Modèle adaptatif pour la prédiction de mots

Adaptation à l'utilisateur et au contexte dans le cadre de la communication assistée pour personnes handicapées

Tonio Wandmacher, Jean-Yves Antoine

Université François Rabelais Tours, Laboratoire d'Informatique
3, place Jean-Jaurès
F-41000 Blois cedex, France

{Tonio Wandmacher ;Jean-Yves.Antoine}@univ-tours.fr

RÉSUMÉ. *Cet article présente SIBYSEM, un moteur de prédiction utilisé pour la complétion de mots dans le système de communication assistée SIBYLLE. SIBYSEM est un modèle adaptatif qui prend en compte les saisies de l'utilisateur et propose une adaptation sémantique au champ sémantique courant du discours basée sur l'analyse sémantique latente. Dans cet article, nous présentons en détail le module de prédiction ainsi que le système Sibylle dans lequel il s'intègre. De nombreuses expérimentations rendent compte des capacités d'adaptation du modèle, que ce soit pour des tâches de prédiction en général ou dans le cadre applicatif spécifique de l'aide à la communication en particulier. Nous rendons compte également de sept années d'expérience acquise avec SIBYLLE au cours de sept années d'utilisation quotidienne au sein du centre de rééducation de Kerpape.*

ABSTRACT. *This paper presents SIBYSEM, a word prediction engine, which was developed for the Sibylle AAC system. SIBYSEM incorporates an adaptive model which 1) can be trained on the messages typed by the user and 2) achieves a semantic adaptation (based on Latent Semantic Analysis) which considers the current topic of communication. Several experiments are described, showing the benefits of this adaptive behaviour. We also summarize in the paper our experience of seven years of daily use of the SIBYLLE system with patients from the Kerpape rehabilitation center.*

MOTS-CLÉS : *communication assistée, prédiction de mots, modèle de langage, adaptation, analyse sémantique latente.*

KEYWORDS: *AAC system, word prediction, language model, latent semantic analysis (LSA).*

1. Introduction : communication assistée et prédiction de mots

La prédiction de mots est une problématique qui représente un intérêt économique croissant du fait de la généralisation d'interfaces de communication à dispositifs d'entrée limités (organiseurs personnels, téléphones mobiles…). Bien que les techniques mises en œuvre soient identiques, nous nous intéresserons dans cet article à un autre champ d'application qui répond à un besoin social fort : l'aide à la communication (orale ou écrite) pour personnes lourdement handicapées.

Les communicateurs, ou systèmes de communication assistée (AAC en anglais, pour *Alternative and Augmentative Communication*) ont pour objectif de restaurer les capacités de communication de personnes souffrant d'un handicap moteur sévère. Quelle que soit la pathologie concernée (Infirmités Motrices Cérébrales, Scléroses Latérales Amyotrophiques, *Locked-In Syndrom*, lésion médullaire, myélopathie…), elle se traduit par une paralysie des membres ou par une athétose limitant de manière très pénalisante le contrôle physique de l'environnement. Les possibilités de communication à l'écrit ou sous forme de saisie de texte sur clavier sont alors excessivement réduites. Par ailleurs, l'insuffisance du contrôle moteur peut concerner jusqu'à l'appareil phonatoire. Dans ce cas, la communication est également privée de son support oral habituel.

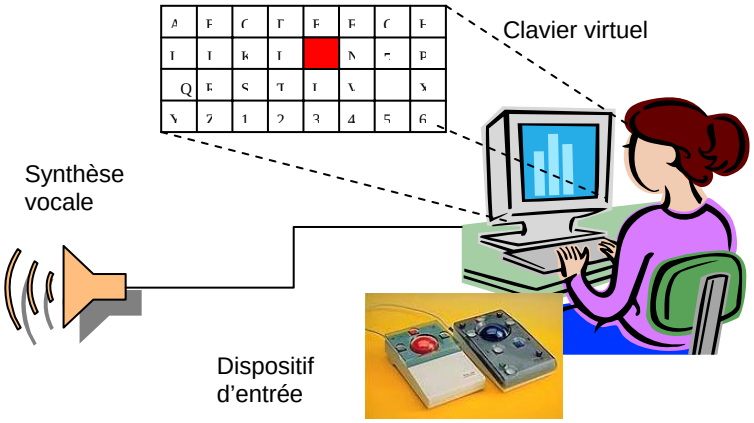

Figure 1. *Système d'aide à la communication pour personnes handicapées*

Les communicateurs proposent une aide à la saisie de messages sur un tableau virtuel de symboles (mots, lettres, phonèmes voire icônes pour les patients aphasiques ou les enfants) qui est affiché sur l'écran d'un ordinateur. Le message est construit en sélectionnant successivement sur le clavier virtuel les symboles qui le composent. Plus précisément, l'utilisation d'un système AAC repose sur trois composants principaux (figure 1). Tout d'abord, un dispositif physique joue le rôle de périphérique d'entrée de l'ordinateur. Cette interface matérielle dépend du geste libre laissé par le handicap. Il peut s'agir d'un joystick, d'une commande oculaire,

d'une commande par souffle, d'un joystick microgravité, d'un simple bouton-poussoir, etc. (figure 2). Une caractéristique importante est le nombre de degrés de liberté autorisé par ce dispositif. Le plus souvent, le patient ne peut plus réaliser que l'équivalent d'un simple clic (commande de l'environnement de type contacteur « tout ou rien »). C'est donc par clics successifs qu'il va sélectionner les éléments de son message sur le clavier virtuel, élément central de l'interaction.


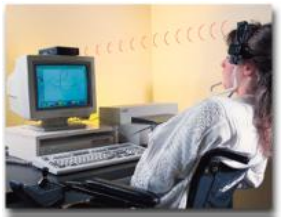

Figure 2. *Deux exemples de dispositifs d'entrée. À gauche, détecteur de souffle. À droite, détecteur de mouvements de têtes*

Différents modes de sélection peuvent ainsi être envisagés. Dans le cas du défilement linéaire, un curseur se déplace successivement sur toutes « touches » du clavier virtuel jusqu'à ce que l'utilisateur sélectionne le symbole désiré. En mode ligne/colonne, l'utilisateur doit au contraire effectuer deux appuis : tout d'abord, le clavier est balayé ligne par ligne par un curseur horizontal. Une fois sélectionnée la ligne où se trouve le symbole recherché, le curseur balaie cette dernière jusqu'à atteindre la touche correspondante. Le clavier virtuel peut être associé à un éditeur intégré, ou au contraire permettre le pilotage de toute application externe (éditeur de texte, navigateur Web, client de messagerie…). Enfin, une synthèse de parole peut-être utilisée pour vocaliser le message saisi.

Ainsi constitué, un système AAC permet à un patient lourdement handicapé de communiquer avec son entourage par écrit ou par oral. Néanmoins, le problème majeur de ces systèmes de suppléance réside, comme on peut s'en douter, dans la lenteur de la composition des messages. La tâche de saisie est généralement longue (1 à 5 mots par minute en moyenne) et extrêmement fatigante pour les patients.

Pour accélérer la composition des messages, deux approches complémentaires sont envisageables. Tout d'abord, on peut chercher à optimiser la sélection sur le clavier virtuel, avec pour objectif de limiter le nombre des défilements du curseur, celui des appuis sur le contacteur ou encore les déplacements de la souris lorsque la personne handicapée peut encore déplacer le curseur sur l'interface. Cette question concerne la disposition des touches sur le clavier (Cantegrit & Toulotte, 2001 ; Vella & Vigouroux, 2006) et plus globalement l'ergonomie de l'interface. Mais elle peut être également abordée du point de vue de l'ingénierie des langues. Dans le cas d'un défilement linéaire, il est ainsi possible de réorganiser dynamiquement le clavier simulé pour afficher en premier les caractères les plus probables compte tenu du contexte de saisie (Schadle *et al.*, 2002).

En dépit de ces optimisations, la composition d'un message symbole après symbole reste très pénible. D'où l'intérêt de techniques de complétion qui évitent la saisie de certains caractères. Une économie de saisie appréciable peut ainsi être obtenue à l'aide de systèmes de désabréviation automatique (Ricco, 2001 ; McCoy, 1995). Une approche complémentaire, très étudiée depuis maintenant plusieurs années, consiste à prédire les mots (ou groupes de mots) à venir en fonction de ceux déjà saisis. Le système SIBYLLE que nous avons développé met ainsi en œuvre une prédiction de mots avancée qui se caractérise par des capacités intéressantes d'adaptation à l'utilisateur et au contexte courant du discours.

Dans cet article, nous allons nous intéresser avant tout au moteur de prédiction du système SIBYLLE. Après une présentation générale du système et de son utilisation en conditions réelles, nous étudierons plus précisément les éléments composant le modèle de prédiction réalisé. Des études expérimentales nous permettront ensuite d'évaluer les performances de ce modèle en terme d'économie de saisie, de caractériser les facteurs ayant une influence sur son comportement et surtout de discuter plus globalement des capacités d'adaptation.

2. Présentation générale du système SIBYLLE

Avant de présenter en détail le moteur de prédiction de mots que nous avons réalisé, il est utile de présenter le fonctionnement du système SIBYLLE tel qu'il apparaît à un utilisateur. La figure 3 (page suivante) présente l'interface utilisateur tenant lieu de clavier virtuel. Bien qu'utilisable avec une souris, SIBYLLE s'adresse en premier lieu à des personnes lourdement handicapées qui ne peuvent actionner qu'un simple contacteur. Dans ce cas, il est utilisable en défilement linéaire ou ligne/colonne. SIBYLLE existe en versions française, allemande et anglaise.

L'interface de l'application regroupe plusieurs sous-claviers qui permettent respectivement de sélectionner des caractères, des nombres, des mots mais aussi des messages préenregistrés pour une communication d'urgence liée le plus souvent aux besoins vitaux de l'utilisateur. Des touches de saut de clavier permettent de naviguer d'un sous-clavier à un autre suivant les besoins de la saisie. La figure 3 permet de distinguer les différents sous-claviers de l'interface :

– le **clavier de lettres**, au centre, est utilisé pour composer les messages caractère par caractère. Il ne comprend que des caractères alphabétiques avec leurs diacritiques, ainsi que l'espace comme marque de fin de mot. Les caractères de ponctuation, très peu prévisibles, sont situés dans un sous-clavier à part (cf. infra). Cette organisation, que l'on retrouve sur les claviers physiques classiques, est dictée par le caractère dynamique du sous-clavier lorsqu'on se trouve en mode de défilement linéaire. Dans ce cas, la disposition des lettres est actualisée après chaque saisie, afin de présenter en priorité les lettres les plus probables compte tenu des caractères déjà saisis. Un module de prédiction basé sur un pentagramme de lettres (Schadle *et al*., 2002) permet ainsi de présenter la lettre attendue dans les trois

premières positions en moyenne (2,7 pour le français, 3 pour l'allemand). Les signes de ponctuation saisis sont par contre considérés par ce pentagramme ;

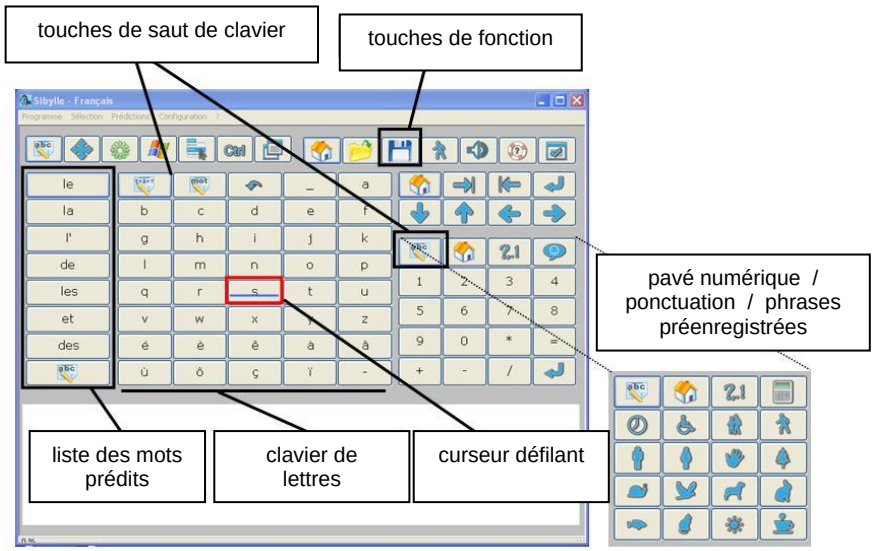

Figure 3. *Interface utilisateur du système* SIBYLLE *(version française 3.1)*

– la **liste de mots**, sur la gauche de l'interface, correspond aux prédictions lexicales données par SIBYLLE en fonction du message déjà composé. Lorsque l'utilisateur sélectionne un de ces mots, le message est alors automatiquement complété, évitant ainsi la saisie des caractères correspondants ;

– la barre horizontale de **touches fonctions**, en haut de l'interface, permet le pilotage complet de l'ordinateur. Les versions antérieures du système SIBYLLE ne comprenaient qu'un éditeur de texte intégré. Le système émule désormais intégralement le clavier physique de l'ordinateur et est donc utilisable pour piloter toute application Windows. Cela permet à l'utilisateur d'utiliser Sibylle pour écrire des textes mais également des courriers électroniques, de communiquer à « l'oral » (synthèse de parole) avec son entourage, de naviguer sur la Toile, etc. Afin de faciliter ce pilotage intégral de l'ordinateur, ce sous-clavier présente des touches de fonction réalisant des actions diverses telles que la sauvegarde d'un fichier, l'appel à la synthèse de parole ou encore l'appel du menu *Démarrer* de Windows ;

– comme sur un clavier ordinaire, le **pavé numérique** présente plusieurs modes d'utilisation qui s'affichent alternativement à la demande. Il sert à saisir des chiffres, mais permet également l'insertion de signes de ponctuation. Enfin, un dernier mode associe chaque touche du sous-clavier à la composition immédiate de messages prédéfinis. Ces messages peuvent être composés par l'utilisateur lui-même (cf. figure 3).

On notera que l'interface utilisateur est largement paramétrable pour répondre au mieux aux besoins du patient. Ce dernier peut ainsi choisir son mode de sélection (souris, ligne/colonne, linéaire), la vitesse de défilement du curseur, les durées minimales et maximales autorisées pour un appui. Ces derniers paramètres sont essentiels pour distinguer les appuis volontaires des actions non maîtrisées. De même, SIBYLLE distingue des clics longs et très longs qu'il est possible d'associer à des actions prédéfinies (effacement, mise en majuscule, aller en début de clavier...), ceci pour des utilisateurs qui arrivent à contrôler leur geste. La configuration de ces actions peut être très fine, puisque l'utilisateur a la possibilité de donner une sémantique différente à ces clics suivant le sous-clavier où se trouve le curseur.

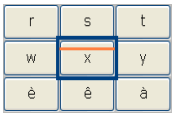

Figure 4. *Curseur avec barre glissante de temporisation*

Enfin, les personnes handicapées ont souvent du mal à anticiper les sauts du curseur d'une touche à l'autre. L'utilisateur sélectionne ainsi fréquemment la touche qui précède ou qui suit celle recherchée. Afin de réduire la fréquence de ces erreurs très pénalisantes, nous avons ajouté une barre de temporisation qui glisse doucement du haut au bas du curseur lors de chaque saut (figure 4). Ce feedback dynamique permet à l'utilisateur de paramétrer temporellement ses actions. Nos expérimentations au centre de rééducation de Kerpape (Mutualité du Morbihan) montrent qu'il permet une baisse significative des erreurs de sélection.

3. Limiter le nombre de saisies

3.1. *Prédiction de mots : modèle de base*

Le système SIBYLLE bénéficie de l'expérience de sept ans d'utilisation au centre de Kerpape. Il propose une interface utilisateur qui répond globalement aux besoins des utilisateurs handicapés. La saisie de messages lettre par lettre n'en reste pas moins pénible et fatigante. C'est pourquoi le cœur d'un communicateur réside dans sa capacité à éviter le maximum de saisies à l'utilisateur.

L'économie de saisies est normalement réalisée par un module de prédiction qui exploite les redondances dans le langage. Shannon (1951) a démontré l'ampleur de cette redondance par une approximation statistique de la probabilité d'occurrence d'un symbole en position n étant donné les $n - 1$ symboles présents à sa gauche. En testant différentes tailles de contexte n, il a pu estimer les bornes de redondance supérieure et inférieure pour une approximation d'ordre n. D'un point de vue théorique un prédicteur de mots ne fait que jouer au » jeu de Shannon » tel que ce dernier l'avait défini : étant donné un contexte gauche ($n - 1$ symboles), il essaie de prédire les mots les plus probables à partir d'une approximation statistique d'ordre n, estimée sur les mots d'un large corpus de texte (modèle de langage).

Le prédicteur de SIBYLLE repose sur un modèle quadrigramme (n = 4), qui estime la probabilité d'occurrence des termes qui peuvent suivre les trois derniers mots saisis : $P(w_i \mid (w_{i-1} w_{i-2} w_{i-3})$. Construit sur le toolkit du SRI (Stolcke, 2002) avec un vocabulaire contrôlé, il utilise un lissage de *Kneser-Ney modifié* (Goodman, 2001) et la technique de pruning proposée par (Stolcke, 1998). Après chaque saisie, SIBYLLE affiche la liste des hypothèses de plus hautes probabilités. Dans le cas d'une prédiction à large vocabulaire, le nombre d'hypothèses qui peuvent compléter une suite de trois mots est généralement important. Considérons l'énoncé suivant :

(1) *Je pense que …*

Les mots les plus probables avec ce contexte de prédiction sont selon SIBYLLE : *les*, *le*, *c'*(*est*), *nous*, *la*. Chaque proposition est cohérente, mais bien d'autres complétions pourraient être imaginées. C'est pourquoi l'utilisateur doit souvent saisir la ou les premières lettres du mot recherché pour que la complétion soit envisageable. Supposons que l'utilisateur saisisse les lettres *n* et *o* :

(2) *Je pense que no …*

SIBYLLE filtre alors les prédictions pour ne garder que les mots commençant par ces lettres. Les trois hypothèses restantes les plus probables sont : *nous*, *notre*, *non*,. Ces propositions sont cohérentes, comme le montrent les exemples ci-dessous :

(3) *Je pense que nous allons gagner.*
Je pense que notre équipe à toutes ses chances.
Je pense que non.

Le filtrage peut être étendu en supposant que les mots prédits non choisis à un moment donné ne répondent pas aux attentes de l'utilisateur. Les propositions délaissées sont donc effacées lors de la saisie suivante, même si elles restent valides dans ce nouveau contexte. Cette stratégie de filtrage est adoptée par défaut par le système SIBYLLE. Il s'agit là encore d'un paramétrage qui peut être modifié.

3.2 *Méthodologie d'évaluation et premiers résultats*

Quelle que soit la technique d'aide envisagée (désabréviation, complétion de mots), elle est évaluée par un taux d'économie de saisie observé (*KSR* pour *Keystroke Saving Rate* en anglais) :

(a) $$KSR_n = \left(1 - \frac{k_p}{k_a}\right) \cdot 100$$

où k_p est le nombre d'appuis effectivement réalisés par l'utilisateur lors de la saisie d'un message, k_a le nombre d'appuis qui auraient été nécessaires sans aide à la composition de mots et n est la taille de la liste de prédiction (normalement n = 5).

L'économie de saisie dépend de l'aide à la composition implémentée, mais également de l'interface utilisateur avec laquelle elle communique. Afin d'évaluer les performances pures de l'aide linguistique, on utilise une norme standard de

comptage des appuis. Dans le cas d'une aide par prédiction de mots, qui nous intéresse ici, nous supposerons comme (Trost *et al*., 2005) et (Trnka *et al*., 2005) que nous travaillons avec une liste de 5 mots prédits (KSR_5), qu'un unique appui supplémentaire est nécessaire pour accéder à la liste des mots, et enfin qu'un espace additionnel est automatiquement inséré en fin de mots. Les résultats que nous présentons dans cet article répondront par ailleurs tous à la stratégie de filtrage des hypothèses non sélectionnées que nous avons présentée au paragraphe précédent.

Dans cet article, nous étudierons les performances des versions française et allemande du système SIBYLLE. Les données concernant l'entraînement du modèle sont présentées dans le tableau 1 ci-dessous.

Modèle	**Données d'apprentissage**	**Vocabulaire**
4-gramme français	*Le Monde* 1998-99 (44 Millions de mots)	141 078 formes
4-gramme allemand	*Tageszeitung* 97-99 (37 Millions de mots)	141 242 formes

Tableau 1. *Modèle quadrigramme de prédiction de mots : caractéristiques.*

Dans ces conditions, le modèle quadrigramme présente des performances déjà intéressantes. Nos expériences montrent que les modèles français et allemand, entraînés sur des corpus journalistiques, permettent d'économiser plus d'un appui sur deux sur des corpus de test correspondant au même registre de langue (figure 5). Le KSR_5 observée pour le français est ainsi de 57,8 %, tandis que celle de l'allemand est de 51,6 %. La différence entre les deux langues est due à la présence de mots composés en allemand, difficilement prédictibles par un n-gramme.

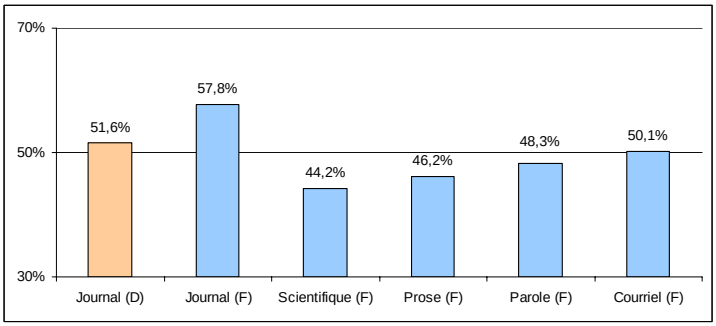

Figure 5. *KSR_5 du modèle quadrigramme de prédiction sur différents corpus correspondant à différents registres de langue (F = français ; D = allemand)*

Ces performances se dégradent toutefois dès que le système est utilisé pour d'autres registres de langue. Nous avons ainsi testé la version française de SIBYLLE sur des corpus correspondant à cinq genres langagiers différents : journal, prose (roman), parole conversationnelle, courrier électronique et rédaction scientifique (tableau 2). Les performances du modèle sont également présentées sur la figure 5. On observe une dégradation très sensible de l'aide réellement apportée par le

système. Le KSR chute ainsi à 44,2 % sur le corpus scientifique. Deux facteurs principaux peuvent être invoqués pour expliquer cette baisse de performances :

– la présence de mots hors vocabulaire, particulièrement sensible dans le cas du corpus scientifique (termes techniques) ;

– l'utilisation d'un style de langage différent de celui du corpus d'apprentissage, particulièrement visible dans le cas de la parole conversationnelle.

Corpus/Langue		**Description**	**Nb. Mots**
Journalistique	F	Journal *L'Humanité* (extrait : janvier 1999)	58 457
	D	Extrait du journal *Süddeutsche Zeitung*	56 031
Scientifique	F	Articles scientifiques (bibliographies comprises)	8 766
Littéraire	F	Extrait de *Germinal* d'Emile Zola	50 251
Parole	F	Transcription du corpus de dialogue oral OTG (Antoine *et al*., 2002) : énoncés oraux prononcés par l'hôtesse d'accueil d'un office de tourisme	15 435
Courriel	F	Ensemble de courriels personnels des auteurs ; en-têtes et réponses attachées filtrées.	44 946

Tableau 2. *Corpus de test des versions française et allemande du système* SIBYLLE

C'est pour répondre à ces limitations que nous avons élaboré un modèle de prédiction plus complexe, qui vise une double adaptation à l'utilisateur (vocabulaire et registre de langue) et au domaine sémantique du thème courant du discours.

4. Adaptation à l'utilisateur : modèle utilisateur

Cette dégradation des performances était prévisible, tant le genre journalistique est éloigné de notre façon de communiquer au quotidien. Si le recours à de tels corpus est nécessaire pour disposer de données d'apprentissage couvrant la langue, il faut ensuite adapter ce modèle général aux productions de l'utilisateur. Ce problème a déjà été largement étudié en modélisation stochastique du langage et notre démarche est de ce point vue classique : nous interpolons linéairement le modèle général avec un modèle appris sur les productions de l'utilisateur. On a :

$$P_{global}(w_i) = \lambda_u \cdot P_{utilisateur}(w_i) + (1-\lambda_u) \cdot P_{général}(w_i) \quad \text{(b)}$$

où le paramètre d'interpolation λ_u est estimé par l'algorithme EM (Jelinek, 1990). Formellement, le modèle utilisateur est un quadrigramme identique au modèle général. Il est mis à jour à la fin de chaque session d'utilisateur, si le patient (ou son thérapeute) choisit de conserver ses messages comme données d'adaptation.

Par rapport aux travaux classiques sur l'adaptation de modèles, la spécificité de la communication assistée réside dans le fait que les données utilisateurs resteront toujours limitées, du fait de l'extrême lenteur de la saisie sur les systèmes AAC. Aussi avons-nous cherché à évaluer la masse de données utilisateur suffisante pour atteindre une adaptation satisfaisante. Nous avons procédé à une évaluation incrémentale suivant un paradigme déjà utilisé dans (Boissière *et al*., 2006). L'idée

est la suivante : le corpus de test est divisé en un nombre *n* de sous-corpus. Initialement, on mesure le *KSR* du système général sur le premier sous-corpus. Ce corpus est alors utilisé comme donnée d'apprentissage par le modèle utilisateur ainsi que pour l'adaptation du coefficient d'interpolation par l'algorithme EM. Le modèle adapté qui en résulte est ensuite évalué sur le second sous-corpus. L'évaluation se poursuit ainsi de manière incrémentale, le modèle adaptatif appris sur les *n* premiers sous-corpus étant évalué sur le (*n*+*1*)-ième sous-corpus.

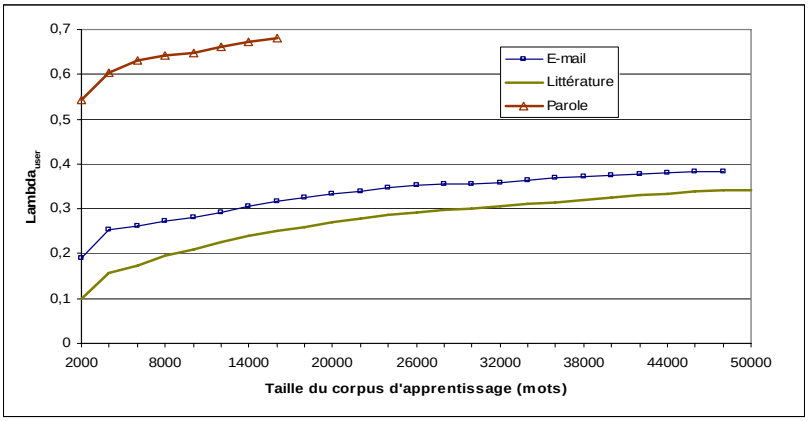

Figure 6. *Apprentissage incrémental du module de prédiction de mots avec adaptation utilisateur : évolution du coefficient d'interpolation*

Cette étude a été menée sur trois corpus : courriel, prose et parole conversationnelle. La figure 6 décrit l'évolution du facteur d'interpolation au cours de l'apprentissage. L'adaptation est manifeste : au bout de 10 000 mots saisis, la part prise par le modèle utilisateur dans l'estimation des probabilités dépasse 20 % dans tous les cas. Le modèle utilisateur prend même une part majoritaire avec le corpus de dialogue oral ($\lambda_u \approx 0{,}7$). Cette extrême adaptation résulte de la spécificité de la parole conversationnelle par rapport à l'écrit. À l'opposé, le modèle de langue général reste prédominant dans les deux autres cas ($\lambda_u \approx 0{,}3$).

On observera toutefois qu'il n'y a pas de corrélation entre l'importance de l'adaptation et la dégradation des performances relevée précédemment (figure 5). L'adaptation à l'utilisateur n'est donc pas la seule source d'amélioration du système. La figure 7 présente précisément l'évolution du *KSR* au fil de l'apprentissage. Plusieurs observations peuvent être faites. Tout d'abord, l'influence du modèle utilisateur est significative rapidement, ce qui avait déjà été observé avec le système VITIPI (Boissière *et al.*, 2007) : avec seulement 2 000 mots de données d'adaptation, on relève au minimum (courriel) une amélioration du *KSR* d'environ 2 %. Chaque sous-corpus de test étant de taille réduite, le *KSR* suit des évolutions localement chaotiques. On observe toutefois une tendance générale à l'amélioration au fil de l'apprentissage, un plateau semblant être atteint à partir de 25 000 mots

d'apprentissage[1]. Enfin, le gain de performances est maximal pour le corpus de parole conversationnelle. Ce résultat est dû à la spécificité de la parole spontanée, mais également au fait que l'interaction est finalisée (tâche de renseignement touristique). La communication se focalise donc sur un vocabulaire restreint relevant d'un champ sémantique bien défini. Le paragraphe suivant étudiera précisément l'intérêt d'une adaptation au contexte sémantique.

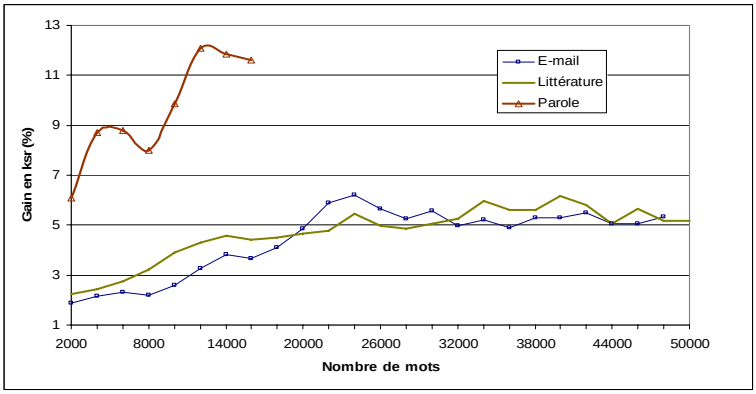

Figure 7. *Apprentissage incrémental de la prédiction de mots avec adaptation utilisateur : évolution du gain en KSR_5 sur trois corpus différents*

Pour le moment, notons que l'ajout d'un modèle utilisateur permet un gain significatif en *KSR*. Le tableau 3 et la figure 8 comparent les résultats globaux[2] avec ou sans adaptation utilisateur. Avec adaptation, le KSR_5 est supérieur à 50 % sur tous les corpus. Cette amélioration est particulièrement sensible lorsque la communication fait appel à un vocabulaire très spécifique (scientifique : + 8,2 % de *KSR*) et/ou à des structures linguistiques particulières (dialogue oral : + 9,4 %). L'adaptation permet même une amélioration, certes limitée, dans les cas les plus favorables (journal français : + 0,6 %). Pour l'allemand, l'apprentissage de certains mots composés récurrents permet des gains de performances substantiels (+ 3,0 %).

Modèle	**Journal** (D)	**Journal** (F)	**Sciences** (F)	**Prose** (F)	**Parole** (F)	**Courriel** (F)
4-gramme seul	51,6 %	57,9 %	44,2 %	46,0 %	48,3 %	48,6 %
4-gramme+MU	54,6 %	58,5 %	52,4 %	50,6 %	57,7 %	53,0 %
Gain KSR_5	*+ 3,0 %*	*+ 0,6 %*	*+ 8,2 %*	*+ 4,6 %*	*+ 9,4 %*	*+ 4,4 %*

Tableau 3. *Performances du 4-gramme interpolé avec un modèle utilisateur (MU)*

[1] Cette affirmation est à confirmer pour la parole, où le corpus de test est de taille limitée.

[2] Les gains globaux présentés sont inférieurs à ceux observés en figure 7, puisqu'ils incluent les tests sur le début des corpus, là où l'adaptation utilisateur n'est pas encore optimale.

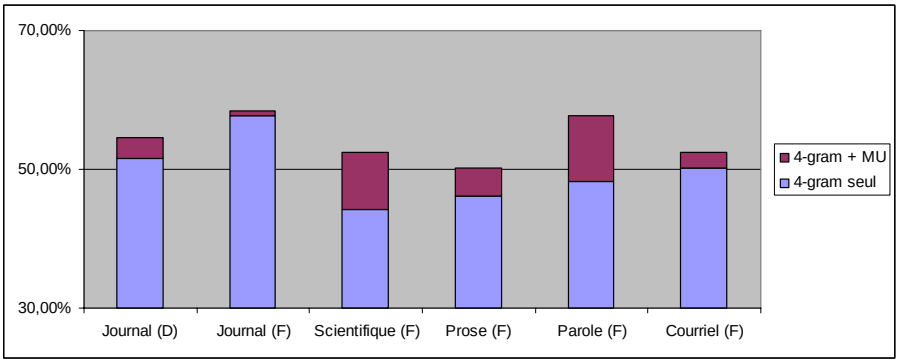

Figure 8. *Amélioration des performances du système Sibylle adaptation à l'utilisateur: comparaison du KSR avec ou sans modèle utilisateur*

En première approximation, l'adaptation à l'utilisateur semble d'autant plus bénéfique que le *KSR* initial est faible. Nous nous sommes demandé où cette adaptation jouait le plus : correspond-elle à une amélioration générale des performances, ou opère-t-elle plus fortement sur des textes où le *KSR* est faible ?

	Ensemble du corpus		**15 derniers sous-corpus**	
	Prose (F)	**Courriel** (F)	**Prose** (F)	**Courriel** (F)
4-gramme seul	σ = 1,44%	σ = 8,38%	σ = 1,95%	σ = 7,34%
4-gramme+MU	σ = 1,28%	σ = 5,05 %	σ = 0,78%	σ = 4,48%

Tableau 4. *Analyse distributionnelle du KSR_5 sur les corpus français : variance*

Le tableau 4 donne la variance de taux d'économie de saisie observé sur les corpus *Prose* et *Courriel*. Les résultats montrent que l'ajout du modèle utilisateur limite de manière significative la dispersion des performances. Le gain en *KSR* est donc modéré pour les passages déjà bien prédits et supérieur pour les textes difficiles. C'est sans doute cette capacité à limiter les situations où la prédiction se comporte mal qui fait que le modèle adaptatif est apprécié des utilisateurs.

5. Adaptation au contexte sémantique du discours : analyse sémantique latente

En dépit de sa rapidité d'entrée en action (2 000 mots d'apprentissage, soit tout de même une dizaine d'heures de saisie pour une personne handicapée), le modèle utilisateur ne permet qu'une adaptation à moyen et long terme. Une adaptation à court terme peut également être intéressante. En effet, lorsque la communication se focalise sur un thème, les mots qui appartiennent au champ sémantique du discours ont plus de chance d'apparaître (Lesher *et al.*, 2002).

Une première approche pour réaliser cette adaptation contextuelle serait de caractériser le thème du discours (Bigi *et al.*, 2001), puis de faire basculer la

prédiction sur un modèle spécifique à ce thème. L'adaptation thématique résultante restera toutefois assez grossière. Surtout, la détection automatique du thème du discours reste encore une question largement ouverte en terme de robustesse.

Une approche moins ambitieuse consiste à favoriser les mots qui ont été saisis récemment, en supposant qu'ils ont une probabilité plus forte de réapparaître. Cette hypothèse est particulièrement pertinente lorsque la communication est finalisée. Elle est à la base du modèle cache (Kuhn &De Mori, 1990). Dans le même esprit, le modèle *trigger* (Rosenfeld, 1996 ; Matiasek & Baroni, 2003) est un peu plus fin. Il fonde l'adaptation non pas sur des mots isolés mais sur des collocations : un mot déclencheur augmente (dès qu'il est utilisé) la probabilité des mots avec lesquels il est déjà apparu dans une certaine fenêtre de contexte. S'il a été montré que ces modèles ont une certaine capacité d'adaptation, les gains rapportés par la littérature restent assez limités, en particulier dans le domaine de l'aide à la saisie de texte.

Nous nous proposons d'étudier un nouveau modèle d'adaptation thématique, qui repose sur une évolution de l'analyse sémantique latente (LSA pour *Latent Semantic Analysis*). Le modèle LSA (Deerwester *et al.*, 1990), qui répond à des motivations cognitives (Landauer *et al.*, 1997), peut être vu comme une technique qui rend compte de relations lexicales sémantiques à partir de la distribution des cooccurrences de mots dans un contexte prédéfini. La LSA peut ainsi prédire des mots lexicaux à partir de termes qui lui sont reliés sémantiquement. Formellement, le modèle LSA constitue une évolution du modèle vectoriel (Salton & McGill, 1983). Dans un premier temps, un corpus de textes est représenté sous la forme d'une matrice (terme × contexte) où quelques milliers de termes d'indexation ont été préalablement choisis et où le contexte peut être aussi bien un document, un paragraphe, ou toute autre fenêtre contextuelle suivant l'application. Chaque élément de la matrice comptabilise la fréquence normalisée (*td.idf*, par exemple) d'occurrence d'un terme d'indexation dans le contexte considéré. L'apport de la LSA consiste à transposer la matrice obtenue dans un espace de dimension réduite. Pour cela, on décompose la matrice en valeurs singulières et on ne conserve que les k principales valeurs singulières obtenues (typiquement, k est compris entre 100 et 300). Chaque mot est représenté par un vecteur de k dimensions dans ce nouvel espace. Il est alors possible de comparer ces vecteurs par une mesure de distance ordinaire, le cosinus de l'angle entre les vecteurs étant le plus souvent utilisé. On montre que cette distance constitue, jusqu'à un certain point (Wandmacher, 2005), une bonne mesure de similarité sémantique.

Dans le cadre de la prédiction de mots, notre objectif est d'estimer la probabilité d'occurrence d'un mot en fonction du thème du discours. Dans l'espace LSA, la somme de plusieurs vecteurs représente la sémantique globale des éléments considérés. Nous faisons l'hypothèse que le champ sémantique correspondant au thème courant peut se décrire par un vecteur de contexte correspondant aux derniers mots déjà saisis :

$$\vec{h} = \sum_{i=1}^{m} \vec{w}_i \qquad \text{(c)}$$

où (w_1, … w_m) sont les mots du contexte courant.

On estime alors la (pseudo-) probabilité sémantique d'occurrence d'un mot à partir du cosinus de l'angle que fait sa représentation vectorielle avec le vecteur contexte :

$$\text{(d)} \qquad P_{LSA}(w_i|h) = \frac{\left(\cos(\vec{w}_i, \vec{h}) - \cos_{\min}(\vec{h})\right)^{\gamma}}{\sum_k \left(\cos(\vec{w}_k, \vec{h}) - \cos_{\min}(\vec{h})\right)^{\gamma}}$$

Le dénominateur de normalisation nous assure que la somme des probabilités est égale à 1. On observe que la distribution de ces pseudo-probabilités est très plate. Suivant (Coccaro et Jurafsky, 1998), nous leur appliquons donc un facteur de température γ pour augmenter le contraste entre les différentes prédictions.

6. Analyse sémantique latente et prédiction de mots : SIBYSEM

Notre modèle sémantique est ainsi supposé prédire des mots sémantiquement cohérents avec le contexte courant de discours. Considérons l'énoncé suivant :

(4) *Mon père était professeur en mathématiques et je pense que…*

Le tableau 5 donne la liste des dix mots qui, selon le modèle, sont les plus appropriés pour poursuivre la saisie. Ces propositions relèvent bien toutes des thématiques (*famille*, *enseignement*, *sciences*) initiées dans le discours.

Rang	Mot	P_{LSA}	Rang	Mot	P_{LSA}
1	*professeur*	0,0117	6	*père*	0,0046
2	*mathématiques*	0,0109	7	*mathématiques*	0,0045
3	*enseigné*	0,0083	8	*grand-père*	0,0043
4	*enseignait*	0,0053	9	*sciences*	0,0036
5	*mathématicien*	0,0049	10	*enseignant*	0,0032

Tableau 5. *Mots de probabilités maximales selon le modèle LSA pour l'énoncé (4)*

Cet exemple est également révélateur des limites d'une prédiction purement sémantique. Celle-ci relève d'une dimension paradigmatique du langage mais est incapable de prendre en compte le déroulement syntagmatique des énoncés : la LSA ne peut prédire l'apparition de mots grammaticaux, de même qu'elle ne saurait prédire la partie du discours d'un mot lexical à venir. C'est pourquoi il faut coupler cette prédiction sémantique avec un modèle de langage tel que celui présenté précédemment. L'intégration d'information sémantique, via la LSA, dans un modèle de langage a été déjà étudiée par (Bellegarda, 1997) et (Coccaro & Jurafsky, 1998). Dans SIBYLLE, nous la réalisons par interpolation géométrique du modèle sémantique et du modèle n-gramme adaptatif :

$$\text{(e)} \qquad P_{global}(w_i) = \frac{P_{base}(w_i)^{\lambda_1} \cdot P_{LSA}(w_i)^{(1-\lambda_1)}}{\sum_{j=1}^{n} P_{base}(w_j)^{\lambda_1} \cdot P_{LSA}(w_j)^{(1-\lambda_1)}}$$

avec *n*, nombre de termes du vocabulaire. L'interpolation géométrique impose que les deux modèles soient en accord pour assigner une probabilité élevée à un mot.

Nos études (Wandmacher & Antoine, 2007) ont montré la supériorité de cette interpolation dans une tâche de prédiction (gain en *KSR* de 0,6 %).

Cependant, l'efficacité de la prédiction sémantique dépend fortement du mot considéré : la LSA est bien capable de prédire des mots spécifiques de contenu, mais pas des mots de fonction (Wandmacher, 2005). C'est pourquoi certains travaux ont proposé d'adapter le coefficient d'interpolation en fonction de la probabilité sémantique obtenue. (Coccaro & Jurafsky, 1998) ont ainsi mis en place une mesure de confiance sur les probabilités LSA basée sur l'entropie, en partant du principe que les mots qui apparaissent dans des contextes variés (donc qui ont une entropie élevée) ne peuvent pas être correctement prédits par la LSA. (Wandmacher, 2005) n'a pas observé de corrélation significative liée à l'entropie dans ses études expérimentales. Il a, par contre, montré qu'il existait une certaine corrélation entre le nombre de termes sémantiquement reliés à un mot et la densité de ses plus proches voisins dans l'espace LSA. Cette observation est assez intuitive : plus les voisins d'un terme sont proches, plus on a de chance que ceux-ci lui soient réellement liés sémantiquement. On définit comme suit la densité des voisins d'un terme w_i :

$$D_m(w_i) = \frac{1}{m} \cdot \sum_{j=1}^{m} \cos(\vec{w}_i, NN_j(\vec{w}_i)) \qquad \text{(f)}$$

où les $NN_j(\vec{w}_i)$ sont les vecteurs w_j les plus proches de w_i. La densité est utilisée pour adapter le facteur d'interpolation du modèle global suivant la formule :

$$\lambda_i = \beta \cdot D(w_i) \text{ si } D(w_i) > 0 \text{ et } \lambda_i = 0 \qquad \text{(g)}$$

sinon, avec β constante de pondération contrôlant l'influence de la LSA. Après étude expérimentale, nous avons fixé ce facteur à une valeur optimale de 0,4. Cela signifie donc que l'influence du modèle LSA sur le prédicteur global est comprise entre 0 et 40 %. Le module de prédiction général obtenu est appelé SIBYSEM.

Afin d'évaluer l'apport de la LSA pour une tâche de prédiction générale, nous avons tout d'abord testé les capacités intrinsèques de la version française du modèle sur le corpus de test journalistique (*L'Humanité*). SIBYSEM a été entraîné sur sept années du journal *Le Monde* (1996-2002, 101 millions de mots) à l'aide du toolkit *InfoMap*[3]. Nous avons utilisé 3 000 termes d'indexation et un vocabulaire contrôlé de 80 000 mots lexicaux. Après plusieurs tests, le contexte de cooccurrence qui a été retenu a été une fenêtre de ±100 mots. Nous n'utilisons donc pas la notion de document ou de paragraphe pour l'apprentissage. Nos études nous ont également conduits à choisir une valeur de k = 150 dimensions pour la réduction par décomposition en valeurs singulières. Ce modèle a été comparé à trois autres :

– *base :* le modèle 4-gramme avec modèle utilisateur précédent (Base) ;
– *cache :* 4-gramme couplé à un modèle cache avec fonction de décroissance temporelle (Clarkson & Robinson, 1997) et fenêtre du cache de 400 mots ;

[3] Infomap Project: http://infomap-nlp.sourceforge.net

– *LSA_IG :* 4-gramme interpolé géométriquement avec le modèle LSA mais sans pondération du coefficient d'interpolation par une mesure de confiance.

Le tableau 6 compare le *KSR* et la perplexité des différents modèles. Les améliorations d'économie de saisie que l'on obtient avec la LSA peuvent sembler assez limitées. Elles sont cependant statistiquement significatives (seuil de significativité $p < 0{,}001$) et il faut comprendre qu'un KSR_5 de base de 57,87 % constitue déjà une excellente performance. On observera en outre que l'adaptation contextuelle due à la LSA est très nettement supérieure à celle du modèle cache. Enfin, la pondération dynamique de l'influence de la LSA permet d'atteindre des gains supplémentaires par rapport à une interpolation géométrique standard.

Modèle	**Base** 4-gramme + MU	**Cache** Base + cache	**LSA_IG** Base + LSA simple	**SIBYSEM** Base + LSA pondéré
KSR_5	57,87 %	58,00 %	58,61 %	58,92 %
Gain KSR	-	*+ 0,13 %*	*+ 0,74 %*	*+ 1,05 %*
Perplexité	109,7	106,1	99,5	99,5
Réduction	-	*- 3,6*	*- 10,2*	*- 10,2*

Tableau 6. *KSR_5 des différents modèles adaptatifs sur le corpus «* L'Humanité *»*

Ces résultats témoignent des capacités de la LSA à améliorer les performances d'un module stochastique de langage, et ce quel que soit le cadre applicatif visé. Si l'on considère maintenant la communication assistée, SIBYSEM a été également testé sur les corpus de tests « écologiques » rencontrés précédemment. Le tableau 7 présente ces résultats pour quatre modèles différents :

– modèle 4-gramme seul (base) ;
– modèle 4-gramme + modèle utilisateur (rappel de résultats déjà présentés) ;
– modèle 4-gramme + LSA mais sans modèle utilisateur ;
– SIBYSEM : modèle 4-gramme + modèle utilisateur + LSA avec pondération.

Modèle	**Journal** (D)	**Journal** (F)	**Sciences** (F)	**Prose** (F)	**Parole** (F)	**Courriel** (F)
Base	51,6 %	57,9 %	44,2 %	46,0 %	48,3 %	48,6 %
Base + MU	54,6 %	**58,5 %**	52,4 %	50,6 %	57,7 %	53,0 %
Base + LSA	52,6 %	**58,9 %**	45,6 %	47,7 %	49,9 %	50,2 %
SIBYSEM	55,4 %	59,4 %	52,9 %	52,0 %	58,8 %	53,7 %
Gain en KSR	*+ 4,2 %*	*+ 1,5 %*	*+ 8,7 %*	*+ 6,0 %*	*+10,5 %*	*+ 5,1 %*

Tableau 7. *Performances (KSR_5) des différents modèles de prédiction*

Globalement, la LSA permet de nouveaux gains dans toutes les situations : son apport est donc complémentaire de celui du modèle utilisateur. Les gains cumulés permettent d'atteindre des valeurs de *KSR* très satisfaisantes (entre 52 % et 59 % suivant le genre testé). En l'absence de données de référence en français, il est

difficile de comparer ces résultats avec d'autres. VITIPI (Boissière *et al.*, 2006) présente un KSR_5 de l'ordre de 35 % sur un corpus de courriels tandis que le système PCA atteint un KSR_5 de 49 % sur un corpus de type récit biographique (Blache & Rauzy, 2007). Enfin, la version française du système FASTY approche les 50 % de KSR_5 sur du texte journalistique (Beck *et al*., 2004).

Si l'on s'intéresse à l'apport de la LSA seule, on remarque que les gains obtenus sont toujours inférieurs à ceux du modèle utilisateur sauf sur le corpus journalistique (en gras dans le tableau 6). Cette observation situe bien les domaines d'intervention de deux modèles d'adaptation :

– le modèle utilisateur sert à rapprocher les données d'apprentissage du langage réellement produit par l'utilisateur. Son apport est donc d'autant plus important que le registre de langue considéré diffère des données d'apprentissage. Lorsque ce n'est pas le cas, son apport est au contraire assez limité[4] ;

– le modèle sémantique LSA permet une adaptation dynamique au contexte du discours. Il n'autorise que des gains limités en *KSR*, mais ceux-ci sont constants dans toutes les situations et deviennent prédominants lorsque le modèle de base est appris sur des données suffisamment représentatives de l'usage réel.

7. Un problème particulier : analyse partielle des mots composés en allemand

SIBYSEM constitue donc un modèle adaptatif intéressant pour la prédiction de mots. Certaines spécificités linguistiques demandent cependant à être étudiées de plus près pour disposer d'un modèle réellement générique. Dans ce paragraphe, nous allons nous intéresser au problème de la gestion des mots composés en allemand. Cette étude concerne d'une manière générale les langues agglutinantes ou à forte capacité de dérivation compositionnelle.

L'allemand est une langue assez difficile pour le TAL, en général, et pour la prédiction de mots, en particulier. Il a une riche morphologie avec trois genres (masculin/féminin/neutre) et quatre cas (nominatif/génitif/accusatif/datif), ce qui multiplie les formes fléchies des termes. Par ailleurs, l'ordre des mots dans la phrase est très libre, les possibilités de continuer une phrase à partir d'un contexte gauche donné sont donc nombreuses. L'exemple suivant illustre cette variabilité :

(5) *Ich habe ihm das gestern schon gesagt.*
Ich habe das ihm gestern schon gesagt.
Das habe ich ihm gestern schon gesagt.
Gestern habe ich ihm das schon gesagt.
Schon gestern habe ich ihm das gesagt.
Gesagt habe ich ihm das schon gestern.

Translittération en français : *Je lui ai dit cela déjà hier*.

[4] Ce résultat est moins frappant sur le corpus journalistique allemand, le modèle utilisateur servant par défaut de gestionnaire d'adaptation des mots composés complexes.

Cependant, le principal problème que pose l'allemand réside dans sa capacité à composer en une seule unité graphématique des termes complexes parfois longs. Cette caractéristique se retrouve dans d'autres langues germaniques (néerlandais, suédois). La création de mots composés est très productive. Elle peut faire intervenir n'importe quel nom commun, ainsi que quelques verbes et adjectifs :

- *Hundenase* (« nez du chien »)
- *Vereinssitzung* (« réunion de l'association »)
- *Schiffshebewerksdirektor* (« directeur à la centrale de levage de bateaux »)
- *Wortprädiktionskomponente* (« module de prédiction de mots »)

Analysant un corpus journalistique (*APA newswire corpus*) allemand (Baroni *et al*., 2002b) rapportent que presque la moitié (47 %) des mots lexicaux rencontrés étaient des mots composés. La plupart de ces mots étaient par contre extrêmement rares, la majorité n'apparaissant qu'une seule fois dans le corpus (hapax). Comme ces mots n'apparaissent dans aucun vocabulaire prédéfini, ils ne sont pas prédictibles. En plus, comme ils sont plutôt longs, leur impact négatif sera assez important sur le *KSR* si on ne les prédit par correctement. Dans le paragraphe 4, nous avons ainsi montré que le *KSR* pour l'allemand était sensiblement plus basse (– 6,1 %) que pour le français sur un corpus de test comparable. Si la variabilité d'ordonnancement linéaire reste un problème pour les modèles de langage (Antoine & Goulian, 2001), on peut espérer traiter les mots composés dans un système de prédiction de mots. (Baroni *et al.*, 2002a, 2002b) proposent un modèle de prédiction (*Split compound model*) qui découpe un mot composé en une tête lexicale et son modificateur en se basant sur une description linguistique détaillée, le modèle ne traite que des mots de type *Nom* + *Nom*, type le plus fréquent en allemand. Par exemple, *Polizeikontrolle* (« contrôle de police ») est analysé en *Kontrolle* (tête) + *Polizei* (modificateur). Les gains de cette analyse morphologique complexe à base de connaissances restent cependant modestes. (Trost *et al*., 2005) rapportent ainsi une amélioration de + 0,3 % du *KSR*.

À l'opposé, nous avons implémenté une stratégie simple mais efficace : la *sélection partielle* (SP) permet de choisir dans la liste de prédiction les parties d'un mot composé et de les agglutiner en entrant la touche *retour* après chaque partie. Les mots peuvent ainsi être complétés constituant par constituant, la prédiction n'étant plus limitée aux mots simples. Sans analyse morphologique, cette approche ne peut toutefois traiter les morphèmes de jointure, qui apparaissent dans certains mots composés (*Hund-**e**-nase*, chien'-***e***-,nez', *Verein-**s**-sitzung*, Association'-***s***-,réunion'). Notre système permet d'ajouter un morphème de jointure ('*-s-*', '*-e-*', '*-en-*', '*-es-*', '*-er-*') dès qu'une partie du mot a été sélectionnée.

	KSR_5 (sans SP)	KSR_5 (avec SP)	Grain KSR_5
4-gramme seul	51,56 %	53,05 %	+1,50 %
4-gramme + PU	54,57 %	56,06 %	+1,49 %
4-gramme + LSA	52,56 %	54,05 %	+1,49 %

Tableau 8. *Performances de la prédiction allemande avec sélection partielle*

Le tableau 8 montre les résultats sans et avec la sélection partielle (SP) pour le corpus journalistique (*Süddeutsche Zeitung*). La sélection partielle permet un gain d'environ + 1,5 % en *KSR* pour chaque modèle considéré. (Trost *et al*., 2005) obtiennent des résultats aussi significatifs (+ 3 %) avec une stratégie équivalente. Même si le modèle de sélection partielle est assez simple, son intérêt est indéniable et sans comparaison avec les approches à base d'analyse morphologique complexe.

8. Prédiction de mots et communication assistée : vers les usages réels

La prédiction de mots peut concerner de nombreuses applications : aide à la communication pour personnes handicapées, communication assistée sur dispositifs à interfaces limitées (organiseurs personnels, téléphones mobiles, GPS) ou encore aide à la traduction automatique. Les tests que nous avons présentés jusqu'ici on cherché à montrer la pertinence générale de notre modèle. Si l'on s'intéresse à une application donnée, l'efficacité de la prédiction dépend néanmoins de son adéquation avec l'interface utilisateur réalisée. Nous n'allons pas développer ici les aspects IHM de cette recherche[5]. Nous nous attarderons, par contre, sur plusieurs expériences qui montrent quels peuvent être les facteurs d'adaptation de la prédiction à l'interface utilisateur, et leurs conséquences sur l'économie de saisie.

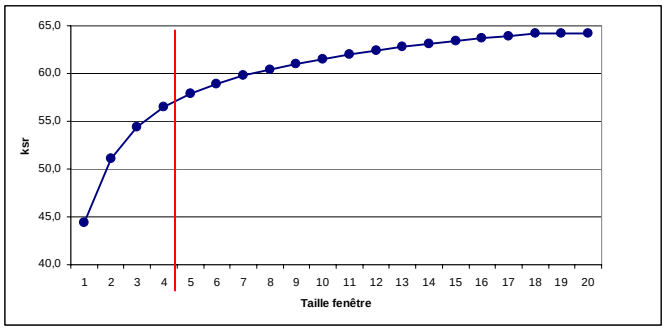

Figure 9. *KSR de SIBYSEM en fonction de la taille de la liste de mots affichés*

Une première expérience concerne la taille de la liste de prédiction présentée à l'utilisateur. Plus cette liste est importante, plus on a une chance d'y trouver le mot recherché. En contrepartie, une liste longue occupe plus d'espace sur l'écran, et son parcours visuel peut s'avérer fatiguant. L'économie de saisie est généralement mesurée avec une liste de cinq mots. L'interface de SIBYLLE comporte désormais sept hypothèses de mots. Afin de chercher une taille optimale, nous avons, quant à nous, étudié l'évolution du *KSR* en fonction du nombre d'hypothèses présentées. Cette expérience a été menée avec le moteur de prédiction SIBYSEM complet sur le corpus *L'Humanité*. Comme le montre la figure 9, le gain marginal que l'on peut espérer gagner en augmentant la taille de la liste de prédiction décroît rapidement avec celle-ci. Une liste de cinq mots constitue un choix d'affichage assez optimal puisque c'est vers cette taille que l'aplatissement de la courbe du *KSR* commence à

[5] Sur ce sujet, consulter nos travaux communs avec le VALORIA (Barhoumi *et al*., 2007).

se manifester. Cette étude montre également qu'on atteint une économie de saisie appréciable avec une liste de un ou deux mots. Cette observation ouvre la porte à des stratégies évitant l'appui nécessaire au changement de clavier :

– complétion intégrée : la meilleure hypothèse est affichée comme unique proposition directement en complétion dans le texte (Boissière & Dours, 2001)

– intégration d'une ou deux hypothèses lexicales au début du clavier de lettres dynamique. Nous réfléchissons à une telle approche pour SIBYLLE, le nombre de mots intégrés pouvant dépendre d'un seuil de probabilité.

Une autre expérience illustre l'interdépendance du moteur de prédiction avec l'interface utilisateur. Comme nous l'avons dit (§ 3.1.), le système SIBYLLE filtre les hypothèses lexicales qui n'ont pas été sélectionnées à un moment donné. Cette approche suppose que l'utilisateur parcourt toujours visuellement l'intégralité de la liste pour sélectionner toute proposition correcte. Or, les études que nous avons menées avec des patients Infirmes Moteurs Cérébraux (IMC) du centre de Kerpape montrent que cette stratégie n'est jamais suivie de manière systématique. Aussi avons-nous étudié l'influence du filtrage sur l'économie de saisie. Cette étude a été menée sur le corpus de test *L'Humanité* avec une liste de cinq mots de prédiction.

	KSR_5 (sans filtrage)	**KSR_5 (avec filtrage)**	**Grain KSR_5**
4-gramme seul	56,90 %	57,87 %	+ 0,97 %
4-gramme+LSA	58,01 %	58,92 %	+ 0,91 %

Tableau 9. *Influence du filtrage sur l'économie de saisie (corpus Humanité)*

Le tableau 9 résume les résultats obtenus avec un 4-gramme couplé ou non avec l'adaptation sémantique. Dans les deux cas, la stratégie de filtrage permet des gains comparables en matière de *KSR* (≈ 0,9 %). Ce gain n'est pas négligeable. Il est au contraire équivalent à celui qu'apporte l'ajout du module LSA. Cela montre, là encore, l'importance de l'interface utilisateur dans l'aide apportée par un communicateur. En revanche, ces résultats ne nous permettent pas de choisir une stratégie optimale d'affichage. Comme toujours dans le monde du handicap, tout dépendra du profil de l'utilisateur. Le choix de la stratégie la plus adaptée pour une personne donnée résulte d'un compromis entre le gain de *KSR* théorique et l'usage effectif que fait l'utilisateur de la liste de mots. Un module de trace (inactif par défaut) a été précisément ajouté à SIBYLLE pour caractériser ces comportements.

9. Utilisation du système SIBYLLE : retour d'expérience et perspectives.

Au fil de sept ans d'utilisation quotidienne au centre de Kerpape, plusieurs dizaines de patients IMC ont pu découvrir SIBYLLE et l'ont conservé comme outil d'aide à la communication. De nombreux enfants en séjour long utilisent SIBYLLE pour leurs activités pédagogiques au sein de l'école adaptée du centre. SIBYLLE est ainsi un outil précieux dans les activités d'apprentissage de la langue mais également en ergothérapie. Les enseignants ont constaté que les enfants composent plus de textes et font moins de fautes d'orthographe lorsqu'ils sont équipés du système. On peut donc affirmer que SIBYLLE, dont l'interface a été conçue avec

Jean-Paul Départe (laboratoire d'informatique de Kerpape), répond aux besoins des utilisateurs. La phase d'apprentissage à cette nouvelle aide est simple et rapide, même pour des personnes lourdement handicapées. Le principe du saut de touche, du saut de clavier, voire du clic long pour ceux qui le maîtrisent, n'ont pas posé de problème particulier. Du point de vue des utilisateurs, SIBYLLE est essentiellement apprécié en termes de « confort », le défilement linéaire et son unique validation retenant leur attention. L'aspect dynamique du clavier de lettres n'a jamais posé de problème, contrairement à nos craintes initiales. Les seules personnes qui ne l'ont pas accepté souffraient en effet de problèmes de poursuite oculaire.

En dépit de ces retours positifs, des pistes d'améliorations restent à explorer. Les performances élevées de SIBYSEM font qu'on ne peut s'attendre à des gains substantiels de *KSR* au niveau linguistique. Nos observations au centre de Kerpape, confirmées par d'autres études (Biard *et al*., 2006), montrent cependant que les patients ignorent souvent les propositions de la prédiction. L'économie de saisie réelle est donc éloignée du potentiel maximal estimé par le *KSR*. C'est pourquoi les améliorations du système résulteront d'une étude approfondie des interactions entre la prédiction et l'interface utilisateur. Nous avons commencé à étudier cette question (taille de la liste, filtrage). Ces recherches doivent être approfondies dans une approche pluridisciplinaire associant TAL, Interaction Homme-Machine et sciences cognitives. Elles permettraient d'avoir une meilleure connaissance des usages et du comportement réel des utilisateurs. C'est l'objet du projet ESAC_IMC, financé par la Fondation Motrice, qui vise le recueil de corpus de traces d'exécution de différents systèmes[6] et leur mise en relation avec le tableau clinique des sujets concernés. Un format commun de fichier de logs a été défini qui permet une interopérabilité entre les données recueillies. Ces corpus permettent de rejouer ou simuler les sessions d'utilisation en faisant varier certains paramètres des systèmes. À terme, il sera ainsi possible de tester le comportement du moteur de prédiction SIBYSEM avec une interface utilisateur développée par l'IRIT.

Une autre piste prioritaire relève spécifiquement du TALN. Elle concerne la prédiction de mots pour des patients présentant, en plus de leur infirmité motrice, d'autres troubles cognitifs (retard d'apprentissage de la langue, dyslexie, aphasie…). S'il semble difficile d'améliorer significativement notre prédiction de mots sur des énoncés bien formés, cette question redevient d'actualité pour des utilisateurs souffrant de troubles langagiers associés. Elle n'a malheureusement été que rarement étudiée par les chercheurs du domaine. On sait pourtant que la prévalence de ces troubles langagiers est loin d'être négligeable. Voici un énoncé composé par un jeune adulte IMC ayant une maîtrise imparfaite de la langue :

(6) *tu a vu quil ya des famille qui porte plinte acose l'handicap de leur fils . j'ai pas tous de suite compris les tenen les abouticen mais papa ma espliqué. je pences si on fait tous ça les génécologue serai derier les baros souven.*

Translittération : *tu as vu qu'il y a des familles qui portent plainte à cause (de) l'handicap de leur fils. Je n'ai pas tout de suite compris les tenants et les*

[6] Participants : laboratoires IRIT, LI, VALORIA, Jacques Lordat et le centre de Kerpape.

aboutissants, mais papa m'a expliqué. Je pense (que) si on fait tous ça, les gynécologues seraient derrière les barreaux souvent

Cet exemple est représentatif des difficultés qui se posent à la prédiction : fautes d'orthographe, mais surtout écriture phonétique (*pences*, *baros*, *abouticen*) et regroupement incorrect de graphèmes. Nous avons dit plus haut que la liste de mots prédits représente une aide orthographique appréciable pour l'utilisateur, et que les personnes IMC faisaient moins de fautes lorsqu'elles utilisaient SIBYLLE. Il faut toutefois noter que l'énoncé (6), très perturbé, a été composé avec ce système ! L'utilisateur n'a pas sélectionné le mot recherché dans la liste de prédiction, alors qu'il était souvent présent (*plainte*, *pense*, *famille*, *aboutissants*, *gynécologue*…) Il est donc illusoire de croire qu'une prédiction de mots, même parfaite, influencera suffisamment l'utilisateur pour lui permettre de composer des énoncés sans fautes. Nos systèmes doivent donc gérer les principales erreurs présentes dans ces énoncés.

Corpus	**Description**	**Nb. Mots**
Courrier	courriers rédigés par des IMC de 12 à 20 avec SIBYLLE. Motricité Fonctionnelle Globale : niveau V (Palisano *et al*., 1997) ; élocuction : anarthrie ou dysarthrie sévère.	1416
Parole	énoncés rédigés sur SIBYLLE par des IMC lors de groupes de discussion. Tableaux cliniques inconnus.	855

Tableau 10. *Corpus francophones de communication assistée recueillis à Kerpape*

Afin d'estimer la dégradation résultante des performances, nous avons étudié deux corpus de productions d'adolescents et de jeunes adultes IMC du centre de Kerpape (tableau 10). Enregistrés dans le cadre du projet ESAC_IMC avec SIBYLLE, ces corpus sont de taille restreinte du fait de la lenteur de composition des messages. Ils constituent toutefois une des premières tentatives d'étude des usages langagiers réels de personnes IMC avec ou sans troubles langagiers associés.

Modèle / Corpus	**Courrier**	**Parole**	**Ensemble**
***KSR*$_5$ 4-gramme seul**	31,8 %	50,9 %	38,9 %
***KSR*$_5$ 4-gramme + MU**	36,6 %	59,6 %	44,8 %
***KSR*$_5$ 4-gramme + MU + LSA (SIBYSEM)**	37,2 %	60,2 %	45,1 %
% mots hors vocabulaire	20,6 %	3,0 %	14,2 %

Tableau 11. *KSR_5 sur des corpus d'énoncé composés par des patients IMC*

Les résultats obtenus sur ces corpus sont synthétisés dans le tableau 11. Ils sont contrastés : on observe une dégradation très sensible des performances sur le corpus *courrier* (*KSR*$_5$ de 37,2 % pour SIBYSEM), alors que celles-ci restent satisfaisantes sur le corpus *parole*. En l'absence de tableau clinique pour certains sujets, seule l'étude de leurs productions langagière peut nous aider à expliquer ces différences :

– le corpus *parole* se caractérise par de forts agrammatismes : les verbes n'y sont ainsi que rarement conjugués (exemple : *pourquoi je être handicapé*), mais les mots sont le plus souvent orthographiés correctement. Au final, la proportion de mots hors du vocabulaire de l'application est limitée (3 %) ;

– le corpus *courrier*, dont est extrait l'énoncé (6), témoigne au contraire d'une bonne maîtrise des structures de la langue mais d'une écriture qui est purement phonétique, avec de nombreux cas d'agglutinations. D'où un taux très élevé de mots hors vocabulaire (20,6 %).

La prédiction semble donc avant tout affectée par les mots hors vocabulaire, ce qui suggère que les pistes majeures d'amélioration se situent au niveau du lexique.

Par ailleurs, l'adaptation utilisateur ou la LSA opèrent toujours sur ces énoncés perturbés (gain en KSR_5 de 5,4 % à 9,3 %). En particulier, les erreurs commises répondent à des régularités qui sont apprises par le modèle utilisateur. Cette capacité d'apprentissage pose une question sur les objectifs de la communication assistée. Doit-elle s'adapter au maximum aux productions de l'utilisateur afin de faciliter la communication ? Ou doit-elle viser en priorité sa rééducation langagière en ne proposant que des hypothèses linguistiquement correctes (Maurel *et al*., 2000) ? Nos discussions avec les thérapeutes et les enseignants de Kerpape montrent que ce débat ne peut être tranché : chaque contexte d'utilisation répond en effet à des objectifs différents. C'est pourquoi nous nous orientons vers la réalisation d'un système mixte pouvant, suivant le paramétrage choisi, être utilisé en rééducation ou en aide à la communication. Selon l'objectif visé, les techniques mises en jeu différeront : méthodes de correction (voir par exemple celles du système VITPI) ou au contraire tolérance aux erreurs. Dans ce second cas, nous réfléchissons à adapter à la problématique du handicap des techniques de modélisation utilisées en reconnaissance de la parole (problèmes de segmentation) ou sur le traitement des SMS (écriture phonétique) et des phrases dysorthographiées (Sitbon *et al*., 2007).

10. Bibliographie

Antoine J.-Y., Letellier-Zarshenas S., Schadle I., Nicolas P., Caelen J. « Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement ». Actes *TALN'2002*, Nancy, France. p. 319-324. 2002.

Baroni M., Matiasek J., Trost H. « Wordform- and class-based prediction of the components of German nominal compounds in an AAC system », Actes *19th COLING*, 2002.

Baroni, M. and Matiasek, J. and Trost, H. « Predicting the Components of German Nominal Compounds » Actes *ECAI 2002*, IOS Press, Amsterdam, p.470-474, 2002.

Barhoumi Z., Poirier F., Antoine J.-Y. « Sibylle : Considérations ergonomiques et adaptation aux besoins des utilisateurs ». Soumis à *ASSISTH'07*, Toulouse, France.

Beck C., Seisenbacher G., Edelmayer G., Zagler W.L. « First user test results with the predictive typing system FASTY ». *Actes ICCHP'04*. Paris. *LNCS* 3118. Springer. 2004.

Bellegarda J. « A Latent Semantic Analysis framework for large-span language modeling », Actes *Eurospeech 97*, Rhodes, Grèce. 1997.

Biard N., Dumas C., Bouteille J., Pozzi D., Lofaso F., Laffont I. « Apports de l'évaluation en situation de vie à partir d'une étude sur l'intérêt de la prédiction de mots auprès d'utilisateurs de synthèse vocale ». Actes *Handicap 2006*, Paris. p. 145-148. 2006.

Bigi, B.; Brun, A.; Haton, J.; Smaili, K. & Zitouni, I. « Dynamic Topic Identification: Towards Combination of Methods », Actes *Recent Advances in Natural Language Processing workshop, RANLP'2001*, p. 255-257. 2001.

Blache Ph., Rauzy S. « Le moteur de prédiction de mots de la Plateforme de Communication Alternative », *Traitement Automatique des Langues*, *TAL*, vol. 48 n° 3. 2007 (à paraître)

Boissière P., Dours D. « VITIPI : Comment un système d'assistance à l'écriture pour les personnes handicapées peut offrir des propriétés intéressantes pour le TALN ? » Actes *TALN'2001*, *atelier TALN et Handicap*, Tours. Vol. 2, p. 183-192. 2001.

Boissiere Ph., Schadle I., Antoine J.-Y. « A methodological framework for writing assistance systems: applications to sibylle and VITIPI systems », *AMSE Journal on Modelling, Mesurement & Control*, *Série C.*, Barcelone, Espagne, Vol 67, p. 167-176. 2006.

Cantegrit B., Toulotte J.-M. « Réflexions sur l'aide à la communication des personnes présentant un handicap moteur ». Actes *TALN'2001*, *atelier Ingénierie des Langues et Handicap*, Tours, France, juillet 2001. Vol. 2, p. 193-202.

Clarkson, P. R., Robinson, A.J. « Language Model Adaptation using Mixtures and an Exponentially Decaying Cache », Actes *IEEE ICASSP-97*, Munich, Allemagne. 1997.

Coccaro, N. and Jurafsky, D. « Towards better integration of semantic predictors in statistical language modelling », Actes *ICSLP-98*, Sydney. 1998.

Deerwester, S. C., Dumais, S., Landauer, T., Furnas, G. and Harshman, R. « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, *JASIS* 41(6), p. 391-407. 1990.

Goodman J. « A Bit of Progress in Language Modeling », Extended Version, *Microsoft Research Technical Report* MSR-TR-2001-72. 2001.

Jelinek, F. « Self-organized Language Models for Speech Recognition ». In Waibel A., Lee K.-F, *Readings in Speech Recognition*, Morgan Kaufman Publ. p. 450-506. 1990.

Kuhn, R. and De Mori, R. « A Cache-Based Natural Language Model for Speech Reproduction » *IEEE Trans. PAMI*, 12 (6), p. 570-583. 1990.

Landauer, T. K., Laham, D., Rehder, B. and Schreiner, M. E. « How well can passage meaning be derived without using word order? A comparison of LSA and humans ». Actes *19th meeting of the Cognitive Science Society*. Mawhwah, NJ. p. 412-417, 1997.

Lesher, G. W., Moulton, B. J, Higginbotham, D. and Alsofrom, B. « Limits of human word prediction performance. » Actes *CSUN 2002*, California State U., Northridge. 2002.

McCoy K.F., Demasco P. « Some applications of natural language processing to the field of augmentative and alternative communication » Actes *IJCAI'95 Workshop on Developing AI Applications for Disabled People*, Montreal, Canada. p. 97-112. 1995.

Matiasek, H. and Baroni, M. « Exploiting long distance collocational relations in predictive typing » Actes *EACL'03 Workshop on Language Modeling for Text Entry Methods*, Budapest. 2003.

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

Palisano R. Rosenbaum P., Walter S., Russel D., Wood E., Galuppi B. « Development and relaibility of a system to classify gross motor function in children with cerebral palsy ». *Dev. Med. Child Neurol*., 39, p. 214-223. 1997.

Rosenfeld, R. « A maximum entropy approach to adaptive statistical language modelling ». Actes *Computer Speech and Language*, 10 (1), p. 187-228. 1996.

Ricco, X., Dutoit T. « Vers un logiciel multilingue et gratuit pour l'aide aux personnes handicapées de la parole : HOOK ». Actes *TALN'2001*, *atelier Ingénierie des Langues et Handicap*, Tours, France, juillet 2001, Vol 2, p. 223-232.

Salton G., McGill M. « Introduction to modern information retrieval » McGraw-Hill, New-York, NJ. 1983.

Schadle I., Antoine J.-Y., Le Pévedic B., Poirier F. « Sibylettre : prédiction de lettre pour la communication assistée, Revue d'Interaction Homme-Machine ». *Revue d'Interaction Homme-Machine, RIHM*, vol. 3, n° 2. 2002.

Shannon, C. E. « Prediction and Entropy of Printed English », *Bell System Technical Journal*, p. 50-64, 1951.

Sitbon L., Bellot P., Blache Ph. « traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées » Actes *TALN'2007*, Toulouse. Vol. 1, p. 263-272. 2007.

Stolcke, A. « Entropy-based pruning of backoff language models ». Actes DARPA Broadcast News Transcription and Understanding Workshop, 1998.

Stolcke, A. « SRILM - An Extensible Language Modeling Toolkit», Actes *International. Conference on Spoken Language Processing, ICSLP'02*, Denver, Colorado, 2002.

Trnka K., Yarington D., McCoy K.F., Pennington C. « Topic Modeling in Fringe word prediction for AAC » Actes *International Conference on Intelligent User Interfaces*, Sydney, Australie. p. 276-278. 2006.

Trost H., Matiasek J., Baroni M., « The language component of the FASTY text prediction system » *Applied Artificial Intelligence*. 19(8). P. 743-781. 2005.

Vella F., Vigouroux N., « Disposition spatiale des touches/caractères des claviers logiciels et fatigue motrice: résultats expérimentaux », Actes *Handicap'2006*, Paris. 2006.

Wandmacher T. « How semantic is Latent Semantic Analysis », Actes *RECITAL'2005*, Dourdan, France. 2005.

Wandmacher T., Antoine J.-Y. « Methods to integrate a language model with semantic information for a word prediction component » Actes *EMNLP-CoNLL'07*, Prague, Tchéquie. 506-513. Juin 2007.