à exploiter le potentiel de l'idée de TransTalk, et nous examinons actuellement diverses façons de l'améliorer.

Références

[1] Peter F. Brown, Stanley F. Chen, Stephen A. Della Pietra, Vincent Della Pietra Andrew S. Kehler, et Robert L. Mercer. Automatic speech recognition in machine aided translation, 1992. (à paraître).

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, et Paul Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, juin 1990.

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent Della J. Pietra, et Robert L. Mercer. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–312, juin 1993.

[4] M. Dymetman, G. Foster, et P. Isabelle. Towards an Automatic Dictation System for Translators (TransTalk). Rapport technique, Ministère des Communications Canada, Centre Canadien de Recherches sur l'Informatisation du Travail, Laval, Québec, 1992.

[5] P. Isabelle, M. Dymetman, G. Foster, J.M. Jutras, E. Macklovitch, F. Perrault, X. Ren, et M. Simard. Translation analysis and translation automation. Dans *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, 1993.

[6] Martin Kay, Jean Mark Gawron, et Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog. Lecture Note No. 33.* Center for the Study of Language and Information, Stanford, 1994.

[7] A. Kurematsu. Automatic telephone interpretation: A basic study. ATR Technical Report TR-I-0001, ATR Interpreting Telephony Research Laboratories, Japan, mai 1987.

[8] R. Lacouture et R. De Mori. Lexical tree compression. Dans *Proceedings, Eurospeech Conference*, p. 581–584, Gênes, Italie, september 1991.

[9] Michel Simard, George F. Foster, et Pierre Isabelle. Using cognates to align sentences in bilingual corpora. Dans *TMI-92*, Montreal, Canada, 1992. Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.

[10] F.W.M. Stentiford et M.G. Steer. Machine translation of speech. Dans C. Wheddon and R. Linggard, éds, *Speech and Language Processing*, pages 183–196. Chapman and Hall, Londres, 1990.

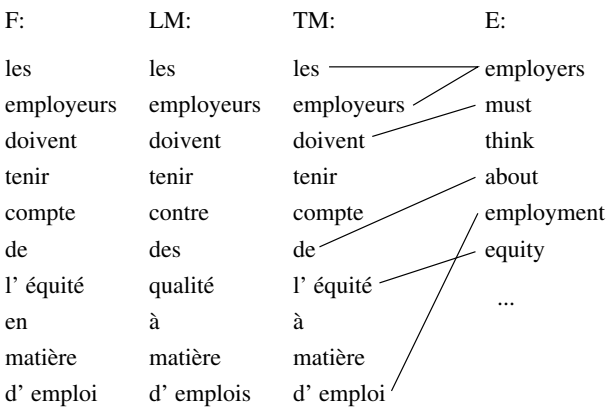

Figure 5: Comparaison des résultats obtenus par le modèle de langue (ML) et par le modèle de traduction (MT) pour une paire de phrases (F, E) du corpus d'essai. (Cette paire a été tronquée pour des raisons d'espace.) Les lignes indiquent les principaux éléments de l'alignement le plus probable entre la phrase de sortie et E. La présence du mot *equity* dans la phrase source anglaise a permis au modèle de traduction de choisir correctement *équité* au lieu de *qualité*.

Modèle	Mots corrects (/918)		Perplexité
	Locuteur 1	Locuteur 2	
langue pur	686 (74,7%)	677 (73,8%)	385
interpolé (.85)	735 (80,1%)	734 (80,0%)	180
maximum	735 (80,1%)	732 (79,7%)	–
essai-e (.30)	742 (80,8%)	734 (80,0%)	–

Figure 6: Résumé des résultats obtenus par TransTalk. La première ligne contient des statistiques sur les résultats obtenus par le modèle de langue pur; les autres lignes contiennent des statistiques sur les résultats obtenus par le modèle de traduction avec chacune des trois méthodes de lissage décrites à la section 3.3 (où .85 était le facteur de pondération optimal pour les paramètres bi-lexicaux, et .30 était le seuil de confiance optimal).

normalisées, elles ne peuvent être comparées que sur le plan de la performance de reconnaissance et non au moyen de la mesure de perplexité (voir la section 5).

4 La recherche

Le but de la composante recherche est de trouver une approximation de la phrase $\hat{\mathbf{f}}$ qui maximise le produit des scores acoustique et traductionnel $p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f} \mid \mathbf{e})$. Notre algorithme de recherche procède en deux étapes, qui sont toutes deux sous-optimales.

La première étape consiste à utiliser le modèle acoustique pour réduire la liste d'hypothèses lexicales pour chaque unité acoustique de 20 000 à un nombre n (actuellement 20). Comme cet élagage s'effectue sans référence au modèle de traduction, il n'y a aucune garantie que $\hat{\mathbf{f}}$ se trouve parmi les $n^{|\mathbf{f}|}$ phrases candidates retenues.

La seconde étape est une recherche Viterbi portant sur le reste des phrases candidates et utilisant le modèle de traduction. Cela nous permet de trouver la paire $(\tilde{\mathbf{f}}, \tilde{\mathbf{c}})$ qui maximise le produit $p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f}, \mathbf{c} \mid \mathbf{e})$ dans un temps qui est proportionnel à $nC^3|\mathbf{f}||\mathbf{e}|$, où C est le nombre de classes de mots dans le modèle de traduction (actuellement 15). Étant donné la nature grossière de nos classes de mots, nous estimons que $\tilde{\mathbf{f}}$ est une approximation raisonnable de $\hat{\mathbf{f}}$.

5 Résultats

Nous avons testé TransTalk sur un petit corpus de 50 paires de phrases françaises/anglaises tirées du Journal des débats qui n'avaient pas été utilisées comme données d'entraînement. Toutes les phrases françaises comportaient de 15 à 20 unités (en comptant la ponctuation) et aucune ne contenait de mots ne faisant pas partie de notre vocabulaire de 20 000 mots. Elles ont été dictées par mots isolés par deux locuteurs différents.

La Figure 5 illustre les résultats obtenus pour une seule paire de phrases. Les statistiques générales sont présentées à la Figure 6. Le modèle de traduction a permis une réduction moyenne du taux d'erreur de l'ordre de 24 p. cent par rapport au modèle de langue pur. Pour les erreurs visant des mots de "contenu" (par exemple *action* au lieu de *section*), cette réduction était de 42 p. cent. L'utilisation du modèle de traduction a permis de réduire de plus de la moitié la perplexité du corpus d'essai.

6 Conclusions

Nos résultats initiaux démontrent qu'il est possible d'utiliser économiquement et efficacement l'information traductionnelle pour la reconnaissance de la parole. Nous croyons que l'approche simple décrite dans cet article commence à peine

f	$p(f \mid c, e)$
gouvernement	0.7363
m.	0.0227
monsieur	0.0134
président	0.0109
canada	0.0081
façon	0.0033
mesure	0.0024
part	0.0023
ministre	0.0023
décision	0.0022

Figure 4: Un échantillon des paramètres bi-lexicaux de TransTalk. Ce sont les dix mots français les plus probables, étant donné la classe NOM et le mot anglais *government*.

la qualité de nos données d'entraînement, nous avons éliminé les alignements comportant plus d'une phrase en langue source ou cible, ainsi que ceux qui contenaient plus de 40 unités dans l'une ou l'autre langue; cela a réduit d'environ 20 p. cent la taille du corpus d'entraînement, à environ 8 millions d'unités dans chaque langue. Ensuite, nous avons utilisé le modèle de langue pur pour étiqueter chaque mot de la partie française du corpus réduit avec sa classe la plus probable. Enfin, nous avons utilisé l'algorithme "EM" pour estimer les paramètres $p(f, c \mid e)$ à partir du corpus étiqueté aligné. Ceux-ci ont été transformés en paramètres bi-lexicaux, comme suit:

$$p(f \mid c, e) = \frac{p(f, c \mid e)}{\sum_f p(f, c \mid e)} \tag{6}$$

La Figure 4 présente un échantillon des résultats.

Étant donné que de nombreuses combinaisons bi-lexicales valides n'apparaissent pas dans notre corpus d'entraînement, il a fallu lisser les paramètres bi-lexicaux. Au lieu de modifier directement la distribution empirique $p(f \mid c, e)$, nous avons choisi de lisser dynamiquement la quantité plus robuste $p(f \mid c, \mathbf{e})$ intervenant dans les calculs basés sur l'équation 4. Nous avons expérimenté trois méthodes simples de combiner cette quantité avec les paramètres lexicaux moins précis mais plus fiables $p(f \mid c)$ du modèle de langue pur: l'interpolation linéaire, l'utilisation du maximum de $p(f \mid c, \mathbf{e})$ et $p(f \mid c)$, et l'utilisation de $p(f \mid c)$ ssi $\max_e p(f \mid c, e)/p(f \mid c, \mathbf{e})$ ne dépassait pas un certain seuil. La justification de la deuxième méthode est que nous nous attendons à ce que les probabilités plus fortes soient estimées de façon plus fiable, en moyenne, que les probabilités plus faibles. La troisième méthode doit servir à rejeter l'information traductionnelle quand aucun mot anglais n'est fortement associé au mot français considéré. Étant donné que les deux dernières méthodes produisent des distributions non

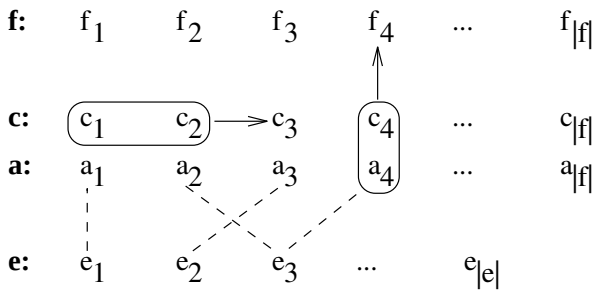

Figure 3: La structure de la source markovienne sous-jacente au modèle de traduction. D'abord, le processus produit **c** en sélectionnant chaque classe en fonction des deux précédentes, la probabilité étant donnée par le paramètre contextuel approprié. Ensuite, il établit **a** en sélectionnant aléatoirement une position dans **e** pour chaque position dans **c**. Enfin, il génère **f** en sélectionnant chaque mot en fonction de sa classe et son correspondant anglais, la probabilité étant donnée par le paramètre bi-lexical approprié.

sommes sur les mots anglais. Le résultat est l'équation suivante:

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{c}} \prod_{i=1}^{|\mathbf{f}|} p(c_i \mid c_{i-2}, c_{i-1}) p(f_i \mid c_i, \mathbf{e}) \tag{4}$$

où $p(f_i \mid c_i, \mathbf{e}) = \sum_{j=0}^{|\mathbf{e}|} p(f_i \mid c_i, e_j)/(|\mathbf{e}| + 1)$. À partir de cela, il devrait être évident que notre modèle de traduction n'est rien de plus qu'un modèle tri-classe standard dans lequel les paramètres lexicaux $p(f \mid c)$ ont été remplacés par $p(f \mid c, \mathbf{e})$.

3.3 Estimation des paramètres

Les deux familles de paramètres du modèle de traduction ont été estimés séparément. Les paramètres contextuels ont été estimés dans le cadre d'un modèle de langue tri-classe pur pour le français, qui a été entraîné sur la moitié française de notre corpus bilingue au moyen de l'algorithme "EM", en utilisant un dictionnaire pour identifier les classes valides pour chaque mot.

Les paramètres bi-lexicaux ont été estimés à partir d'un modèle de traduction simplifié, dans lequel l'information contextuelle a été présumée explicite:

$$p(\mathbf{f}, \mathbf{c} \mid \mathbf{e}) = \frac{1}{A_{\mathbf{f},\mathbf{e}}} \sum_{\mathbf{a}} \prod_{i=1}^{|\mathbf{f}|} p(f_i, c_i | e_{a_i}) \tag{5}$$

Pour entraîner ce modèle, nous avons d'abord aligné le corpus d'entraînement au niveau de la phrase, en utilisant la méthode décrite en [9]. Pour améliorer

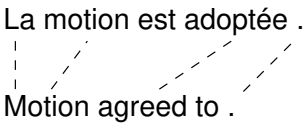

Figure 2: Un exemple d'alignement, une des 5^5 possibilités pour cette paire de phrases.

à aucun. On peut représenter un alignement comme un vecteur $\mathbf{a}$ de longueur $|\mathbf{f}|$ qui contient, pour chaque mot français, la position dans $\mathbf{e}$ du mot anglais auquel il est relié, ou zéro si le mot français n'est connecté à aucun mot. Nous supposons que tous les $A_{\mathbf{f},\mathbf{e}}$ alignements possibles sont équiprobables, avec une probabilité de $p(\mathbf{a} \mid \mathbf{c}, \mathbf{e}) = 1/A_{\mathbf{f},\mathbf{e}}$, de sorte que nous avons:[3].

$$p(\mathbf{f} \mid \mathbf{c}, \mathbf{e}) = \sum_{\mathbf{a}} \frac{1}{A_{\mathbf{f},\mathbf{e}}} \cdot p(\mathbf{f} \mid \mathbf{a}, \mathbf{c}, \mathbf{e}) \tag{2}$$

Il s'agit d'une approximation grossière qui contredit notre intuition que certains alignements—comme ceux dans lesquels tous les mots français sont connectés au même mot anglais, ou ceux dans lesquels les verbes français ne sont connectés qu'à des prépositions anglaises—seront beaucoup moins probables que d'autres. Son but est la simplification et nous la justifions en raison du fait qu'un modèle raisonnable de $p(\mathbf{f} \mid \mathbf{a}, \mathbf{c}, \mathbf{e})$ minimisera, de toute façon, la contribution (de la plupart) des mauvais alignements.

L'étape finale consiste à supposer que les mots de $\mathbf{f}$ sont conditionnellement indépendants étant donné $\mathbf{a}$, $\mathbf{c}$ et $\mathbf{e}$, et en outre, que chaque mot dépend seulement de sa classe et du mot anglais auquel il est connecté dans l'alignement:

$$p(\mathbf{f} \mid \mathbf{a}, \mathbf{c}, \mathbf{e}) = \prod_{i=1}^{|\mathbf{f}|} p(f_i \mid c_i, e_{a_i}) \tag{3}$$

Notre modèle complet est une source markovienne (voir la Figure 3) qui dépend de deux ensembles de paramètres: les paramètres *contextuels* de la forme $p(c_i \mid c_{i-2}, c_{i-1})$, qui prédisent une classe à partir des deux classes qui la précèdent, et les paramètres *bi-lexicaux* de la forme $p(f \mid c, e)$, qui prédisent un mot français à partir de sa classe et de son correspondant anglais.

Il est possible de réaménager la combinaison simple des équations 1, 2 et 3 d'une façon qui permette des calculs plus efficients. L'observation clé est que la somme sur l'ensemble des alignements peut être réorganisée en un produit de

[3]Où $A_{\mathbf{f},\mathbf{e}} = (|\mathbf{e}|+1)^{|\mathbf{f}|}$

3.1 Les approches de modélisation

Il existe au moins deux approches distinctes permettant de modéliser cette distribution. Dans [3], Brown *et al.* l'assimilent au produit $p(\mathbf{f}) \cdot p(\mathbf{e} \mid \mathbf{f})$, auquel elle est proportionnelle dans le cas de la maximisation par rapport à $\mathbf{f}$. Le principal avantage de cette méthode est qu'elle permet une "division du travail" entre $p(\mathbf{f})$, qui est responsable de la structure correcte de $\mathbf{f}$, et $p(\mathbf{e} \mid \mathbf{f})$, qui est chargé de faire en sorte que $\mathbf{e}$ et $\mathbf{f}$ soient des traductions acceptables, sans avoir à se préoccuper outre mesure de la structure interne de l'un ou de l'autre. Bien qu'elle soit puissante, cette technique présente un inconvénient qui la rend inappropriée à nos fins: elle se prête difficilement à des recherches efficientes sur de vastes ensembles de phrases françaises candidates.

À cause de cela, nous avons choisi de modéliser $p(\mathbf{f} \mid \mathbf{e})$ plus directement, comme une famille de modèles de langue markoviens paramétrés, $p_{\lambda(e)}(\mathbf{f})$, où chaque $\mathbf{e}$ spécifie un vecteur de paramètres λ, pas nécessairement de façon unique. Cette approche pose le défi d'incorporer l'information provenant de $\mathbf{e}$ sans interférer avec la connaissance de la structure du français que possède le modèle de langue—en particulier dans le cas des modèles de langue qui sont précis au départ. Dans les travaux décrits ici, nous avons en grande partie évité cette difficulté en utilisant un modèle de langue passablement faible, notre principal but étant d'examiner dans quelle mesure il est possible d'améliorer la performance d'un tel modèle sans augmenter substantiellement son faible coût calculatoire.

3.2 Dérivation

Le modèle de traduction est basé sur un modèle de langue tri-classe standard dépendant de $\mathbf{e}$. La première hypothèse clé que nous posons est que la séquence de classes de mots $\mathbf{c}$ associé à $\mathbf{f}$ est indépendante de $\mathbf{e}$, ce qui nous permet d'écrire:

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{c}} p(\mathbf{c}) \cdot p(\mathbf{f} \mid \mathbf{c}, \mathbf{e}) \tag{1}$$

Cette approximation est motivée par l'intuition que $\mathbf{e}$ sera surtout informatif à propos des mots particuliers que pourra contenir $\mathbf{f}$, mais seulement faiblement informatif sur la structure syntaxique brute du genre de celle que $\mathbf{c}$ saisit. Parce que le modèle est surtout valide quand $\mathbf{c}$ est constitué de classes générales,[2] nous utilisons un ensemble minimal de 15 classes qui correspondent aux principales catégories grammaticales (nom, verbe, etc.).

Pour incorporer l'information traductionnelle, nous supposons, suivant Brown et al. [3], que $\mathbf{f}$ et $\mathbf{e}$ sont reliés par un *alignement* (voir la Figure 2), dans lequel chaque mot français est soit connecté à un seul mot anglais de $\mathbf{e}$, soit connecté

[2]Cette hypothèse devient de plus en plus intenable pour les classifications plus fines; à la limite où les classes sont identiques aux mots, le modèle se réduit à un pur tri-gramme sans aucune composante traductionelle.

anglais indique fortement que le choix correct devrait être *chevaux*, et c'est cette connaissance de la traduction probable que TransTalk cherche à exploiter.

Conceptuellement, la principale différence entre un système traditionnel de reconnaissance de la parole à "voie bruitée" pour le français et TransTalk est que, au lieu de maximiser dans $\mathbf{f}$ le produit $p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f})$ d'un "modèle acoustique" et d'un "modèle de langue" du français (où $\mathbf{s}$ correspond au signal acoustique et $\mathbf{f}$ à la phrase française), nous maximisons le produit $p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f} \mid \mathbf{e})$ d'un modèle acoustique et d'un "modèle de traduction" de l'anglais au français (où $\mathbf{e}$ correspond à la phrase anglaise à traduire). Voir la Figure 1.

Nous avons réalisé une version prototype de TransTalk qui fonctionne en mode de dictée de mots isolés sur un vocabulaire de 20 000 mots français. Cette version est spécialisée dans le domaine des débats parlementaires canadiens, qui sont transcrits sous forme bilingue dans le Journal des débats de la Chambre des communes (*Hansard*). Nous avons utilisé les transcriptions de deux années du Journal des débats (soit environ 10 millions de mots français et 10 millions de mots anglais) comme corpus d'entraînement pour le modèle de traduction.

2 Le modèle acoustique

Nous utilisons un modèle markovien caché (MMC) basé sur des modèles de phones indépendants du contexte pour décrire $p(\mathbf{s} \mid \mathbf{f})$. Le vocabulaire de TransTalk est représenté au moyen d'un ensemble de 47 phonèmes comprenant 20 voyelles et 27 consonnes. Les prononciations de base ont été obtenues à l'aide d'un ensemble de règles de graphème-à-phonème qui tiennent compte des particularités phonétiques du français parlé au Québec, notamment l'assibilation et le relâchement des voyelles.

La reconnaissance s'effectue par la recherche des n-meilleurs candidats dans un graphe phonétique comprimé représentant l'intégralité du vocabulaire de 20 000 mots [8]. Dans ce graphe, il n'y a pas deux chemins qui produisent la même séquence de phones et chaque chemin correspond à une représentation phonétique valide dans le dictionnaire. Un chemin donné correspondra par conséquent à toutes les entrées du lexique partageant la même phonétique. La recherche produit une liste des 20 mots les plus probables acoustiquement pour chaque unité acoustique (isolée).

3 Le modèle de traduction

Le modèle de traduction vise à décrire $p(\mathbf{f} \mid \mathbf{e})$, soit la probabilité que le traducteur produise la phrase française $\mathbf{f}$ comme traduction de la phrase anglaise $\mathbf{e}$.

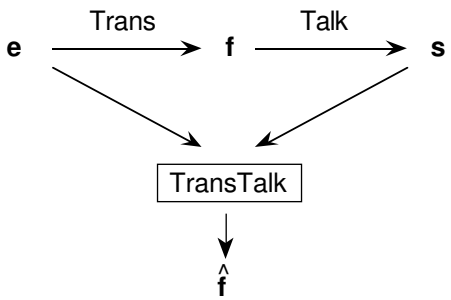

Figure 1: Modèle sous-jacent à TransTalk. Partant d'une phrase anglaise $\mathbf{e}$, le traducteur formule mentalement sa traduction française $\mathbf{f}$, puis produit sa réalisation acoustique $\mathbf{s}$. Le but du système est de trouver $\hat{\mathbf{f}} = \text{argmax}_f\ p(\mathbf{f} \mid \mathbf{e}, \mathbf{s})$, ou, de façon équivalente, suivant la formule de Bayes, $\hat{\mathbf{f}} = \text{argmax}_f\ p(\mathbf{s} \mid \mathbf{e}, \mathbf{f}) \cdot p(\mathbf{f} \mid \mathbf{e})$. En négligeant l'influence de $\mathbf{e}$ sur $\mathbf{s}$ une fois que $\mathbf{f}$ est connu, nous pouvons poser que $\hat{\mathbf{f}} = \text{argmax}_f\ p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f} \mid \mathbf{e})$.

opérations de reconnaissance de la parole, de traduction automatique et de synthèse de la parole. Cependant, à l'heure actuelle, la reconnaissance de la parole et la traduction automatique ne peuvent être réalisées de façon fiable que moyennant l'imposition de strictes contraintes lexicales, syntaxiques et sémantiques, ce qui oblige les concepteurs de systèmes de traduction directe de la parole à s'en tenir à des domaines d'application pour lesquels il est possible de définir naturellement des sous-langages restreints.

Dans le projet TransTalk, nous tentons d'intégrer la traduction automatique et la reconnaissance de la parole de manière à faire une utilisation maximale de leurs forces complémentaires, plutôt que d'aggraver leurs faiblesses respectives. Nous ne cherchons pas à remplacer le traducteur humain (ce qui est en général une vaine entreprise), mais plutôt, de façon plus réaliste, à fournir un *outil de dictée* au traducteur. Notre but est d'utiliser la traduction automatique pour faire des prédictions probabilistes quant aux énonciations susceptibles d'être produites librement par le traducteur et de nous servir de ces prédictions pour réduire la difficulté de la reconnaissance de la parole jusqu'au point où la reconnaissance complète de ces énonciations pourra être réalisée.[1]

Par exemple, dans le cas de la traduction de l'anglais au français, supposons que le traducteur décide de rendre la phrase "what splendid horses you have" par "tes chevaux sont vraiment magnifiques". Un système de reconnaissance de la parole n'ayant pas accès au texte source pourrait avoir de la difficulté à faire la distinction entre *chevaux* et *cheveux*, qui est acoustiquement semblable et contextuellement plus probable. Or, la présence du mot *horses* dans le texte source

[1] Cette idée a été avancée indépendamment par nous [4] ainsi que par des chercheurs du Thomas J. Watson Research Center d'IBM [1].

Vers un système de dictée automatique pour traducteurs: le projet TransTalk

Marc Dymetman[†*], Julie Brousseau[‡], George Foster[†],
Pierre Isabelle[†], Yves Normandin[‡], Pierre Plamondon[†]

[†] Centre d'innovation en technologies de l'information (CITI)
1575, boul. Chomedey, Laval H7V 2X2 (Quebec) Canada

[‡] Centre de Recherche Informatique de Montreal (CRIM)
1801, av. McGill College, Montreal H3A 2NA (Quebec) Canada

Le 16 septembre 1994

Résumé

Les traducteurs professionnels dictent souvent leurs traductions, qu'ils font ensuite transcrire. Le projet TransTalk vise à automatiser cette deuxième partie du processus. Comme système de dictée, son originalité réside dans le fait que le signal acoustique produit par le traducteur et le texte source en voie de traduction sont tous deux accessibles au système. Les traductions probables du texte source peuvent être prédites et ces prédictions peuvent être mises à contribution pour aider le système de reconnaissance de la parole à faire ses choix lexicaux. Nous présentons ici les résultats du premier prototype, qui révèlent une nette amélioration de la performance de la reconnaissance de la parole quand les prédictions de traduction sont prises en compte.

1 Introduction

L'intégration de la traduction automatique et des technologies vocales fait actuellement l'objet d'importants projets de recherche dans plusieurs pays [6, 7, 10]. Ces efforts visent habituellement le développement de systèmes de traduction directe de la parole (de parole à parole) réalisant séquentiellement les

*Address actuelle: Centre de recherche Rank Xerox, 6 chemin de Maupertuis, 38240 Meylan, France.