Le moteur de prédiction de mots de la Plateforme de Communication Alternative

Philippe Blache, Stéphane Rauzy

Laboratoire Parole et Langage
CNRS & Aix-Marseille Universités
pb@lpl.univ-aix.fr, stephane.rauzy@lpl.univ-aix.fr

RÉSUMÉ. *Nous présentons le modèle de langage sous-jacent au moteur de prédiction de mots de la Plateforme de Communication Alternative (PCA), un logiciel d'aide à la communication pour personnes handicapées. Le moteur s'appuie sur un lexique général du français très couvrant qui donne pour chaque entrée la fréquence d'usage du mot et ses traits morphosyntaxiques associés. Il intègre un modèle utilisateur (lexique personnel mémorisant les mots inconnus saisis par l'utilisateur, calcul des fréquences d'usage propres, stockage des phrases produites) et un module de prédiction morphosyntaxique qui pondère les fréquences des mots prédits en fonction du contexte syntaxique de la phrase en cours de composition. L'évaluation du modèle de langage donne des résultats satisfaisants, le taux d'économie de saisies est d'environ* 55% *pour une liste de* 9 *propositions. La contribution dominante provient de la prédiction brute basée sur les fréquences d'usage du lexique général.*

ABSTRACT. *We present the model of language implemented in the words prediction engine of the "Plateforme de Communication Aternative" (PCA), an assistive communication software designed for impaired persons. The model relies on a large coverage lexicon for french langage which provides for each entry its word frequency and its set of associated morphosyntactic categories. The engine includes a user model (personal lexicon for unknown words typed by the user, computation of personal word frequencies, storage of the generated sentences) and a morphosyntactic prediction system which weight the word frequencies of predicted words according to the syntactical context of the sentence. The evaluation of the model of language is quite encouraging : a keystrokes saving rate of* 55% *for* 9 *propositions. The major contribution comes from the raw prediction obtained by using the words frequencies of the general lexicon.*

MOTS-CLÉS : *Communication assistée pour personnes handicapées, système de prédiction de mots, modèle utilisateur, prédiction morphosyntaxique.*

KEYWORDS: *Assistive communication, words prediction, user model, morphosyntactic prediction.*

1. Introduction

La communication alternative désigne un ensemble d'outils d'aide à la communication pour des personnes handicapées atteintes dans leur motricité et leur capacité de production de parole. Il s'agit par exemple de patients atteints de pathologies neurodégénératives totalement paralysantes ou encore de personnes victimes d'accidents vasculaires cérébraux. Ces patients ne gardent le contrôle que de quelques muscles (comme la paupière) et ne peuvent plus parler. Pour d'autres pathologies, certains types d'aphasies par exemple, les capacités linguistiques et cognitives sont affectées et des stratégies alternatives comme la communication non verbale à base d'icônes doivent être utilisées. L'objectif de ce type de système est de permettre à l'utilisateur d'améliorer voire de rétablir la possibilité de communication avec son entourage en offrant la possibilité de composer des messages, de piloter un système de synthèse de parole ou encore de désigner des objets ou des actions. Il s'agit donc de prendre en compte les besoins effectifs des utilisateurs dans une situation réelle de communication, et d'intégrer des modalités multiples d'interaction pour le support de la communication et le contrôle de l'environnement (voir par exemple (Vaillant, 1997) et (Brangier et Gronier, 2000)).

L'aide à la communication de personnes handicapées est un problème majeur, mais qui peut aujourd'hui bénéficier de la maturité technologique des travaux menés dans le domaine de la linguistique, la linguistique-informatique, l'ergonomie et la psychologie cognitive. Les réponses apportées à ce jour ne sont pas totalement satisfaisantes, notamment pour ce qui concerne les modalités d'interaction entre l'utilisateur handicapé et son environnement humain ou électronique (voir par exemple (Maurel *et al.*, 2000)).

Quelques systèmes d'aide à la communication en langue française sont aujourd'hui proposés sur le marché. Citons par exemple pour le verbal : WiViK, clavier virtuel avec prédiction de mots et principe de défilement en option, permettant également le contrôle du système d'exploitation ; Eurovocs Suite, claviers virtuels et prédiction de mots basée sur un dictionnaire contenant 35 000 formes. Et pour la communication non verbale : Clicker 4, outil d'aide à la communication à base d'icônes ; Mind Express, un système de communication non verbale à base d'icônes qui intègre une reformulation rudimentaire ; Axelia, destiné aux jeunes enfants IMC (Infirme Moteur Cérébraux) et aphasiques, accessible via une interface graphique évoluée, qui base sa reformulation sur l'application du modèle de la grammaire applicative et cognitive (voir (Abraham, 2000) et (Abraham, 2006)).

Il existe enfin un certain nombre d'applications expérimentales développées dans le milieu académique : par exemple, Vitipi (Boissière et Dours, 2000), HandiAS (Le Pévédic, 1997) ou Kombe (Pasero et Sabatier, 1995), mais qui ne sont pas véritablement distribués au grand public. Signalons toutefois le système Sibylle (Schadle, 2003) (Wandmacher *et al.*, 2007), un clavier orthographique optimisé pour la modalité défilement et muni d'un moteur de prédiction de mots très efficace, qui sera prochainement disponible en libre accès.

La Plateforme de Communication Alternative (PCA), a été developpée au Laboratoire Parole et Langage par une équipe pluridisciplinaire constituée de chercheurs en informatique et en linguistique, de psychologues cognitivistes et d'ergonomes, d'électroniciens, de médecins et rééducateurs ainsi que de centres d'accueil et d'associations de handicapés. Distribuée depuis début 2004, la PCA est caractérisée par son homogénéité, sa généricité et son évolutivité, trois points clés pour tout système de communication assistée (Copestake, 1997). Le logiciel PCA permet la composition assistée de messages selon deux modes principaux : le mode verbal et le mode non verbal (Blache et Rauzy, 2003) (Bellengier *et al.*, 2004). Ces deux types de composition sont accessibles par le clavier, la souris, ou une procédure de défilement controlée par un capteur binaire, selon le degré de motricité des utilisateurs (Blache et Rauzy, 2004) (une version de démonstration de la PCA est disponible sur le site www.aegys.com).

La composition en mode non verbal s'effectue à l'aide d'un clavier d'icônes. La base d'icônes générale partagée par tous les utilisateurs regroupe environ 750 pictogrammes qui ont été dessinés à partir d'une chartre graphique et sémantique élaborée par le Laboratoire Parole et Langage. Elle couvre des besoins communicationnels variés. La base comprend environ 200 verbes (les verbes les plus courants et des verbes spécialisés utilisés par exemple dans le domaine médical), environ 200 noms communs (désignant des objets, des lieux, des personnes, etc.), une cinquantaine d'adjectifs, les pronoms, les adverbes, les déterminants, les prépositions les plus courantes, et les nombres. La base comprend de plus les icônes représentant les lettres et les phonèmes qui permettent de créer des claviers alphabétiques ou phonétiques. Chaque utilisateur pourra ensuite créer et ajouter, via une interface facile d'accès, ses propres icônes (à partir de photos numériques par exemple). Le système incorpore un module de reformulation iconique qui génère, à partir de la séquence d'icônes composée, une phrase en langage naturel syntaxiquement et sémantiquement correcte (Blache et Rauzy, 2007).

La composition en mode verbal s'effectue à l'aide d'un clavier virtuel orthographique statique complété par un clavier dynamique de proposition de mots. Le moteur de prédiction implanté dans PCA utilise un lexique très couvrant du français et propose une prédiction contextuelle incluant l'information sur les traits morphosyntaxiques associées aux entrées du lexique ainsi qu'un modèle utilisateur qui prend en compte les habitudes langagières de l'utilisateur par apprentissage.

Nous décrivons dans cet article le modèle de langage sous-jacent au moteur de prédiction de mots de la PCA. Dans la section 2, nous présentons le principe de fonctionnement de la PCA Orthographique. Le modèle de langage du moteur de prédiction de mots est décrit section 3 et son évaluation est conduite section 4. Nous résumons finalement section 5 les principaux résultats obtenus, et proposons des pistes en vue d'améliorer l'efficacité du système.

2. Présentation de la PCA Orthographique

La PCA dans sa version orthographique est principalement destinée aux personnes possédant une bonne maîtrise de l'écrit, mais qui souffrent de sévères troubles moteurs les empêchant de la mettre en œuvre naturellement. Le problème est tout d'abord de proposer une solution adaptée à chaque utilisateur handicapé pour piloter le logiciel d'aide à la communication. L'évaluation du niveau de motricité de l'utilisateur est donc fondamentale et la recherche de capteurs adaptés doit tenir compte des facteurs tels que le taux d'échecs, la charge cognitive, le confort, la rapidité et la fatigabilité de l'utilisateur. D'autre part, le problème réside dans la lenteur de saisie des messages écrits (en moyenne 1 à 5 mots par minute sur un clavier non dédié (Wandmacher et Antoine, 2006)). Le logiciel d'aide à la communication doit donc proposer à l'utilisateur des procédés permettant d'accélérer la vitesse de composition des messages. Dans cette section, nous décrivons les modalités d'accessibilité permettant de piloter l'application, le clavier de lettres optimisé pour accélérer la saisie des messages et la stratégie à suivre pour composer des messages avec la PCA Orthographique.

2.1. *Les modalités d'accessibilité*

Trois modalités d'accessibilité sont proposées pour contrôler le logiciel PCA, selon le degré de motricité de chaque utilisateur. Quelle que soit la modalité d'accessibilité employée pour composer un message, les grands principes d'utilisation de la PCA ainsi que l'interface graphique restent inchangés. Ainsi, un utilisateur amené à changer de modalité d'accessibilité au cours de l'évolution de sa pathologie n'aura pas à réapprendre le fonctionnement du logiciel (c'est le cas par exemple pour les utilisateurs atteints d'une pathologie neurodégénérative ou pour les utilisateurs en phase de remédiation).

La modalité *Clavier* : l'utilisateur interagit avec l'ordinateur en utilisant le clavier physique de l'ordinateur, c'est-à-dire en appuyant sur les lettres et les touches de fonction du clavier. Cette modalité requiert de l'utilisateur la capacité de déplacer le bras au-dessus du clavier (le mouvement peut être lent), de sélectionner la touche (un guide-doigts peut être utilisé pour pallier des tremblements non contrôlés) et d'exercer une pression sur la touche.

La modalité *Souris* : l'utilisateur a la possibilité de contrôler le déplacement du curseur de la souris à l'écran, en utilisant la souris standard de l'ordinateur, un trackball ou un joystick. Cette modalité peut être proposée à un utilisateur possédant encore la motricité du poignet (même de faible amplitude). Des solutions alternatives, consistant par exemple à associer les déplacements du pointeur de la souris aux mouvements de la tête de l'utilisateur filmé par une webcam, peuvent aussi être retenues comme procédé de pointage. La touche virtuelle pointée est ensuite sélectionnée par une interaction binaire déclenchée par un capteur ou un contacteur. Dans ce cas, l'utilisateur a la capacité de contrôler un mouvement intentionnel (pression du doigt, émission d'un souffle, clignement de la paupière, etc.) détecté par le capteur. Un mode pour sé-

Figure 1. *L'interface de la Plateforme de Communication Alternative en mode orthographique, modalité d'accessibilité défilement*

lectionner automatiquement la touche virtuelle pointée au bout d'un certain délai est aussi proposé.

La modalité *Défilement* : l'utilisateur a la capacité de contrôler un mouvement intentionnel qui est transformé en interaction binaire. Un curseur défile automatiquement sur les touches des claviers virtuels. Lorsque le curseur passe sur la touche désirée, l'utilisateur sélectionne la touche virtuelle en activant le capteur.

2.2. *Le clavier de saisie de lettres*

Le clavier virtuel de lettres est présenté figure 1. Il s'agit d'un clavier statique dont la disposition des caractères est optimisée pour la modalité d'accessibilité défilement. Dans ce cas, la sélection des touches se fait selon un défilement ligne-colonne. Deux interactions avec le capteur sont ainsi nécessaires pour saisir un caractère, une interaction pour sélectionner la ligne contenant la touche et une interaction pour sélectionner la touche sur cette ligne. Le temps d'accès est proportionnel à la somme de la position de la ligne et de la colonne de la touche dans le clavier (voir par exemple

(Copestake, 1997)). Les lettres sont donc disposées en diagonale suivant leurs fréquences d'utilisation (pour le français), des plus fréquentes en haut à gauche aux moins fréquentes en bas à droite (i.e. deux temps de défilement sont nécessaires pour accéder à la lettre *a*, trois pour *e* et *s*, quatre pour *i*, *r* et *n*... , neuf pour -, *k* et *z*). Nous avons choisi d'enfreindre localement cette organisation en fréquences afin de regrouper les voyelles sur la première ligne du clavier.

La touche intitulée «fin» est utilisée pour signaler la fin de saisie du mot lorsqu'une nouvelle entrée est ajoutée au lexique personnel de l'utilisateur. Un système d'onglets (les touches bleu-gris au bas du clavier) permet à l'utilisateur d'accéder aux autres jeux de caractères. On trouve notamment le clavier des majuscules, celui des chiffres, parenthèses, et symboles mathématiques, celui des caractères accentués, et le clavier des caractères spéciaux.

Afin d'éviter de changer d'onglets à chaque saisie de lettres accentuées ou majuscules, les touches du clavier de saisie représentent en fait plusieurs caractères. La touche `e` par exemple permet de saisir le caractère *e* ou ses diacritiques *è*, *é*, *ê* et *ë* ou le caractère *E*. La désambiguïsation entre ces différents caractères est réalisée automatiquement lorsque l'utilisateur sélectionne le mot souhaité dans la liste des mots proposés par le moteur de prédiction. Ce mécanisme permet de limiter l'utilisation des claviers de lettres accentuées ou de majuscules à la saisie des nouveaux mots ajoutés par l'utilisateur à son lexique personnel. Il offre de plus à l'utilisateur rencontrant à l'usage des problèmes d'accentuation la possibilité d'accéder à la bonne orthographe du mot.

2.3. *La composition de messages*

Le clavier dynamique de prédiction de mots occupe une place centrale dans la procédure de composition de messages. Il permet d'une part de réaliser une économie d'interactions et un gain de temps considérable. D'autre part, compte tenu de la très bonne couverture du lexique, la consultation de la liste des mots proposés offre à l'utilisateur un procédé pour vérifier l'orthographe du mot en cours de composition. La composition d'une phrase s'effectue mot à mot en s'appuyant sur les propositions du système. Pour composer une phrase, il est recommandé de respecter le cycle d'opérations suivant :

– lecture de la liste des mots proposés par le moteur de prédiction ;

– sélection du mot si il est présent dans la liste ; le mot est ajouté à la phrase en cours de composition, la composition du mot suivant débute ; les ajouts des espaces entre les mots et des majuscules en début de phrase sont gérés automatiquement ;

– le mot n'est pas présent dans la liste, ajout d'une lettre au mot en cours de composition ; l'opération est à reproduire jusqu'à l'apparition du mot souhaité dans la liste des propositions ;

– en fin de phrase, saisie d'une ponctuation.

À tout moment, une touche de correction permet à l'utilisateur d'annuler la ou les dernières opérations effectuées. Une fois le message composé, la PCA offre la possibilité de faire prononcer le texte par une synthèse vocale, de mémoriser le texte dans un fichier, d'envoyer le message par e-mail ou d'effectuer des opérations d'édition sur une partie du message.

3. Description du moteur de prédiction de mots

Le système de prédiction de mots implanté dans la PCA utilise un lexique général du français de 320 000 formes orthographiques dont les fréquences d'usage et les traits morphosyntaxiques associés sont fournis. Un lexique personnel permet de compléter le lexique général en mémorisant les mots inconnus composés par l'utilisateur. Un module d'apprentissage stocke les phrases produites au cours de la composition des messages et calcule les fréquences d'usage propres à l'utilisateur. Le moteur de prédiction possède ainsi un modèle utilisateur qui s'enrichit au fur et à mesure de l'utilisation du système. Le principe directeur du moteur de prédiction de mots est de proposer les mots les plus fréquents sur la base des informations de fréquence d'usage contenues dans les lexiques. Un module de prédiction morphosyntaxique vient ensuite modifier les fréquences lexicales en fonction du contexte syntaxique du mot en cours de saisie.

Le schéma de fonctionnement du moteur de prédiction est présenté figure 2. Le système propose à l'utilisateur N_{prop} propositions affichées dans un clavier dynamique (le nombre de propositions affichées est réglable, de 1 à 9 propositions). La liste des

```
Si événement Nouveau mot ou ponctuation
   lettres ← début de mot
   Si début de phrase, contexte ← début de phrase
   Sinon, contexte ← ajout du mot ou de la ponctuation

   freq_L ← fréquences des lexiques
   freq_M ← prédiction morphosyntaxique(freq_L, contexte)
   freq_P ← prédiction phrases utilisateur(freq_M, contexte)
   fréquences ← élision(freq_P, contexte)

Si événement Nouvelle lettre
   lettres ← ajout de la lettre
   freq_PP ← filtre les mots précédemment proposés(fréquences)
   fréquences ← filtre(freq_PP, lettres)

liste de mots ← les N_prop mots les plus fréquents(fréquences)
propositions ← tri par ordre alphabétique(liste de mots)
```

Figure 2. *Le schéma global de fonctionnement du moteur de prédiction de mots*

mots proposés évolue à chaque fois que l'utilisateur sélectionne une ponctuation ou un mot dans le clavier de propositions (`événement Nouveau mot ou ponctuation`) ou ajoute une lettre au mot en cours de composition (`événement Nouvelle lettre`).
Lorsqu'un nouveau mot ou une ponctuation est saisi, il est ajouté au contexte de la phrase en cours de composition (dans le cas d'un début de phrase, ce contexte est réinitialisé). Les fréquences provenant des lexiques sont initialisées (voir sections 3.1 et 3.2 pour plus de détails). Le module de prédiction morphosyntaxique modifie la fréquence de chaque entrée lexicale en fonction des traits morphosyntaxiques qui lui sont associés, et compte tenu du contexte syntaxique de la phrase (voir section 3.3). Le module permettant de donner plus de poids aux phrases déjà composées par l'utilisateur est ensuite appliqué (voir section 3.4). Enfin, dans les cas d'élision, les fréquences des graphies à rejeter sont mises à zéro.
Lorsqu'une nouvelle lettre est ajoutée au mot en cours de composition, deux opérations sont effectuées. La première consiste à mettre à zéro les fréquences des entrées correspondant aux propositions précédemment affichées (et donc non sélectionnées par l'utiliseur). La seconde opération consiste à appliquer un filtre qui élimine toutes les entrées qui ne commencent pas par les lettres déjà saisies.
Pour les deux types d'événements, les $N_{\texttt{prop}}$ mots les plus fréquents sont ensuite extraits sur la base des fréquences recalculées et finalement triés par ordre alphabétique pour être affichés dans le clavier de proposition.

3.1. *Le lexique général*

Le lexique utilisé par le moteur de prédiction de mots de PCA est extrait du lexique DicoLPL (VanRullen *et al.*, 2005). Il s'agit d'un lexique très couvrant du français comportant 440 000 entrées défactorisées correspondant à 320 000 formes orthographiques différentes. Pour chaque entrée, DicoLPL fournit la forme orthographique, la forme phonétisée, la catégorie morphosyntaxique, le lemme, et la fréquence d'usage. Les fréquences d'usage ont été calculées sur un corpus de 143 millions de mots environ tirés du journal *Le Monde* (VanRullen *et al.*, 2005).

La figure 3 illustre quelles sont les propriétés de couverture associées au lexique DicoLPL. Sur l'axe des abscisses, les formes ont été classées par fréquences d'occurrence décroissantes. La forme 1 correspond ainsi à la préposition *de*, la forme apparaissant le plus fréquemment en français, suivie de la deuxième forme la plus fréquente *et* , etc. Sur l'axe des ordonnées est noté le taux de couverture des n-premières formes les plus fréquentes. Ainsi, les 10 formes les plus fréquentes composent en moyenne $21\ \%$ des énoncés du français, les 10 000 formes les plus fréquentes composent en moyenne $90\ \%$ des énoncés, etc. Un lexique limité aux 10 000 formes les plus fréquentes couvrirait en moyenne $90\ \%$ du français, ou autrement dit, pour des énoncés de 100 mots, 10 mots en moyenne seraient absents de ce lexique de 10 000 formes. Il est intéressant de noter que les 54 000 formes les plus fréquentes couvrent $99\ \%$ du français et que seules 181 000 des 444 000 formes composant notre lexique ont été observées dans le corpus de 143 millions de mots analysés. Les formes peu usitées,

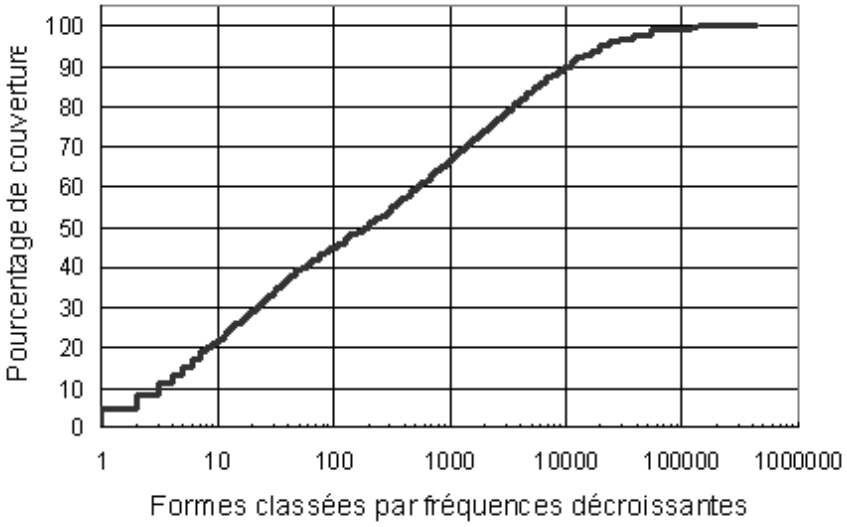

Figure 3. *Le taux de couverture du lexique dicoLPL en fonction des n-premières formes classées par fréquence décroissante*

au nombre de 263 000, soit 60 % du lexique environ, sont en majorité composées de verbes (à 80 %), de noms communs (à 15 %) et d'adjectifs (à 5 %).

Chaque entrée du lexique général de PCA correspond à une forme orthographique à laquelle est associée sa fréquence générale f_G et la liste des catégories morphosyntaxiques caractérisant le mot. Cette liste possède ainsi plusieurs éléments en cas d'ambiguïté syntaxique (e.g. l'entrée `montre` propose `{Noun, Verb}` comme catégories morphosyntaxiques).

3.2. *Le lexique personnel et les fréquences propres à l'utilisateur*

Lorsque l'utilisateur compose un mot inconnu, c'est-à-dire absent du lexique général, le mot est mémorisé, après une demande de confirmation, dans le lexique personnel de l'utilisateur. Par défaut, les mots inconnus ajoutés sont classifiés comme nom propre (c'est souvent pertinent en pratique compte tenu de la bonne couverture du lexique général). Les entrées du lexique général et du lexique personnel sont fusionnées pour former la liste des mots du lexique global.

Après chaque utilisation, la production de l'utilisateur est analysée. Les fréquences d'usage propres à l'utilisateur sont mémorisées dans un fichier qui s'enrichit au cours

des utilisations répétées. La prise en compte des fréquences propres est modélisée pour chaque entrée i de la liste des mots du lexique global (de taille N_L) par l'équation :

$$f_{L,i} = \frac{1}{A_L}(N_G \times f_{G,i} + N_A \times f_{P,i}) \text{ avec } A_L \text{ tel que } \sum_{i=1}^{N_L} f_{L,i} = 1 \qquad [1]$$

La fréquence $f_{L,i}$ affectée à l'entrée i est une moyenne pondérée entre sa fréquence générale $f_{G,i}$ et sa fréquence propre $f_{P,i}$, le poids affecté aux fréquences propres étant proportionnel au nombre N_A de mots produits par l'utilisateur depuis la première utilisation du logiciel, le poids N_G affecté aux fréquences générales étant constant. L'influence des fréquences propres augmente donc linéairement avec le nombre N_A de mots produits au cours de l'apprentissage. La valeur de la constante N_G détermine la vitesse d'apprentissage du système. Lorsque le nombre de mots produits N_A atteint N_G, les fréquences propres contribuent en moyenne à même hauteur que les fréquences du lexique général.

Nous avons adopté une valeur de $N_G = 800\,000$ pour la constante régissant la vitesse d'apprentissage. Avec une telle valeur pour N_G, on s'attend à ce que les fréquences ne soient que peu modifiées en moyenne (e.g. après un apprentissage de $8\,000$ mots, la contribution moyenne des fréquences propres s'élève à $1\ \%$). Les mots les plus fréquents du lexique général ne verront donc pas leurs fréquences significativement se modifier. En revanche, les entrées moins fréquentes (disons les formes de la figure 3 dont le rang en fréquence décroissante est supérieur à $5\,000$) seront les plus affectées par les variations de fréquence propre. Nous illustrons cet effet en prenant l'exemple concret suivant. L'adjectif *handicapées* apparaît dans le lexique général avec un rang en fréquence de $19\,398$ correspondant à une fréquence de $2,871\,10^{-6}$. Après un apprentissage de $8\,000$ mots, l'utilisateur a composé 25 fois la forme *handicapées*, soit une fréquence propre de $3,125\,10^{-3}$. La nouvelle fréquence d'usage donnée par l'équation [1] est égale à $3,412\,10^{-5}$, ce qui fait remonter la forme *handicapées* au $2\,978$ rang dans la liste des formes classées par fréquence décroissante. La prise en compte des fréquences propres de l'utilisateur par le modèle décrit équation [1] a donc pour effet de favoriser les formes peu fréquentes dans le lexique général mais pourtant produites par l'utilisateur, après un nombre de mots appris de taille raisonnable.

3.3. *Le module de prédiction morphosyntaxique*

Le moteur de prédiction intègre un module morphosyntaxique qui modifie la fréquence de chaque entrée du lexique selon la liste des catégories morphosyntaxiques associée à l'entrée et en fonction du contexte syntaxique dans la phrase en cours de composition. Le modèle de langage adopté est ici un modèle stochastique sans notion de constituants ni de structures imbriquées (le système HandiAS (Le Pévédic, 1998) par exemple utilise une grammaire plus complexe). La grammaire probabiliste est apprise sur un corpus de phrases annotées morphosyntaxiquement. Nous avons utilisé le corpus du projet CLIF (des extraits tirés du journal *Le Monde* contenant $370\,000$

Adjectif	$\texttt{Adj}_{\texttt{i}}$ indéfini, $\texttt{Adj}_{\texttt{q}}$ qualificatif
Conjonction	$\texttt{Conj}_{\texttt{c}}$ coordination, $\texttt{Conj}_{\texttt{s}}$ subordination
Déterminant	Det
Nom	$\texttt{Noun}_{\texttt{c}}$ commun, $\texttt{Noun}_{\texttt{p}}$ propre
Pronom	Pro
Adverbe	Adv
Préposition	$\texttt{Prep}_{\texttt{a}}$ à, $\texttt{Prep}_{\texttt{d}}$ de, $\texttt{Prep}_{\texttt{o}}$ autres prépositions
Auxiliaire avoir	$\texttt{V}_{\texttt{an}}$ infinitif, $\texttt{V}_{\texttt{ap}}$ part. présent , $\texttt{V}_{\texttt{as}}$ part. passé, $\texttt{V}_{\texttt{ac}}$ conjugué
Auxiliaire être	$\texttt{V}_{\texttt{en}}$ infinitif, $\texttt{V}_{\texttt{ep}}$ part. présent, $\texttt{V}_{\texttt{es}}$ part. passé, $\texttt{V}_{\texttt{ec}}$ conjugué
Verbe	$\texttt{V}_{\texttt{mi}}$ infinitif, $\texttt{V}_{\texttt{mp}}$ part. présent, $\texttt{V}_{\texttt{ms}}$ part. passé, $\texttt{V}_{\texttt{mc}}$ conjugué
Ponctuation	$\texttt{Punct}_{\texttt{d}}$ dure, $\texttt{Punct}_{\texttt{m}}$ molle

Tableau 1. *Les* 24 *catégories morphosyntaxiques utilisées*

mots environ (Abeillé *et al.*, 2001)). L'information syntaxique est représentée sous la forme de 24 catégories morphosyntaxiques (voir tableau 1) pour lesquelles les traits de genre, de nombre et de personne sont disponibles lorsque pertinents.

Nous utilisons le modèle des patrons (Blache et Rauzy, 2006) pour calculer la probabilité de chaque entrée du lexique global compte tenu du contexte syntaxique constitué par la séquence des formes orthographiques déjà saisies dans la phrase. Le modèle des patrons, une sous-classe des modèles de Markov cachés (voir par exemple (Rabiner, 1989)), présente des avantages par rapport aux modèles de N-grammes que nous décrivons appendice A. La phase d'apprentissage de la grammaire probabiliste sur le corpus CLIF fournit 224 patrons de taille variable qui permettent de calculer la probabilité en contexte à affecter aux entrées du lexique.

À chaque entrée i du lexique global est associé un vecteur de taille 24 (le nombre total de catégories) noté $\{\mathrm{c_{MS}}\}_i$. La j-ième valeur du vecteur est mise à 1 si la catégorie j fait partie des catégories morphosyntaxiques associées à la forme orthographique i. Le vecteur $\{\mathrm{c_{MS}}\}_i$ représente ainsi les catégories possibles pour l'entrée i. Le modèle des patrons se comporte comme un automate à états finis dont l'évolution est régie par l'ajout successif des vecteurs $\{\mathrm{c_{MS}}\}_t$ associés aux mots composant la phrase en cours de composition (le contexte syntaxique). À chaque étape, le modèle permet d'évaluer la probabilité pour que le contexte soit suivi de tel ou tel vecteur $\{\mathrm{c_{MS}}\}_i$. Nous notons cette probabilité $P_{\mathrm{t}}(\{\mathrm{c_{MS}}\}_i)$ (voir équation [9]). Elle est calculée pour chacune des entrées du lexique. La prise en compte de la prédiction morphosyntaxique est finalement donnée par l'équation :

$$f_{M,i} = \frac{1}{A_M}(f_{L,i} \times P_{\mathrm{t}}(\{\mathrm{c_{MS}}\}_i)) \text{ avec } A_M \text{ tel que } \sum_{i=1}^{N_L} f_{M,i} = 1 \qquad [2]$$

La fréquence $f_{M,i}$ affectée à l'entrée i est la fréquence $f_{L,i}$ obtenue équation [1] qui rend compte des fréquences propres de l'utilisateur, pondérée par l'influence du contexte syntaxique de la phrase en cours de composition.

3.4. *L'arbre des phrases de l'utilisateur*

Nous avons vu section 3.2 que la production de l'utilisateur était analysée après chaque utilisation afin d'extraire les fréquences d'usage propres. De même, les phrases produites sont mémorisées sous forme d'arbre. Le nœud racine correspond au début de phrase. Les nœuds fils de la racine sont identifiés par la forme orthographique de tous les premiers mots débutant les phrases produites, et ainsi de suite. Chaque phrase produite est donc représentée par une branche de l'arbre. Chaque nœud possède un compteur qui est incrémenté à chaque fois que le chemin passant par ce nœud est emprunté (e.g. si deux phrases ont été produites *je vais bien* et *je suis fatiguée*, le compteur du nœud `je` en première position dans la phrase vaut $n_{\texttt{je}} = 2$, les autres valant $n_{\texttt{je,vais}} = 1$, $n_{\texttt{je,suis}} = 1$, $n_{\texttt{je,vais,bien}} = 1$, $n_{\texttt{je,suis,fatiguée}} = 1$).

L'arbre des phrases est utilisé par le moteur de prédiction de mots pour proposer en priorité les mots complétant les débuts de phrases déjà produits par l'utilisateur. Le modèle implanté fonctionne comme suit. À chaque début de phrase composé par l'utilisateur, le nœud `contexte` est réinitialisé sur la racine de l'arbre. Les N_s fils du nœud `contexte` sont examinés et une table donnant la fréquence relative de chacun des fils est calculée à partir des valeurs n_j des compteurs des nœuds fils :

$$\text{Pour } j \in [1, N_s],\ f_j = \frac{n_j}{N_T} \text{ avec } N_T = \sum_{i=1}^{N_s} n_j \qquad [3]$$

Lorsqu'un mot est ajouté à la phrase en cours de composition, ce mot est comparé à l'identifiant de chacun des fils du nœud `contexte`. Si un fils correspond au mot ajouté, on descend dans l'arbre en associant le nœud `contexte` au fils correspondant. L'opération est répétée tant qu'il existe un chemin dans l'arbre décrivant la phrase en cours de composition. Les fréquences évaluées équation [3] sont intégrées au modèle général. Pour chaque entrée i du lexique, la quantité $f_{T,i}$ vaut f_j si l'entrée est le j-ième fils du nœud `contexte`, $f_{T,i}$ vaut 0 autrement. La fréquence $f_{S,i}$ de chaque entrée du lexique est déduite de la fréquence $f_{M,i}$ donnée équation [2] par l'équation :

$$f_{S,i} = \frac{1}{A_S}(N_S \times f_{M,i} + \text{Min}(N_{\text{seuil}}, N_T) \times f_{T,i}) \text{ avec } A_S \text{ tel que } \sum_{i=1}^{N_L} f_{S,i} = 1 \qquad [4]$$

La fréquence $f_{S,i}$ est une moyenne pondérée entre $f_{M,i}$, qui rend compte des fréquences d'usage propres à l'utilisateur et de la prédiction morphosyntaxique, et $f_{T,i}$ qui donne la distribution en fréquence des mots complétant les débuts de phrases déjà produits. Le poids associé à la prédiction par complétion de phrases est $\text{Min}(N_{\text{seuil}}, N_T)$, qui vaut N_T si $N_T < N_{\text{seuil}}$ et N_{seuil} sinon. Ce poids est proportionnel au nombre N_T de fois que la phrase a déjà été produite, jusqu'à un nombre seuil que nous avons fixé à $N_{\text{seuil}} = 20$. Le poids alloué à la prédiction au sortir du module morphosyntaxique est fixé à $N_S = 5$. Le seuil est incorporé au modèle afin de limiter l'influence de la prédiction par complétion de phrases. La contribution maximum du module est obtenue lorsque le contexte a été produit plus de N_{seuil} fois par l'utilisateur, et atteint la valeur $f_{\max} = N_{\text{seuil}}/(N_S + N_{\text{seuil}})$, soit 80% des fréquences.

4. L'évaluation du système de prédiction

L'évaluation écologique d'un système d'aide à la communication pour personnes handicapées est un problème complexe qui fait appel à des domaines de connaissances aussi variés que la psychologie, l'ergonomie, la psycholinguistique et la linguistique.
Le premier problème rencontré est celui du dispositif d'interaction entre l'utilisateur et la machine support du système. Un certain nombre de solutions techniques (contacteurs, capteur de mouvements, capteur de souffle,...) sont aujourd'hui proposées mais il est difficile d'évaluer leur efficacité en termes de taux d'échecs, de fatigabilité, de confort. L'évaluation et le diagnostic du dispositif le mieux adapté à la motricité de l'utilisateur sont pourtant fondamentaux, ils conditionnent en pratique l'utilisation ou non du système d'aide à la communication.
L'évaluation de l'ergonomie de l'interface du logiciel est un autre problème. Il s'agit ici de quantifier en termes de durée et de nombre d'interactions les différentes tâches effectuées par l'utilisateur au cours de la composition de messages. Les coûts associés à la tâche de répérage d'un mot dans une liste, d'un caractère dans un clavier dynamique de lettres ou le coût associé à la procédure d'annulation en cas d'erreur de saisie, doivent être estimés. De nombreuses métriques d'évaluation sont proposées dans la littérature (voir à ce sujet (Vigouroux *et al.*, 2004)), mais qui ne couvrent qu'une partie de la totalité des processus cognitifs mis en œuvre.
Enfin, la procédure d'évaluation doit prendre en compte les compétences linguistiques et la stratégie de communication de chaque utilisateur. Il est montré dans (Wandmacher et Antoine, 2006) que les performances des aides à la prédiction de mots varient de façon importante suivant le type de registres de langage considérés. Comment modéliser alors le registre de langage produit par l'ensemble des utilisateurs d'un système d'aide à la communication ? De même, quel poids doit-on accorder aux erreurs linguistiques produites par l'utilisateur durant le processus d'évaluation (voir à ce sujet (Boissière *et al.*, 2007)) ?

Nous avons adopté dans cet article la démarche proposée par (Boissière et Schadle, 2006) qui consiste à n'évaluer qu'une sous-partie du système, à savoir l'impact du moteur de prédiction de mots sur les performances de l'utilisateur. La métrique utilisée pour évaluer le modèle de langage du système sur un corpus est le *taux d'économie de saisies* défini par l'équation :

$$\text{Keystrokes saving rate} = 1 - \frac{\text{Nbr de caractères saisis} + \text{Nbr de validations}}{\text{Nbr de caractères total du corpus}}$$

Le nombre de caractères saisis ne tient pas compte du procédé employé pour sélectionner les caractères sur le clavier virtuel de lettres (voir section 2.2). Nous n'évaluons donc pas ici l'apport de l'organisation du clavier de lettres. Le nombre de validations correspond au nombre de sélections de mot dans la liste de mots proposés. Le modèle ne rend pas compte du coût cognitif associé au repérage du mot dans la liste des propositions et des possibles carences en orthographe de l'utilisateur.

Le taux d'économie de saisies est une mesure de la performance du moteur de prédiction de mots du système. Il possède une valeur optimale théorique qui est atteinte

si dans $100\ \%$ des cas, le mot que souhaite composer l'utilisateur est présent dans la liste des propositions sans avoir à utiliser le clavier de lettres. Pour ce cas idéal, le taux d'économie de saisies est égal au nombre de mots saisis (identique au nombre de validations) divisé par le nombre de caractères total du message. Cette limite est en moyenne d'environ $80\ \%$ pour le français.

4.1. *Le corpus de test et la procédure d'évaluation*

Le corpus que nous avons utilisé pour effectuer l'évaluation du modèle de langage sous-jacent au moteur de prédiction de mots de la PCA Orthographique est extrait du corpus Multitag (Paroubek et Rajman, 2000). Il s'agit d'une autobiographie d'environ $52\,000$ mots relatant les mémoires de guerre d'un général durant la Première Guerre mondiale. Certains passages sont à la première personne du singulier, d'autres sont descriptifs et incluent de nombreux noms propres faisant référence à des personnes ou des lieux-dits. Les temps employés alternent entre le passé et le présent et quelques envolées lyriques sont même présentes. Nous avons sélectionné ce corpus sur les bases de son hétérogénéité temporelle à mesure que l'auteur avance dans le récit. Le modèle utilisateur de notre système fonctionnant par apprentissage, il est en effet important de tester si il est aussi capable de «désapprendre». Notre corpus d'évaluation ne fait pas partie des textes ayant servis à calculer les fréquences du lexique général ni à apprendre la grammaire probabiliste du module de prédiction morphosyntaxique.

Le corpus a été découpé en tranches de 500 mots sur lesquelles a été appliquée la procédure d'évaluation. Nous avons été amenés à adapter le programme du logiciel PCA afin d'automatiser cette évaluation. Néanmoins, certaines contraintes techniques rendent la procédure coûteuse en temps, de l'ordre de 30 mn par tranche de 500 mots. Il n'a pas été possible, pour cette raison, d'évaluer notre système pour différents types de registres de langage comme dans (Wandmacher et Antoine, 2006) ou de traiter le corpus de référence proposé dans (Boissière et Schadle, 2006). Les valeurs obtenues pour le taux d'économie de saisies que nous présentons dans la suite de l'article sont donc propres à notre corpus de test.

4.2. *Influence du modèle utilisateur*

La figure 4 présente l'évolution du taux d'économie de saisies en fonction du nombre de mots engrangés par le modèle utilisateur. Les mesures ont été réalisées par tranches de 500 mots, la taille de la liste de mots proposée à l'utilisateur a été fixée à 9. Le taux d'économie de saisies augmente avec la taille de l'échantillon d'apprentissage jusqu'à atteindre un plateau aux alentours de $2\,000$ mots appris environ. Le taux d'économie de saisies fluctue ensuite autour de sa valeur moyenne (environ $55\ \%$) avec un écart-type entre tranches de $1,5\ \%$. Les différents modules d'apprentissage implantés dans le moteur de prédiction de mots (lexique personnel, fréquences propres à l'utilisateur, et arbre des phrases de l'utilisateur) contribuent à hauteur de $3\ \%$ au gain en performances du système. L'apprentissage est rapide, le système at-

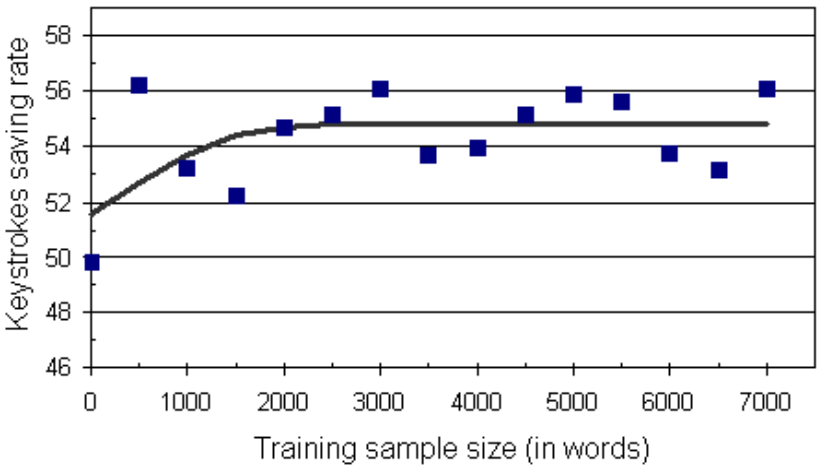

Figure 4. *Évolution du taux d'économie de saisies en fonction du nombre de mots déjà composés par l'utilisateur (pour 9 propositions). Un plateau apparaît lorsque 2000 mots environ ont été appris par le module d'apprentissage du modèle utilisateur*

teint son régime de croisière lorsque l'utilisateur a composé $2\,000$ mots. Le plateau observé correspond à une saturation du système de prédiction, certaines formes orthographiques sont favorisées par le module d'apprentissage, mais au détriment d'autres formes tout aussi pertinentes. Nous avons testé en avancant dans le récit le comportement du système à plus grande échelle, jusqu'à $25\,000$ mots appris. Le plateau perdure avec une dispersion moyenne de $2,5\,\%$ entre tranches. Le modèle utilisateur que nous avons adopté ne semble donc pas trop invasif.

4.3. *Influence du nombre de propositions*

La figure 5 présente la variation du taux d'économie de saisies en fonction de la taille de la liste des propositions. Les résultats sont obtenus sur une tranche de 500 mots et lorsque $8\,000$ mots ont été appris par le modèle utilisateur. Les barres en gris foncé représentent les résultats pour le moteur complet. Le taux d'économie de saisies augmente comme attendu avec le nombre de propositions, mais un tassement est observé à partir de 5 propositions (augmentation de $4\,\%$ entre 3 et 5, de $2,5\,\%$ entre 5 et 7, et de $1,5\,\%$ entre 7 et 9 propositions). Il n'apparaît donc pas nécessaire d'augmenter plus encore le nombre de propositions, le gain en performance étant contrebalancé par le coût cognitif associé à la tâche de répérage du mot recherché dans la liste des propositions.

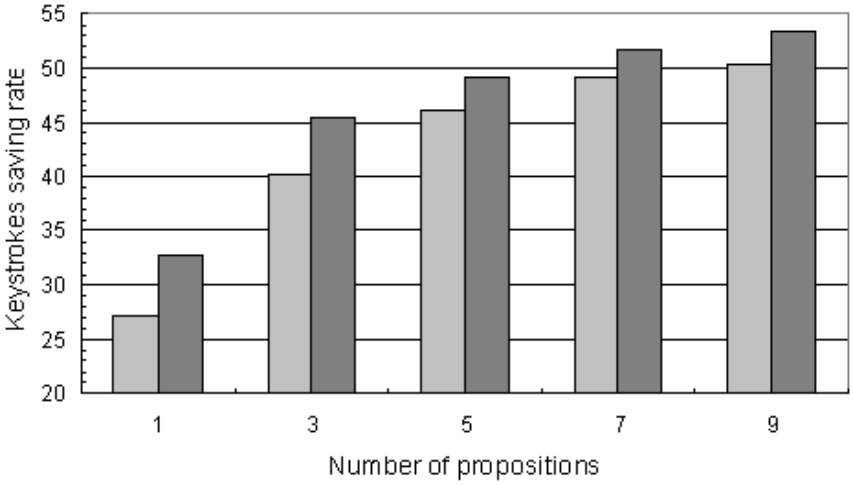

Figure 5. *Le taux d'économie de saisies en fonction de la taille de la liste des propositions (1, 3, 5, 7 ou 9 mots proposés). En gris clair, les valeurs pour une prédiction basée seulement sur les fréquences du lexique général. En gris foncé, l'économie réalisée pour le moteur complet (fréquences + apprentissage + module morphosyntaxique)*

La performance du moteur de prédiction de mots de la PCA Orthographique est bonne, comparable avec les résultats obtenus par le système Sibylle (Wandmacher *et al.*, 2007). Pour une liste de 5 propositions, le taux d'économie de saisies s'élève environ à 50 %, c'est le même taux que celui mesuré pour Sibylle sur les corpus de type journaux et sans intégrer le module basé sur l'analyse sémantique latente.

Nous avons essayé d'analyser l'apport des différentes composantes de notre moteur de prédiction de mots. Les barres en gris clair sur la figure 5 représentent le taux d'économie de saisies pour notre moteur de prédiction amputé de son module de prédiction morphosyntaxique, de la prise en compte des fréquences propres à l'utilisateur et de l'arbre des phrases (pour des raisons techniques, le lexique personnel a été conservé). Le moteur de prédiction établit alors la liste des mots proposés sur la seule base des fréquences du lexique général, sans notion de contexte. On observe que c'est le facteur dominant du modèle (pour 5 propositions, le taux d'économie de saisies est supérieur à 45 %). Pour les cas à 5, 7 ou 9 propositions, l'intégration des autres modules dans le moteur élève le taux d'économie de saisies de 3 % environ. Nous avons aussi testé le moteur sans le module morphosyntaxique et avec apprentissage (gain de 2 %) et avec la prédiction morphosyntaxique et sans apprentissage (gain de 2 %).

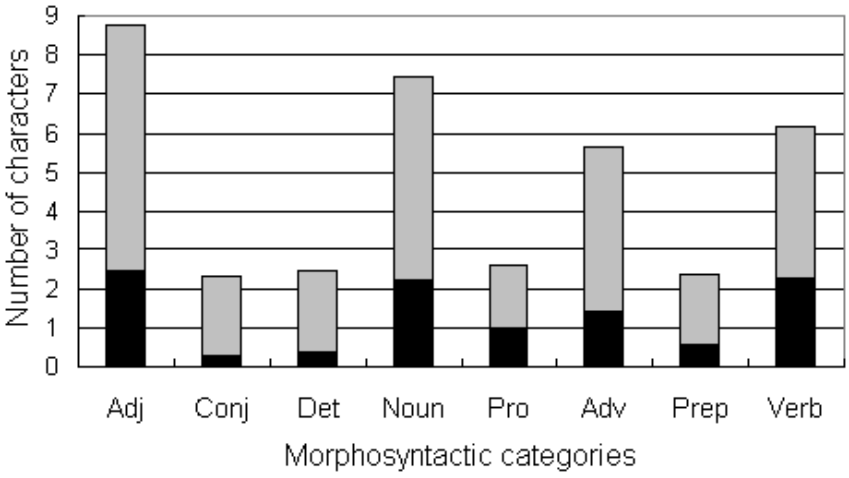

Figure 6. *Le nombre moyen de caractères par mots en fonction de leurs catégories morphosyntaxiques (en gris). En noir, le nombre moyen de caractères à saisir pour faire apparaître le mot dans la liste des mots proposés (pour 9 propositions)*

4.4. *Les résultats par catégories morphosyntaxiques*

Dans cette section, nous analysons nos résultats en prenant en compte l'information sur les catégories morphosyntaxiques. La figure 6 présente le nombre moyen de caractères par mots en fonction de la catégorie morphosyntaxique (en gris). Les barres noires représentent le nombre de caractères à saisir en moyenne pour faire apparaître le mot dans la liste des propositions (pour 9 propositions). Ces résultats sont obtenus sur une tranche de 500 mots et lorsque 8 000 mots ont été appris par le modèle utilisateur. On distingue nettement les différents rôles joués dans la phrase par les catégories morphosyntaxiques. Les catégories *Nom*, *Verbe* et *Adjectif* portant une fonction sémantique importante nécessitent la saisie d'un peu plus de 2 caractères en moyenne avant d'être trouvées par l'utilisateur dans la liste des propositions. Au contraire, les classes fermées représentées par les catégories morphosyntaxiques *Conjonction*, *Déterminant* et *Préposition* sont dans la plupart des cas proposées par le système sans avoir à saisir la première lettre du mot. Pour les catégories intermédiaires dont la fonction sémantique est moins importante, comme les catégories *Pronom* et *Adverbe*, la saisie d'environ 1 caractères en moyenne suffit à le faire apparaître dans la liste des propositions.

Nous présentons figure 7 le taux d'économie de saisies réalisé par le moteur de prédiction (9 propositions) pour chacune des catégories morphosyntaxiques (en noir). Les barres grises représentent la valeur optimale que peut atteindre ce taux, c'est-à-

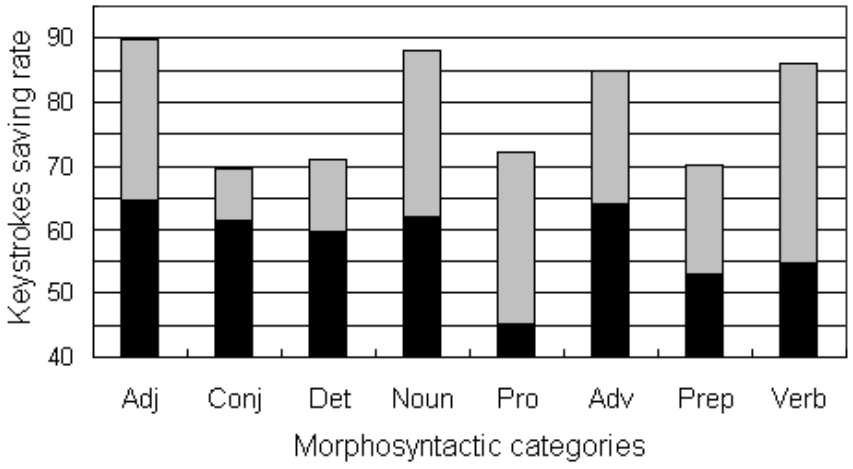

Figure 7. *Le taux d'économie de saisies en fonction de la catégorie morphosyntaxique. En gris, la valeur optimale que le taux peut atteindre. En noir, le taux mesuré pour le moteur de prédiction*

dire le cas limite idéal ou chacun des mots que souhaite composer l'utilisateur est prédit dans la liste de propositions sans avoir à utiliser le clavier de lettres. L'écart entre le taux optimal limite et le taux mesuré nous donne des pistes sur les progrès pouvant être réalisés afin d'améliorer les performances du moteur de prédiction de mots. Pour les catégories *Nom*, *Verbe* et *Adjectif* portant une fonction sémantique forte, le taux d'économie de saisies peut encore être amélioré de façon significative, par exemple en introduisant un module de prédiction sémantique dans le système. On observe d'autre part que le moteur de prédiction dans son état actuel est proche de la limite optimale pour les catégories morphosyntaxiques *Conjonction* et *Déterminant*. Le taux d'économie de saisies pour la catégorie morphosyntaxique *Pronom* est particulièrement bas dans notre système. Nous l'expliquons par le fait qu'une seule classe de pronom a été considérée dans notre modèle de prédiction morphosyntaxique. Le *type* et le *cas* du pronom, qui portent une information syntaxique forte, doivent être aussi pris en compte.

Les taux d'économie de saisies de la figure 7 sont en moyenne supérieurs au résultat de $55\ \%$ annoncé section 4.2. Dans notre évaluation, les ponctuations molles (comme les virgules) ou les ponctuations dures (comme les points finaux) sont d'abord saisies puis validées. Le taux d'économie de saisies pour les ponctuations, qui représentent en moyenne $10\ \%$ du nombre de mots du corpus, est ainsi de $-100\ \%$, ce qui fait chuter la mesure du taux d'économie de saisies global d'environ $2\ \%$.

5. Conclusion et perspectives

Nous avons présenté dans cet article le moteur de prédiction de mots de la Plateforme de Communication Alternative, version orthographique. Il utilise un lexique général du français très couvrant, 320 000 formes orthographiques pour lesquelles les fréquences d'usage et les traits morphosyntaxiques associés sont connus. Un lexique personnel permet de compléter le lexique général en mémorisant les mots inconnus composés par l'utilisateur. Le moteur de prédiction possède un modèle utilisateur qui calcule les fréquences d'usage propres à l'utilisateur et mémorise les phrases produites au cours des utilisations répétées du système. Le moteur intègre de plus un module de prédiction morphosyntaxique qui pondère les fréquences des mots prédits en fonction du contexte syntaxique de la phrase en cours de composition.

Nous avons conduit l'évaluation du système, plus précisément l'évaluation du modèle de langage du moteur de prédiction de mots, en mesurant le taux d'économie de saisies sur un corpus de test. Les résultats sont satisfaisants, le taux d'économie de saisies est d'environ $55\,\%$ pour une liste de 9 propositions et d'environ $50\,\%$ pour une liste de 5 propositions. Une analyse des contributions respectives des différents modules montre que les contributions du modèle utilisateur et du module de prédiction morphosyntaxique, même si bien réelles, sont dominées par la contribution principale provenant de la prédiction brute basée sur les fréquences d'usage du lexique général (plus de $45\,\%$ du taux d'économie de saisies pour 5 propositions).

L'analyse de ces résultats par classe de catégories morphosyntaxiques nous donne des pistes sur les moyens d'améliorer les performances du moteur de prédiction de mots actuel. Le module de prédiction morphosyntaxique montre notamment des faiblesses quant à la définition de certaines catégories morphosyntaxiques comme les pronoms par exemple. Une nouvelle version du module de prédiction morphosyntaxique sera prochainement intégrée dans la PCA Orthographique. Basée sur l'analyse du corpus Multitag (Paroubek et Rajman, 2000), un corpus annoté morphosyntaxiquement de 700 000 mots environ, la prise en compte d'informations syntaxiques plus précises devrait considérablement améliorer les performances du module.

6. Bibliographie

Abeillé A., Clément C., Kinyon A., Toussenel F., « Un corpus français arboré : quelques interrogations », *Actes de Traitement Automatique des Langues Naturelles*, vol. 1, Tours, France, p. 33-42, 2-5 juillet, 2001.

Abraham M., « Reconstruction de phrases oralisées à partir d'une écriture pictographique », *European Journal of Automation*, vol. 34, num. 6-7, p. 883-901, 2000.

Abraham M., « Altérations de la communication dialogique : Le statut de la langue dans la palliation des troubles de la parole », *Actes de la conférence IFRATH, Handicap 2006*, Paris, France, 7-9 juin, 2006.

Bellengier E., Blache P., Rauzy S., « PCA : un système d'aide à la communication alternatif évolutif et réversible », *Actes de la conférence ISAAC 2004*, Neuchâtel, Suisse, p. 78-85, 6-8 mai, 2004.

Blache P., Rauzy S., « Linguistic resources and cognitive aspects in alternative communication », *Proceedings of SICS-8*, Santiago de Cuba, p. 431-436, 2003.

Blache P., Rauzy S., « Une plateforme de communication alternative », *Actes des Entretiens Annuels de l'Institut Garches*, Issy-Les-Moulineaux, France, p. 82-93, 26-27 novembre, 2004.

Blache P., Rauzy S., « Mécanismes de contrôle pour l'analyse en Grammaires de Propriétés », *Actes de Traitement Automatique des Langues Naturelles*, Leuven, Belgique, p. 415-424, 10-13 avril, 2006.

Blache P., Rauzy S., « Le module de reformulation iconique de la Plateforme de Communication Alternative », *Actes du workshop RLCAA, conférence TALN*, vol. 2, Toulouse, France, p. 519-528, 5-8 juin, 2007.

Boissière P., Bouraoui J.-L., Vella F., Lagarrigue A., Mojahid M., Vigouroux N., Nespoulous J.-L., « Méthodologie d'annotation des erreurs en production écrite. Principe et résultats préliminaires », *Actes du workshop RLCAA, conférence TALN*, vol. 2, Toulouse, France, p. 529-538, 5-8 juin, 2007.

Boissière P., Dours D., « VITIPI : Un système d'aide à l'écriture basé sur un principe d'auto-apprentissage et adapté à tous les handicaps moteurs », *Actes de la conférence IFRATH, Handicap 2000*, Paris, France, 7-9 juin, 2000.

Boissière P., Schadle I., « Proposition d'un cadre méthodologique d'évaluation des systèmes d'assistance à la saisie de textes : Applications aux systèmes Sybille et VITIPI », *Actes de la conférence IFRATH, Handicap 2006*, Paris, France, p. 81-86, 15-16 juin, 2006.

Brangier E., Gronier G., « Conception d'un langage iconique pour grands handicapés moteurs aphasiques », *Actes de la conférence IFRATH, Handicap 2000*, Paris, France, p. 93-100, 15-16 juin, 2000.

Copestake A., « Augmented and Alternative NLP Techniques for Augmentative and Alternative Communication », *Proceedings of ACL workshop on NLP for Communication Aids*, Madrid, Spain, July 12th, 1997.

Katz S., « Estimation of probabilities from sparse data for the language model component of a speech recognizer », *IEEE Trans. ASSP*, vol. 35, p. 400-401, 1987.

Le Pévédic B., « Prédiction Morphosyntaxique évolutive dans un système d'aide à la saisie de textes pour des personnes handicapées physiques », *Thèse de Doctorat I.R.I.N. (No. ED-82-269)*, 1997.

Le Pévédic B., « Le niveau syntaxique dans le système HandiAS », *Actes de Traitement Automatique des Langues Naturelles*, Paris, France, p. 132-141, 10-12 juin, 1998.

Maurel D., Fourche B., Briffault S., « Aider la communication en facilitant la saisie rapide de textes », *Actes de la conférence IFRATH, Handicap 2000*, Paris, France, p. 87-92, 15-16 juin, 2000.

Paroubek P., Rajman M., « MULTITAG, une ressource linguistique produit du paradigme d'évaluation », *Actes de Traitement Automatique des Langues Naturelles*, Lausanne, Suisse, p. 297-306, 16-18 octobre, 2000.

Pasero R., Sabatier P., « Guided Sentences Composition : Some problems, solutions, and applications », *Proceedings of NLULP'95*, Lisbonne, Portugal, p. 97-110, 1995.

Rabiner L., « A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition », *Proceedings of the IEEE*, vol. 77, p. 257-286, 1989.

Ron D., Singer Y., Tishby N., « The power of amnesia : Learning probabilistic automata with variable memory length », *Machine Learning*, vol. 25, p. 117-149, 1996.

Schadle I., « Sibylle : Système linguistique d'aide à la communication pour personnes handicapées », *Thèse de l'Université de Bretagne Sud, Vannes, France*, 2003.

Vaillant P., « Interaction entre modalités sémiotiques : de l'icône à la langue », *Thèse de l'Université Paris XI, Orsay, France*, 1997.

VanRullen T., Blache P., Portes C., Rauzy S., Maeyhieux J.-F., Guénot M.-L., Balfourier J.-M., Bellengier E., « Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales », *Actes de Traitement Automatique des Langues Naturelles*, Paris, France, p. 41-48, 2005.

Vigouroux N., Vella F., Truillet P., Raynal M., « Evaluation of ACC for text input by two groups of subjects : Able-bodied subjects and disabled motor subjects », *8th ERCIM Workshop, User Interface for All*, Vienne, Autriche, 28-29 juin, 2004.

Wandmacher T., Antoine J.-Y., « Adaptation de modèles de langage à l'utilisateur et au registre de langage : expérimentations dans le domaine de l'aide au handicap », *Actes de Traitement Automatique des Langues Naturelles*, Leuven, Belgique, p. 630-639, 10-13 avril, 2006.

Wandmacher T., Béchet N., Barhouim Z., Poirier F., Antoine J.-Y., « Système Sibylle d'aide à la communication pour personnes handicapées : modèle linguistique et interface utilisateur », *Actes du workshop RLCAA, conférence TALN*, vol. 2, Toulouse, France, p. 539-548, 5-8 juin, 2007.

A. Le modèle des patrons

Le modèle des patrons décrit une sous-classe des modèles de Markov cachés (HMM, voir par exemple (Rabiner, 1989)) pour laquelle chaque état du système est identifié par une séquence de catégories qui sont toutes «visibles» dans l'échantillon d'apprentissage. Le modèle des N-grammes est par exemple une sous-classe du modèle des patrons pour laquelle tous les états du modèle sont identifiés par des séquences de catégories de taille identique N. Le modèle des patrons relâche cette contrainte en acceptant des états identifiés par des séquences de longueur variable (voir par exemple (Ron *et al.*, 1996)). Cette caractéristique permet en pratique d'extraire du corpus d'apprentissage un ensemble d'états, les *patrons* du modèle, qui capture de façon optimale l'information contenue dans le corpus. Après extraction de la liste des patrons, le modèle permet de probabiliser l'espace $\mathcal{S}$ composé de toutes les séquences de catégories pouvant être produites.

A.1. *Principe général*

On considère dans la suite que l'*alphabet*, c'est-à-dire l'ensemble des catégories $\mathcal{C}$ est de taille N, $\mathcal{C} = \{c_1, ..., c_j, ..., c_N\}$. L'objectif est de donner une estimation de la probabilité d'une séquence de catégories quelle qu'elle soit, par exemple la séquence

$(c_3, c_{14}, c_{12}, c_3, c_5, c_{14})$ représentant la séquence $(\mathrm{Det}, \mathrm{Noun}, \mathrm{Conj}, \mathrm{Det}, \mathrm{Adj}, \mathrm{Noun})$ de l'énoncé *la maison et le grand chêne*. La probabilité de la séquence est donnée par la formule canonique des probabilités conditionnelles, i.e.

$$P(c_3, c_{14}, c_{12}, c_3, c_5, c_{14}) = \pi_1 \times \pi_2 \times \pi_3 \times \pi_4 \times \pi_5 \times \pi_6 \qquad [5]$$

avec les termes $\pi_1 = P(c_3) = P(c_3|\emptyset)$; $\pi_2 = P(c_{14}|c_3)$, $\pi_3 = P(c_{12}|c_3, c_{14})$, $\pi_4 = P(c_3|c_3, c_{14}, c_{12})$, $\pi_5 = P(c_5|c_3, c_{14}, c_{12}, c_3)$ et $\pi_6 = P(c_{14}|c_3, c_{14}, c_{12}, c_3, c_5)$. Dans le terme π_4 par exemple $P(c_3|c_3, c_{14}, c_{12})$ est la probabilité de c_3 conditionnée par la séquence (c_3, c_{14}, c_{12}).

Le modèle des patrons vise à fournir des approximations fiables de la valeur des probabilités conditionnelles formant le produit du terme droit de l'équation [5]. Pour ce faire, nous disposons d'un corpus d'apprentissage constitué par la réalisation d'une séquence de catégories de taille importante, contenant de l'information sur la distribution des catégories et sur leurs interdépendances. Les états du système, les *patrons* σ_i, sont définis par deux informations :
- l'identifiant s_i, c'est-à-dire la séquence de catégories qui les constitue, e.g. $s_i = (c_3, c_{14}, c_{12})$;
- un vecteur de taille N, $(P_{i,1}, ..., P_{i,j}, ..., P_{i,N})$ donnant la probabilité de chaque catégorie c_j conditionnée par l'identifiant du patron σ_i, $P_{i,j} = P(c_j|s_i)$. On a par définition $\sum_{j=1}^{N} P_{i,j} = 1$.

La liste des patrons est extraite du corpus d'apprentissage en prenant garde aux problèmes d'échantillonnage. Un patron sera inclus dans le modèle si le nombre d'occurrences de sa séquence identifiante dans le corpus d'apprentissage permet une évaluation fiable des probabilités conditionnelles $P_{i,j}$. Contrairement aux modèles de N-grammes, la stratégie n'est pas ici de fixer une taille commune aux patrons, puis d'appliquer des méthodes d'interpolation pour estimer les paramètres $P_{i,j}$ pour les patrons qui ne sont pas présents dans le corpus d'apprentissage (l'information étant manquante, des choix d'heuristiques sont nécessaires pour créer cette information à partir des données). Le modèle des patrons est composé de patrons de taille variable. Le critère d'inclusion du patron σ_i dans le modèle repose sur la fiabilité de l'estimation statistique des paramètres $P_{i,j}$ à partir du corpus d'apprentissage. L'objectif est ici d'extraire une information optimale du corpus, en conservant notamment les patrons de taille importante possédant une séquence identifiante fréquemment rencontrée dans le corpus. Parmi les patrons retenus, les problèmes standard liés à l'estimation des paramètres d'une loi multinomiale (*zero-frequency problem*, etc.) sont traités par des techniques de lissage du type *backoff model* (Katz, 1987).

On considère dans la suite que nous avons extrait M patrons de taille variable définissant le modèle des patrons $\mathcal{M} = \{\sigma_1, ..., \sigma_i, ..., \sigma_M\}$. Pour chaque patron σ_i et pour chaque catégorie c_j, nous introduisons la notion de *patron cible* $\sigma_k = succ(\sigma_i, c_j)$: c'est le patron appartenant au modèle dont l'identifiant est le plus proche (du point de vue du conditionnement) lorsqu'on ajoute la catégorie c_j à l'identifiant s_i du patron σ_i. Par exemple, si $s_i = (c_{14}, c_{12}, c_3)$ et $c_j = c_5$, on recherchera dans le modèle si le patron d'identifiant $(c_{14}, c_{12}, c_3, c_5)$ existe, puis (c_{12}, c_3, c_5), puis (c_3, c_5),

et ainsi de suite, jusqu'à rencontrer le patron cible dans la liste des patrons du modèle (le patron cible est le plus grand suffixe de $(c_{14}, c_{12}, c_3, c_5)$ dans les identifiants des patrons du modèle). Le modèle des patrons $\mathcal{M}$ est finalement caractérisé par la matrice des probabilités conditionnelles $P_{i,j}$ de taille $M \times N$, et la matrice $I_{i,j}$ de taille $M \times N$ qui donne pour chaque patron σ_i et pour chaque catégorie c_j, l'indice du patron cible dans la liste des patrons du modèle.

Pour calculer la probabilité d'une séquence, la méthode consiste à approximer la formule donnée équation [5] par un produit des probabilités conditionnelles des patrons du modèle. Par exemple, imaginons que le modèle contienne 9 patrons d'identifiants $s_1 = \emptyset$, $s_2 = c_3$, $s_3 = c_{12}$, $s_4 = c_5$, $s_5 = c_{14}$, $s_6 = (c_3, c_{14})$, $s_7 = (c_3, c_5)$, $s_8 = (c_{12}, c_3)$ et $s_9 = (c_{12}, c_3, c_5)$. Les valeurs de la matrice $I_{i,j}$ utilisées pour calculer la probabilité de la séquence $S = (c_3, c_{14}, c_{12}, c_3, c_5, c_{14})$ sont successivement $I_{1,3} = 2$, $I_{2,14} = 6$, $I_{6,12} = 3$, $I_{3,3} = 8$, $I_{8,5} = 9$ et la probabilité de la séquence est approximée par :

$$P(c_3, c_{14}, c_{12}, c_3, c_5, c_{14}) \approx P_{1,3} \times P_{2,14} \times P_{6,12} \times P_{3,3} \times P_{8,5} \times P_{9,14} \quad [6]$$

La comparaison termes à termes des équations [5] et [6] permet d'obtenir les approximations effectuées : $P(c_{14}|c_3, c_{14}, c_{12}, c_3, c_5) \approx P(c_{14}|c_{12}, c_3, c_5)$, $P(c_5|c_3, c_{14}, c_{12}, c_3) \approx P(c_5|c_{12}, c_3)$ et $P(c_3|c_3, c_{14}, c_{12}) \approx P(c_3|c_{12})$. Cette approximation représente la meilleure estimation que l'on peut faire de la probabilité de la séquence, compte tenu des informations fournies par le corpus d'apprentissage.

A.2. *Calcul de la probabilité d'une séquence d'observations*

Le modèle des patrons permet de calculer la probabilité d'une séquence de catégories. Plus généralement, il offre la possibilité d'estimer la probabilité de séquences dans des cas sous-spécifiés. Nous introduisons ici la notion d'*observation*, une vecteur de taille N noté $\mathbf{o}_t = (\alpha_{t,1}, ..., \alpha_{t,j}, ..., \alpha_{t,N})$, qui caractérise l'information disponible à la position t de la séquence :
- *Mélange*[1] : l'information sur la catégorie à la position t est ambiguë et propose plusieurs solutions possibles. Dans ce cas $\mathbf{o}_t = (\alpha_{t,1}, ..., \alpha_{t,j}, ..., \alpha_{t,N})$ est caractérisée par une distribution de probabilité d'observation sur l'ensemble des catégories $\mathcal{C}$, avec par définition $\sum_{j=1}^{N} \alpha_{t,j} = 1$. C'est le cas par exemple lorsque l'information sur $\mathbf{o}_t$ est extraite d'un lexique de formes proposant plusieurs catégories pour une forme observée (e.g. la forme `montre` en entrée propose deux catégories morphosyntaxiques en sortie `{Noun, Verb}`)
- *Cas pur* : l'information sur la catégorie à la position t n'est pas ambiguë et donne c_{12} par exemple. Dans ce cas, on a pour les coefficients de $\mathbf{o}_t$: $\alpha_{t,12} = 1$ et $\alpha_{t,j} = 0$ pour $j \neq 12$.

1. Cette terminologie est employée en mécanique statistique pour modéliser les systèmes quantiques. La notion de vecteur de densité d'états et les calculs présentés dans la suite sont d'ailleurs directement issus de ce formalisme.

Cette notation permet de prendre en compte le cas particulier où aucune information n'est disponible sur la catégorie à la position t. La distribution des coefficients sur les catégories est alors équiprobable, i.e. $\mathbf{o}_t = \mathbf{o}_{NI} = (1/N, ..., 1/N, ..., 1/N)$.

L'objectif est de calculer la probabilité d'une séquence composée de la succession de T observations $S_T = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$. Le système est décrit par M états, les patrons σ_i du modèle $\mathcal{M}$, et un vecteur *densité d'états* qui caractérise la distribution de probabilité des états du système à la position t de la séquence, $\rho_t = (\rho_{t,1}, ..., \rho_{t,i}, ..., \rho_{t,M})$ avec $\sum_{i=1}^{M} \rho_{t,i} = 1$ par définition. L'évolution du système est régie par l'ajout successif des observations à la séquence. On peut montrer que la probabilité de la séquence d'observations $S_T = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$ est donnée par :

$$P(\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T) = P(\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_{T-1}) \times \sum_{i=1}^{M} \rho_{T,i} \sum_{j=1}^{N} \alpha_{T,j} P_{i,j} \qquad [7]$$

$$= \prod_{t=1}^{T} \sum_{i=1}^{M} \rho_{t,i} \sum_{j=1}^{N} \alpha_{t,j} P_{i,j}$$

avec une équation d'évolution du système gouvernant la densité ρ_t de la forme :

$$\rho_{t+1,k} = \frac{1}{A} \sum_{i=1}^{M} \rho_{t,i} \sum_{j=1}^{N} \alpha_{t,j} P_{i,j} \delta[k, I_{i,j}] \qquad [8]$$

où $\delta[k, I_{i,j}]$ est la distribution de Kroeneker (i.e. $\delta[k, I_{i,j}] = 1$ si $k = I_{i,j}$, $\delta[k, I_{i,j}] = 0$ sinon) et A un facteur de normalisation garantissant $\sum_{k=1}^{M} \rho_{t+1,k} = 1$. Notons que le vecteur densité au temps T ne dépend que des observations postérieures et n'utilise pas l'information apportée par l'observation $\mathbf{o}_T$. En pratique, le vecteur densité d'état initial ρ_1 doit être initialisé, par exemple en affectant un coefficient égal à 1 pour le patron défini par l'identifiant $\emptyset$ (patron non conditionné) et des coefficients égaux à 0 pour les autres patrons.

Le modèle des patrons permet de calculer la probabilité de toutes séquences d'observations. La complexité du calcul est faible, de l'ordre de $M \times N \times T$ opérations, pour une séquence de longueur T, dans un modèle à M patrons et N catégories. La probabilité de toutes les séquences d'une longueur T donnée est normalisée, i.e. $\sum_{S_T \in \mathcal{S}_T} P(S_T) = 1$. L'utilisation du modèle des patrons à des fins prédictives découle directement de l'équation [7]. La probabilité P_{T-1} qu'une séquence $S_{T-1} = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_{T-1})$ soit suivie de l'observation $\mathbf{o}_T = (\alpha_{T,1}, ..., \alpha_{T,j}, ..., \alpha_{T,N})$ est :

$$P_{T-1}(\mathbf{o}_T | S_{T-1}) = \sum_{i=1}^{M} \rho_{T,i} \sum_{j=1}^{N} \alpha_{T,j} P_{i,j} \qquad [9]$$