth Amir Al Mu Minin accompanied by his Highness Prince Moulay Ismaïl enquired about the results and action programme of the national initiative for human development in the province of Nador	at the end of the prayer his majesty king mohamed six accompanied by his highness prince moulay ismail asked about the assessment and the action program of the national initiative for human development in the nador province
à cette occasion des explications ont été fournies au souverain sur le projet de reconstruction d ' un centre relevant de l ' organisation alaouite	you are men explanations have been provided to the sovereign on the construction of a centre under the organisation go quickly	on this occasion, explanations were given to the King about the project to reconstruct a centre belonging to the Alaouite organisation	on this occasion explanations were provided to the sovereign about the reconstruction project of a center within the alawite organization
pour la protection des aveugles à nador qui sera réalisé pour un coût de près de quatre millions de dirhams et qui profitera à quarante malvoyants	for the protection of was a player that will be done at the cost of four million dirhams and that will benefit forty visually impaired	for the protection of the blind in Nador which will be carried out at a cost of nearly four million dirhams and will benefit forty visually impaired people	for the protection of the blind in nador which will be achieved at a cost of nearly four million dirhams and that will benefit to forty visually impaired
toujours ce vendredi sa majesté le roi mohamed six chef suprême et chef d ' état major général des forces armées royales accompagné de son altesse le prince moulay ismaïl a procédé à la pose de la première pierre	always this friday his majesty the king mohammed six supreme leader and head of general staff of armed forces with his royal highness prince moulay ismail proceeded to the poses the first stone	Also on Friday, His Majesty King Mohammed the Sixth, Supreme Leader and State Head and Major General of the royal armed forces accompanied by his Highness Prince Moulay Ismaïl laid the first stone	still on friday his majesty king mohamed six supreme leader and chief of general staff of the royal armed forces accompanied by his highness prince moulay ismail laid the first stone
pour la construction d ' une cité intégrée au profit du personnel militaire en activité à la place d ' armes de nador un projet pour lequel ont été consacrés des fonds de l ' ordre de soixante millions de dirhams	for the construction of an integrated city to the benefit of military staff activity instead of arms nador a project for which were dedicated funding of about sixty million dirham	for the construction of an integrated housing estate for military personnel serving at the Place d'Armes in Nador, a project which has received a budget of around sixty million dirhams	for the construction of a city built for the benefit of the active military staff of nador fortress a project for which sixty million dirhams funds were spent
direction à présent la palestine le premier ministre israélien ehud olmert et le président palestinien mahmoud abbas se rencontreront demain lundi à al qods	it on gaza palestine the israeli prime minister ehud olmert and palestinian president mahmoud abbas meet tomorrow monday while is said	Now, direction Palestine. The Israeli Prime Minister Ehud Olmert and the Palestinian president Mahmoud Abbas will be meeting tomorrow, Monday, in Al Gods	now direction palestine israeli prime minister ehud olmert and palestinian president mahmoud abbas will meet tomorrow monday in Jerusalem
et selon un haut responsable sécuritaire palestinien cent quatre vingt neuf membres des brigades des martyrs d ' al aqsa ont accepté de cesser leurs attaques anti israéliennes en cisjordanie	and in a responsible palestinian security one hundred quatre-vingt-neuf member of ' brigades al-aqsa martyrs to have agreed to stop their attacks opposed israeli in west bank and	and according to a top Palestinian security chief, one hundred and eighty nine members of the Al-Aqsa Martyrs' Brigades have agreed to stop their anti-Israeli attacks on the West Bank	and according to a high palestinian security official one hundred eighty nine members of al aqsa martyrs brigades agreed to cease their attacks against israel in west bank

mohamed aït ichkar	mahmoud kitsch	Mohamed Aït Ichkar	mohamed aït ichkar
france inter	france inter	France Inter	france inter
la différence la différence	the difference the difference	the difference the difference	the difference the difference
douze heures six pile ça vous dérange c ' est avec nicolas stoufflet et c ' est tout de suite	there are twelve gold six thousand to it disturbs you this with nicolas cissé trying immediately	It's six minutes past twelve exactly. "Ca vous dérange" is with Nicolas Stoufflet and it's next	dead on six past twelve 'ça vous dérange' it is with nicolas stoufflet and this is immediately
merci mickaël thebault les prochaines infos le journal de treize heures avec fabrice drouel	thank you michael thébault the next infos newspaper thirteen hours with fabrice=drouelle	Thank you Mickaël Thebault, the next news programme will be at one p.m. with Fabrice Drouel	thank you mickaël thébault upcoming news the one p.m news with Fabrice Drouelle
un petit d ' extrait du film ennemi d ' état de tony scott bonjour à toutes et à tous les caméras vous regardent	a small part of the film enemy piles of tony scott hello to all the cameras you look	A short extract from the film "Enemy Of The State" by Tony Scott. Hello everyone, the cameras are looking at you	a small extract from the movie 'State enemy' by tony scott hello to you all the cameras are watching you
sans tomber dans la paranoïa et sans doute sans se référer forcément au big brother du du roman de science fiction de george orwell mille neuf cent quatre vingt quatre on peut légitimement s ' interroger sur le développement de la vidéo surveillance	without falling into the paranoia will undoubtedly without necessarily refer to big brother of the novel fiction of george orwell nine hundred thousand quatre-vingt-quatre we can legitimately be questioned on the development of the vidéosurveillance	without falling into paranoia and probably without necessarily referring to the Big Brother of George Orwell's science fiction novel Nineteen-Eighty-Four, it is still legitimate to wonder about the development of closed circuit television	without falling into paranoia and probably without necessarily referring to the big brother of the science fiction novel by george orwell nineteen eighty-four we can legitimately question the development of video surveillance
dans les rues les parkings les bureaux les magasins les squares	in the streets car parks offices shops the squares	in the streets, car parks, offices, shops, squares	in streets car parks offices stores public gardens
on estime à quatre cents mille le nombre de caméras actuellement en france	an estimated four hundred thousand the number of cameras currently in france	there are an estimated four hundred million cameras in France currently	it is estimated there are four hundred thousand cameras currently in France
nous sommes dans un pays démocratique certes mais on voit que les démocraties menacées par le terrorisme	we are in a democratic country certainly but we see that the democracies threatened	we are in a democratic country for certain but we can see that democracies threatened by terrorism	we are certainly in a democratic country but we see that democracies threatened by terrorism
ont besoin de limiter peut être un peu la liberté de ses citoyens pour les protéger	by terrorism need to limit perhaps a little freedom of its citizens to protect them	need to limit the liberty of their citizens a little to protect them	may need to limit a little the freedom of their citizens to protect them
en grande bretagne où la vidéo surveillance est très répandue depuis longtemps	in britain vidéosurveillance widespread long	in Great Britain where closed circuit television has been widespread for a long time now	in Britain where video surveillance has been widespread for a long time
des personnes suspectées d ' actes terroristes ont été récemment arrêtées grâce à la présence de caméras	of those suspected of terrorist acts have been recently arrested thanks to the presence of cameras	people suspected of terrorist acts were recently arrested thanks to the presence of cameras	persons suspected of terrorist acts have recently been arrested because of the presence of cameras
en france pour prévenir les actes de délinquance ou pour identifier des individus certaines villes se sont dotées d ' un réseau de caméras et ce n ' est que le début semble t il	in france to prevent acts of delinquency or to identify certain individuals cities have with a network of cameras and this is the beginning seems t it	In France, to prevent crime or to identify people, certain towns have equipped themselves with a network of cameras on public transport and this is just the start it seems	in France to prevent criminal acts or identify individuals some cities have installed a network of cameras and it seems this is just the beginning

le président de la république nicolas sarkozy a parlé récemment dans une interview au journal du dimanche de l ' installation possible de caméras dans les transports en commun	the french president nicolas sarkozy appears recently in an interview sunday newspaper of the plant possible cameras in public transport	The president of the republic Nicolas Sarkozy recently gave an interview to the "Journal de Dimanche" in which he spoke about the possible installation of cameras on public transport	president nicolas sarkozy spoke recently in an interview with 'journal du dimanche' newspaper of the possible installation of cameras in public transports
pensez vous que la vidéo surveillance nous protège ou qu ' elle empiète sur nos droits qu ' elle viole nos droits parfois	vous think that the vidéosurveillance protects us ok will pierrette on our rights that it violates our rights sometimes	Do you think closed circuit television protects us or does it encroach on our rights, violate our rights sometimes ?	do you think the video surveillance protects us or that it infringes on our rights that it violates our rights sometimes
vous pouvez donner votre avis comme vous l ' avez déjà fait d ' ailleurs comme vous avez commencé à le faire tout le week end sur france inter point com	you can give your opinion as you yourself have already done elsewhere as you have begun to make the all the weekend on franceinter .com	You can give your opinion, as you have already have and moreover have started doing, all weekend on France Inter dot com	you can give your opinion as you have already done as you started to do all weekend on france inter dot com
vous pouvez aussi nous appeler au zéro un quarante cinq vingt quatre sept mille tout de suite questions opinions témoignages tout cela est bienvenu	you can also call us at a zero forty-five twenty-four seven thousand immediately question opinion testimony all this is welcome	You can also call us on "oh" one forty-five twenty-four seven thousand straight away questions, opinions, anecdotes, all that is welcome	you can also call us at zero one four zero five two zero four seven zero zero zero immediately questions opinions evidence all that is welcome
invité de ça vous dérange aujourd 'hui patrick balkany député maire ump de levallois perret dans les hauts de seine bonjour patrick balkany	invited that disturbs you today patrick balkany member ump mayor of levallois perret in senior de seine hello patrick balkany	The guest on "Ca Vous Dérange" today is Patrick Balkany, UMP deputy mayor of Levallois Perret in the Hauts de Seine region. Hello Patrick Balkany	invited of 'ça vous dérange' today ump member of parliament and mayor of levallois perret in the hauts de seine patrick balkany hello patrick balkany
bonjour	hello	Hello	hello
levallois une ville surveillée par des caméras depuis une vingtaine d ' années maintenant	levallois a city monitored by cameras since another 20 years	Levallois, a town with camera surveillance for the last twenty years or so now	levallois a city monitored by cameras for twenty years now
sécurisée	made safe	secured	
sécurisée dites vous par les caméras depuis une vingtaine d ' années vous avez été des des pionniers en la matière à levallois perret	secure say -vous by the cameras since the twenty you have the pioneers in the area to levallois perret	made safe, you say, by cameras over the last twenty years. You were a pioneer in this field in Levallois Perret	secured you say by the cameras for the past twenty years you have been pioneers in this field in levallois perret
oui il y a élu	yes viard	Yes there is elected	yes there is elected
parce que le problème de la sécurité de mes administrés se posait déjà à ce moment là	because the security problem of my constituents for would already in that time	because the problem of the safety of my constituents had arisen at that time	because the problem of my constituents' security was already raised at that time
créer un réseau de vidéo surveillance qui pour moi est un outil	creating a network of video surveillance which for me is a tool	for me, setting up a closed circuit television network is a tool	to create a video surveillance network that is a tool for me
les policiers à faire leur travail	policemen to do their work	the police to do their job	police to do their job
face à vous marie claude bonneville secrétaire de l ' association souriez vous êtes filmés	addressing you marie-claude bonneville secretary of the association souriez you are recorded on	opposite you, Marie-Claude Bonneville, secretary of the association "Smile - You're On Camera"	facing you marie claude bonneville secretary of the association souriez vous êtes filmés
bonsoir la province de nador de nouveau à la une	good evening the province of my crystal again to the one	Good evening, the Province of Nador is once again in the headlines	good evening nador province again at the front page

à beni ensar sa majesté le roi mohamed six accompagné de son son altesse royale le prince moulay rachid a procédé	with a room to her majesty king mohammed six accompanied by his royal highness prince moulay rachid proceeded	In Beni Ensar, His Majesty King Mohammed the Sixth accompanied by his Highness Prince Moulay Ismaïl laid	at beni ensar his majesty king mohamed six accompanied by his royal highness prince moulay rachid laid
à la pause de la première pierre pour la construction d ' un centre de qualification professionnelle maritime et à el aroui le souverain a inauguré l ' hôpital mohamed six	to break the foundation stone for the construction of a centre of professional qualification maritime and not to the louis the sovereign opened the hospital mohamed six hours	the first stone for the construction of a professional maritime qualification centre and in El Aroui, the sovereign inaugurated Mohammed the Sixth hospital	the foundation stone for the construction of a maritime professional qualification center and at el aroui the sovereign inaugurated mohamed six hospital
le port de beni ensar près de neuf cents barques actives dans le domaine de la pêche artisanale	the recovery life it only almost nine hundred barques active in the field of small-scale fisheries	The port of Beni Ensar, nearly nine hundred boats working in the field of artisan fishing	port beni ensar nearly nine hundred boats operating in the field of artisanal fisheries
deux cents autres opèrent dans le port de ras kebdana à côté d ' une flotte tout aussi importante de la pêche côtière	its centre will lose honduras gives side of a fleet equally important inshore fisheries	two hundred other boats work out of the port of Ras Kebdana alongside just as large a coastal fishing fleet	two hundred others operate in the port of ras kebdana next to an as important coastal fisheries fleet
ce secteur emploie quatorze mille pêcheurs et pour accompagner son développement un outil essentiel	this sector employs fourteen thousand fishermen to support its development an essential tool	This sector employs fourteen fishermen and to help its development, an essential tool	this sector employs fourteen thousand fishermen and to support its development an essential tool
le nouveau projet de centre de qualification professionnelle maritime dont les travaux de construction ont été lancés aujourd 'hui par sa majesté le roi mohamed six	the new project the centre of professional qualifications maritime which the construction work it is denounced today by his majesty the king mohammed vi	the new professional maritime qualification centre project whose construction was launched today by His Majesty King Mohammed the Sixth	the proposed new maritime professional qualification center whose construction has been launched today by his majesty king mohamed six
la nouvelle structure sera opérationnelle dès l ' année prochaine	the new structure will be operational as of next year	this will be operational as early as next year	the new structure will be operational next year
à cette occasion les programmes de formation dans les cycles de spécialisation et de qualification au niveau de ce nouveau centre	on this occasion the training programmes whose rounds of specialisation qualification of level of this new centre	on this occasion the training programmes for the specialist and qualification courses at the new centre	on this occasion the training programs of the specialization and qualification courses at the new center
ont été remis à sa majesté le roi mohamed six	to her majesty king mohammed	were presented to His Majesty King Mohammed the Sixth	were presented to his majesty king mohamed six
le nouveau centre dispensera une formation annuelle à deux cents stagiaires dont cent cinquante dans les disciplines de spécialisation et cinquante dans celles de qualification	the new centre provides training linked to one of its trainees one hundred and fifty in disciplines specialisation and fifty in disciplines qualification	the new centre will provide a year's training for two hundred trainees, one hundred and fifty in the specialized disciplines and fifty in those of qualification	the new center will provide annual training to two hundred students including one hundred fifty in the fields of specialization and fifty in those of qualification
le centre mènera des actions de formation continue de vulgarisation et d ' alphabétisation fonctionnelle au profit des gens de la mer de la région	the centre will lead training actions continues to popularise and literacy functional to the benefit of seemen and of the region	the centre will also run continuous training programmes in popular science and literacy for the peoples of the sea in the region	the center will conduct in-house training popularization extension and elimination of functional illiteracy for the benefit of the area seafarers
par la suite sa majesté le roi mohamed six toujours accompagné de son altesse royale le prince moulay rachid	over the days of their of the rich subsequently his majesty king mohammed vi of the day wali and his royal highness prince moulay rachid	next, His Majesty King Mohammed the Sixth accompanied by his Highness Prince Moulay Rachid	thereafter his majesty king mohamed six always accompanied by his royal highness prince moulay rachid

s ' est rendu dans la ville d ' el aroui où le souverain a inauguré l ' hôpital mohamed six réalisé pour un coût de près de trente neuf millions de dirhams	it is true said in the crisis lakes louis where the sovereign opened the hospital it is done at a cost of nearly thirty-nine million dirhams	visited the town of El Aroui where the sovereign inaugurated Mohammed the Sixth hospital which was built at a cost of nearly thirty-nine million dirhams	went to the town el aroui where the sovereign opened mohamed six hospital achieved at a cost of nearly thirty nine million dirhams
le souverain a effectué une tournée dans les différentes dépendances de cette unité hospitalière construite sur une superficie de deux hectares	the sovereign made a tour of the different buildings that make up this two-hectare hospital site	the sovereign toured the various dependencies of the hospital unit built on an area of two hectares	
d ' une capacité de quarante sept lits cet hôpital comprend notamment un bloc opératoire une salle de diagnostic un laboratoire des services des urgences de réanimation d ' hémodialyse et d ' imagerie	a capacity of forty-seven this hospital beds includes a block operandi a chamber of the diagnostic laboratories services emergency resuscitation of globalised of imagery	with its capacity of forty-seven beds, this hospital notably boasts an operating theatre, a diagnosis room, a laboratory, intensive care, haemodialysis and imaging departments	with a capacity of forty-seven beds this hospital includes an operating room a diagnosis laboratory a laboratory for emergency intensive care hemodialysis and imaging services.
outre des pavillons de chirurgie générale de pédiatrie et de gynécologie obstétrique	outrage of flag of surgery general paediatrics of gynécologie reproductive health	as well as the general surgery, paediatrics and gynaecology and obstetrics buildings	in addition to pavilions of general surgery of paediatrics and of obstetric gynaecology
cette unité hospitalière profitera à une population estimée à soixante treize mille habitants	it is following hospital benefit an estimated population of soixante-treize thousand inhabitants	this hospital will benefit an estimated population of seventy-three thousand inhabitants	this hospital unit will benefit to an estimated population of seventy-three thousand inhabitants
à l ' occasion de l ' inauguration de l ' hôpital mohamed six sa majesté le roi a également suivi des explications sur le projet de construction	on the occasion of the inauguration of the hospital mohamed vi his majesty the king also followed explanations regarding the construction	on the occasion of the inauguration of Mohammed The Sixth Hospital, his Majesty the King also heard explanations about the planned construction	on the occasion of the inauguration of mohamed six hospital his majesty the king also followed explanations about the construction project
d ' une clinique de maladies psychiatriques à nador	of a clinical it is a disease psychiatric at the request	of a psychiatric hospital in Nador	of a psychiatric clinic in nador
pour laquelle ont été mobilisés des investissements de l ' ordre de trois millions trois cents mille dirhams	why were mobilised investment in the order of three million three hundred thousand dirhams	which will benefit from investments of around three million three hundred thousand dirhams	for which investments of around 3.3 million dirhams were mobilized
un projet qui vise à renforcer la déconcentration et à encourager l ' intégration des traitements des maladies psychiatriques dans les programmes et activités des établissements de santé	a project which seeks to strengthen the deconcentration encourage the integration of the treatment of diseases mentioned in the programmes and activities of institutions centre	a project which aims to reinforce deconcentration and encourage the integration of treatment for psychiatric disorders into health establishments' programmes and activities	a project that aims to strengthen devolution and to encourage the integration of psychiatric disease treatment in the programs and activities of health establishments
réunion du conseil de gouvernement sous la présidence du premier ministre driss jettou marqué	council meeting of government under the presidency of prime minister should be added to the marked	a meeting of the Government council presided by the Prime Minister Driss Jettou marked	meeting of the governing council chaired by prime minister Driss Jettou marked
notamment par deux exposés l ' un fait par le secrétaire d ' état chargé de l ' alphabétisation et de l ' éducation non formelle sur le bilan des programmes de son département	in particular by exposed has indicated the secretary of state for literacy and education non-formal on the results of the programmes of his department	notably by two presentations, one by the Secretary Of State for Literacy and Informal Education on the results of his department's programmes	by two presentations one made by the secretary of state in charge of elimination of illiteracy and non-formal education about the outcome his department's programs

et l ' autre par le ministre des finances et de la privatisation relatif à la préparation de la loi des finances deux mille huit	and the other by the finance minister and privatisation on the preparation of the finance law 2008	and the other by the Finance and Privatisations Minister concerning the two thousand and eight finance law	and the other one by the minister of finance and privatization about the preparation of finance act two thousand eight
nisrine el hamdaoui	these women yes	Nisrine El Hamdaoui	nisrine el hamdaoui
dans son exposé le ministre des finances et de la privatisation a évoqué les résultats de l ' application de la loi de finances à fin mai dernier	in his explanatory statement the finance minister and privatisation has mentioned the results of the application of the law of finances hungry dismiss	In his presentation, the Finance and Privatisations Minister spoke of the results of the application of the finance law at the end of last May	in his presentation the finance and privatization minister described the results of the finance act implementation at end of last may
le téléphone sonne	telephone rings	Le Téléphone Sonne	le téléphone sonne
le téléphone sonne bonsoir	telephone rings good evening	Le Téléphone Sonne, good evening	téléphone sonne good evening
zéro un quarante cinq vingt quatre sept mille	at a zero forty-five twenty-four seven thousand	”oh” one forty-five twenty-four seven thousand	zero one four zero five two zero four seven zero zero zero
vos questions vos réactions sur france inter point com	questions your reactions on france-inter point	Your questions and comments on www dot France Inter dot com	your questions your feedback on france inter dot com
des chambres quatre étoiles à trente euros la nuit la liste est longue de ces prix cassés que l ' on trouve sur internet pour paraît il voyager moins cher	chambers four-star thirty eur night the list is long these give-away prices as too on the internet to seems - travel less expensive	four star rooms at thirty euros a night, there is a long list of these slashed prices you can find on the internet to apparently travel more cheaply	four-star accommodation at thirty euros a night the list of these slashed prices found on the internet to travel cheaper
mais peut on vraiment voyager à petits prix	but subsidiarity really can travel to small price	but is it really possible to travel at low cost	but can we really travel cheaply
toutes ces affaires sont elles vraiment de bonnes affaires ne cachent elles pas de mauvaises surprises et disons le mot parfois des arnaques	all these cases are -elles really good affairs does not hide t it not nasty surprises edition of words sometimes scams	Are all these bargains really bargains ? Aren’t there bad surprises hidden and, let’s say the word, sometimes cons	are all these deals really good deals do not they hide bad surprises and let’s say it sometimes frauds
et plus généralement les lowcosters et voyagistes proposant ces petits prix ont ils changé considérablement notre façon de consommer le tourisme	but more generally the low-cost killer and tour-operators proposing its small price that they have changed considerably our way to consume tourism	and more generally, have low-cost companies and travel agencies offering these low prices considerably changed our way of consuming as tourists	and more generally did lowcosters and tour operators offering these low prices dramatically change the way we consume tourism
c ’ est ce que nous allons voir avec les invités du téléphone sonne qui répondront à toutes vos questions au zéro un quarante cinq vingt quatre sept mille	that is what we shall see with the guests of telephone rings which will answer all your questions at a zero forty-five twenty-four seven thousand	This is what we will see with the guests of Le Téléphone Sonne who will answer your questions on ”oh” one forty-five twenty-four seven thousand	that is what we will see with the guests of le téléphone sonne who will answer all your questions at zero one four zero five two zero four seven zero zero zero
dans ce studio dominique vaucy vous êtes membre du bureau exécutif du syndicat national des agents de voyages	in this session dominique you also you member the executive bureau the national union of travel agencies tour-opérateur	In this studio, Dominique Vaucy, you are a member of the executive bureau of the National Union of Travel Agents	in this studio dominique vaucy you are a member of the executive board of the travel agents’ national union
arnaud de blauwe vous êtes rédacteur en chef adjoint du magazine que choisir et plus particulièrement chargé des questions de tourisme et de transport	arnaud backs white you deputy chief editor the magazine that choose and more particularly responsible for issues of tourism transport	Arnaud de Blauwe you are joint editor of the magazine ”Que Choisir” and more particularly the tourism and transport correspondent	arnaud de blauwe you are the associate editor of the magazine que choisir especially on issues of tourism and transport
et andré longuet des diguères vous êtes directeur départemental de la direction générale de la concurrence la consommation et la répression des fraudes du val de marne	and andré longuet dykes and you are départemental director of the competition directorate-general consumption and repression of fraud and the val-de-marne	and André Longuet des Diguères you are departmental director of the Val de Marne General Office of competition, consumer affairs and fraud	and andré longuet des diguères vous êtes directeur départemental de la direction générale de la concurrence la consommation et la répression des fraudes du val de marne

et tout de suite direction le standard	immediately direction in place in the standard	and straight away, let's go to the switchboard	and immediately the standard
le téléphone sonne	telephone rings	le téléphone sonne	le téléphone sonne
première question nous retrouvons gilles gilles bonsoir	first question we meet gilles gilles goodbye	First question is from Gilles, good evening Gilles	first question we meet again gilles gilles good evening
oui bonsoir	yes evening	Yes good evening	yes good evening
vous vous remercier d ' avoir pris mon appel et puis je voulais témoigner parce que il y a un petit peu plus d ' une semaine j ' ai été sur internet pour pour chercher des un voyage pour partir en tunisie	i am glad that the môme we i find my account that it explains that the men that is what its which we conceive that the euro does include more six	thank you for taking my call and I wanted to tell you something because a little over a week ago I was on the internet to, to, look for a trip to Tunisia	thank you for taking my call and then i wanted to testify because a little more than a week ago i looked for a trip to tunisia on the internet
c ' était c ' était assez difficile enfin il y avait il y avait tout un tas de de possibilités	because it is ten of our neighbours i say that in so many conflicts	it was, it was quite difficult, there was, there was a load of possibilities	it was quite difficult there were a lot of opportunities
et puis finalement à partir d ' éléments que j ' ai que j ' ai pu trouver j ' ai contacté mon agence de voyages traditionnelle	a few moments when it was curiously that i find that the group is happy	and then in the end using information I had found I contacted my traditional travel agency	and then finally from elements that i had been able to find I contacted my traditional travel agent
et j ' ai pu avoir des des conditions à ce que j ' avais vu sur internet et surtout avec quelque chose de complet	to meet the conditions i have therefore i can convictions what the kingdom must be	and I managed to get prices that I'd seen on the internet and above all with full information	and i was able to get the conditions i had seen on the internet and especially with something complete
parce que le le gros inconvénient quand on a des des propositions sur internet on ne sait pas si ce si les prix sont des prix qui comprennent les les taxes	yes it is inappropriate that we find these does its account common we can think that president of true for trains	because the big problem with offers on the internet is that you don't know whether the prices are prices including taxes	because a big disadvantage when you have proposals on the internet you do not know if this if the prices include taxes
on on n ' a pas d ' information sur les hôtels qui sont proposés on ne connaît pas les la société enfin le l- de de voyages qui qui va faire qui va faire le circuit	it has three information are before the you propose in common as i a few children and therefore they what will become not without results	you can't get information on the hotels that are on offer, you don't know the travel company that will be, will be running the trip	you have no information on the hotels that are offered you do not know the company that will organise the tour
alors donc demander son son avis sa réaction à dominique vaucy hein pour le le syndicat national des agents de voyages	so that we will all would ask without its opinion its reaction to dominique is one for the national union of travel agents	so therefore ask the opinion of Dominique Vaucy, right, for the National Union of Travel Agents	then to ask his opinion his reaction to dominique vaucy for the travel agents' national union
le manque de transparence parfois sur les les sites internet est ce que c ' est réel	a what said - it if it is a little about the lack of transparency sometimes on the internet sites because it is real	Is this lack of transparency you get sometimes with internet sites real ?	sometimes the lack of transparency on the websites is that real
c ' est à dire que toute agence de voyage qu ' elle soit passez moi l ' expression en dur ou qu ' elle soit sur le net se doit de déjà de donner un minimum de renseignements à son client	i.e. travel agencies that it is -moi mentions the expression endurants or that it is on the net must already give a minimum of intelligence his client	I mean that any travel agency, whether its, excuse my expression, real-life or on the net, should first of all give a minimum of information to its client	that is to say that any travel agency whether it is on internet or not must already provide minimal information to its client

ce qui est notamment sa sa comment dirais je son adresse sa raison sociale son numéro de licence	which is particularly it is a his address its social reasons the number of licences	which is particularly, how should I say, its address, its company name, its licence number	in particular what its address its corporate name its license number are
de façon à donner les garanties minimum auxquelles toute agence de voyages dûment licenciée est à est est obligée de donner	way will the guarantees given minimum which travel agencies duly dismissed is a objects to give	so as to provide the minimum guarantees which any correctly licensed travel agency is obliged to give	so as to provide the minimum guarantees that any duly licensed travel agency has to give

Résumé

Cette thèse de doctorat aborde les problématiques de l'estimation de confiance pour la traduction automatique, et de la traduction automatique statistique de la parole spontanée à grand vocabulaire. J'y propose une formalisation du problème d'estimation de confiance, et aborde expérimentalement le problème sous le paradigme de la classification et régression multivariée. Je propose une évaluation des performances des différentes méthodes évoquées, présente les résultats obtenus lors d'une campagne d'évaluation internationale et propose une application à la post-édition par des experts de documents traduits automatiquement. J'aborde ensuite le problème de la traduction automatique de la parole. Après avoir passé en revue les spécificités du medium oral et les défis particuliers qu'il soulève, je propose des méthodes originales pour y répondre, utilisant notamment les réseaux de confusion phonétiques, les mesures de confiances et des techniques de segmentation de la parole. Je montre finalement que le prototype proposé rivalise avec des systèmes état de l'art à la conception plus classique.

Mots-clés: traduction automatique statistique, mesures de confiance, traduction de la parole, post-édition, segmentation, réseau de confusion, phonèmes

Abstract

In this thesis I shall deal with the issues of confidence estimation for machine translation and statistical machine translation of large vocabulary spontaneous speech translation. I shall first formalize the problem of confidence estimation. I present experiments under the paradigm of multivariate classification and regression. I review the performances yielded by different techniques, present the results obtained during the WMT2012 internation evaluation campaign and give the details of an application to post edition of automatically translated documents. I then deal with the issue of speech translation. After going into the details of what makes it a very specific and particularly challenging problem, I present original methods to partially solve it, by using phonetic confusion networs, confidence estimation techniques and speech segmentation. I show that the prototype I developped yields performances comparable to state-of-the-art of more standard design.

Keywords: statistical machine translation, confidence estimation, speech translation, post edition, segmentation, confusion network, phoneme