La prédiction lexicale et syntaxique à partir du contexte gauche : Application au système *HandiAS*

Denis MAUREL, Brigitte LE PÉVÉDIC et Jonathan YAVCHITZ

Abstract. We develop a method for an adaptable morphosyntaxic prediction from the left context. This predictive method uses an hybrid technic combining symbolic and statistic. The aim is to predict the end of a word after the first input letters. We propose to develop a communication-aid software for disabled people, called *HandiAS*.

Résumé. Nous présentons dans cet article une méthode adaptable de prévision morphosyntaxique à partir du contexte gauche à partir d'une technique de prédiction hybride, à la fois symbolique et statistique. Notre but est de prédire la fin d'un mot à partir de la saisie de ses premières lettres. Nous illustrons notre propos par la présentation d'un prototype de logiciel d'aide à la communication pour des personnes handicapées physiques, logiciel nommé *HandiAS*.

Keywords: Acceptability, stochastic automata, electronic dictionnary, frequency, morphology, syntax.

Mots-clés : Acceptabilité, automates stochastiques, dictionnaire électronique, évolution, fréquence, morphologie, syntaxe, techniques hybrides.

1. Introduction

Peut-on prédire un mot en connaissant seulement son contexte gauche ? L'exercice, qui semble *a priori* quasiment impossible, devient plus abordable si l'on autorise la connaissance de quelques-unes des

✉ Denis MAUREL et Jonathan YAVCHITZ ; LI/E3i ; Université François Rabelais ; 64,avenue Jean-Portalis ; F–37200 Tours (France).
E-mail : maurel@univ-tours.fr

✉ Brigitte LE PÉVÉDIC ; IRIN ; Université de Nantes ; 2, rue de la Houssinière ; F–44072 Nantes cedex 03 (France).
E-mail : brigitte.lepevedic@irin.univ-nantes.fr

premières lettres et, surtout, si une étude syntaxique permet de deviner au mieux sa catégorie grammaticale... Il reste quand même une gageure, sauf si on se limite au vocabulaire courant d'un individu particulier... Car, comme chacun le sait, nous n'utilisons régulièrement qu'un nombre réduit de mots et recourons fréquemment aux mêmes structures de phrase.

Nous allons donc présenter, dans cet article, un système informatique de prédiction lexicale et syntaxique, à partir du contexte gauche. Ces travaux ont conduit à la réalisation d'un prototype de logiciel d'aide à la communication pour des personnes handicapées physiques, le système *HandiAS* [Le Pévédic, 1997]. Il s'agit d'un système hybride, à la fois symbolique et statistique. La partie symbolique est basée sur les notions de *schéma de phrases* et d'*acceptabilité*, notions introduites par Z. S. Harris [Harris, 1968]. La partie statistique, quant à elle, repose sur différentes études concernant la fréquence des mots, en particulier celles de N. Catach [Catach, 1984] et d'A. Juilland [Juilland *et al.*, 1970], et sur la notion d'automates stochastiques (ou chaînes de Markov) [Ménézo, 1992; El-Bèze, 1993].

Dans une première partie (§ 2), nous décrirons notre dictionnaire et la prédiction lexicale qui lui est associée; puis, nous parlerons de l'apport des notions syntaxiques (§ 3); enfin, nous expliquerons comment le système évolue et s'adapte à l'utilisateur (§ 4). Ceci avant de conclure en citant quelques résultats et perspectives.

2. Un dictionnaire fréquentiel

Notre dictionnaire de base est, à peu de chose près, la fusion de deux dictionnaires : celui d'A. Juilland (*opus cit.*) qui est constitué d'environ 18 000 mots fléchis (5 083 lemmes), ce qui représente une couverture lexicale de 92,43 % des mots d'un texte; et celui de N. Catach (*opus cit.*) qui est, *grosso modo*, un sous-ensemble du précédent et qui ne contient plus que 4 000 mots fléchis (1 620 lemmes), mais couvre encore 90,51 % des potentialités d'occurrences des textes français.

Bien sûr, les 10 % d'erreurs restants représentent environ deux mots non prédits par phrase, ce qui est énorme. C'est le module d'adaptation à l'utilisateur (§ 4) qui comblera progressivement ce déficit. Au simple niveau lexical, cette adaptation consiste à introduire des mots nouveaux.

Cette introduction peut être manuelle, décidée par l'utilisateur, ou automatique, avec demande de validation[1].

Mais le rôle principal de ce dictionnaire est de gérer l'aspect fréquentiel du vocabulaire courant d'une personne. Pour cela, nous associons à chaque mot fléchi un *nombre d'occurrences* et à chaque lemme un *numéro de coefficient d'usage*, calculé à partir des occurrences respectives de chaque flexion. Ce numéro est dû à A. Juilland qui s'en sert pour corriger la faible apparition de certaines formes fléchies de mots pourtant courants. Ainsi, la véritable fréquence prise en compte est d'abord celle du lemme, par le coefficient d'usage, la distinction entre les différentes formes fléchies n'intervient qu'ensuite. En fait, ce système de tri à deux clefs ne sera que peu utilisé, puisque la prédiction syntaxique nous permet, en général[2], de choisir une seule proposition parmi les flexions possibles d'un mot.

Prenons un exemple[3], issu du fonctionnement de *HandiAS* : la phrase « *Il danse le twist.* » Rappelons que, dans cette partie, il s'agit uniquement de prédiction lexicale, sans critères syntaxiques; nous reprendrons ce même exemple par la suite (§ 3), ce qui nous permettra de constater l'importance de l'apport de la syntaxe dans la prédiction. Lorsque l'utilisateur saisit la première lettre du deuxième mot, le *d*, *HandiAS* lui propose la liste : *de*, *d'*, *des*; cette liste est composée des trois mots les plus fréquents commençant par *d*. La saisie du *a* permet une nouvelle liste de propositions : *dans*, *dame*, *dames*. Viens la saisie du *n* avec *danger*, *dangers* et *dangereux*. Et, enfin, celle du *s*.

Détaillons la suite du processus. La tableau 1 présente les mots de notre dictionnaire commençant par *dans* : à gauche, dans l'ordre alphabétique, et, à droite, triés suivant deux clefs, le numéro de coefficient d'usage (NCU) et le nombre d'occurrences. *HandiAS* propose trois mots; comme le mot *dans* a déjà été proposé, ce sera donc ici les mots *danser* (Numéro de coefficient d'usage, 1558; nombre d'occurrences, 6), *danse* (Numéro de coefficient d'usage, 1558; nombre d'occurrences, 4) et *dansait* (Numéro de coefficient d'usage, 1558; nombre d'occurrences, 3). Remarquons que le nom *danse* n'est pas sélectionné, malgré son grand nombre d'occurrences, car son numéro de coefficient d'usage est plus élevé.

---

[1] Cette procédure de validation n'est pas encore implantée au niveau du prototype actuel.

[2] Sauf pour les verbes...

[3] Cet exemple est encore académique : il est sujet à modification lors de l'implantation véritable.

Tableau 1

Un extrait du dictionnaire (les mots commençant par *dans*)

Extrait du dictionnaire			
Mots	Catégories	NCU	Occ.
dans	Prep	16	4238
dansaient	V 3p	1558	2
dansait	V 3s	1558	3
dansant	Vant	1558	3
danse	Nom f-s	2151	14
danse	V 1s	1558	4
danser	Vinf	1558	6
dansera	V 3s	1558	1
danses	Nom f-p	2151	3
dansèrent	V 3p	1558	1
dansons	V 1p	1558	2

Tri suivant deux clefs			
Mots	Catégories	NCU	Occ.
dans	Prep	16	4238
danser	Vinf	1558	6
danse	V 1s	1558	4
dansait	V 3s	1558	3
dansant	Vant	1558	3
dansaient	V 3p	1558	2
dansons	V 1p	1558	2
dansera	V 3s	1558	1
dansèrent	V 3p	1558	1
danse	Nom f-s	2151	14
danses	Nom f-p	2151	3

3. Schémas de phrase et automates stochastiques

Voyons maintenant notre modèle syntaxique. Nous représentons une phrase par un schéma directement inspiré de celui proposé par Fathi Debili [Debili, 1982]. Celui-ci, que nous appellerons ici *automate de schéma de phrases*, comporte des chaînes nominales, des chaînes verbales et des ruptures (certaines conjonctions, ponctuations ou prépositions). C'est sur ce schéma que nous insérons les probabilités de succession entre ces différents éléments. Celles-ci proviennent d'une étude de corpus que nous avons réalisée et seront progressivement modifiées en fonction de l'utilisateur. Prenons l'exemple des deux phrases : « *Il danse le twist.* » et « *danse-t-il le twist?* ». Leur syntaxe est reconnue par l'automate de la figure 1. Pour simplifier cette présentation, nous n'avons pas placé sur cette figure, ni sur les suivantes, les fréquences associées à chaque transition ; mais ces automates sont tous stochastiques.

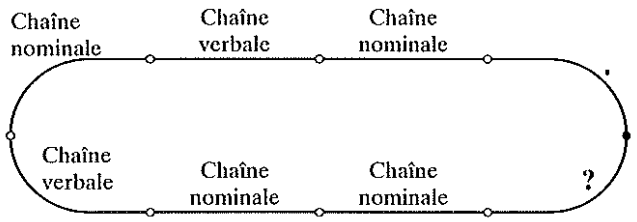

Fig. 1.– Un exemple d'automate de schéma de phrases

Cet automate n'est pas figé, il est, lui aussi, évolutif (voir § 4). Les chemins (ou *transitions*) étiquetées par CN ou CV renvoient à deux autres automates, l'automate des chaînes nominales et celui des chaînes verbales. Les étiquettes de ces deux automates sont alors uniquement des catégories syntaxiques (fig. 2). Sur cette figure, nous constatons que ces trois automates ne sont pas déterministes, ce qui est normal, puisque chaque transition est complétée par une fréquence. C'est elle qui détermine justement le choix du chemin à suivre. Nous avons hésité entre deux possibilités : soit prendre d'abord tous les mots de la catégorie grammaticale de plus forte fréquence avant d'envisager une autre catégorie, soit considérer les mots des catégories grammaticales ayant des fréquences voisines. La dernière solution, testée en considérant les fréquences inférieures à moins de 30 % de la plus élevée, est plus performante. Mais d'autres heuristiques de pondérations pourraient sans doute être envisagées dans la suite.

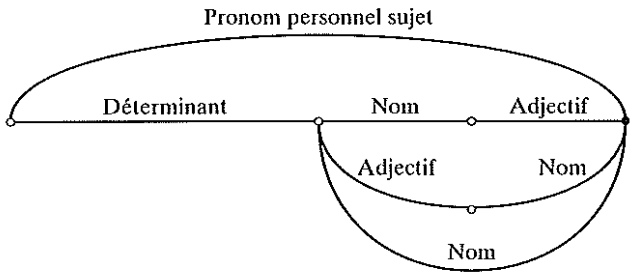

Fig. 2.– Un exemple d'automate de chaîne nominale

Reprenons l'exemple présenté précedemment (§ 2) après la saisie du mot *Il*. Dès la lecture du pronom personnel sujet, *HandiAS* va proposer une liste des trois verbes les plus fréquents, à la troisième personne du singulier (*est*, *était* et *a*). Puis, après la saisie de la lettre *d*, nous obtiendrons, comme le montre le tableau 2, les trois propositions *dit*, *disait* et *dira*. Enfin, dès la lettre *a*, se trouve le mot que nous souhaitons écrire, parmi les trois propositions *danse*, *dansait* et *dansera*.

Tableau 2

Le numéro de coefficient d'usage des verbes commençant par *d* et des deux verbes commençant par *da*

Lemmes	NCU
dire	41
donner	64
devoir	66
... 29 verbes...	
danser	1558
... 11 verbes...	
dater	2454
... 45 verbes...	
dérouter	5013

4. Évolution et adaptation

Bien sûr, au fur et à mesure des utilisations, *HandiAS* va évoluer et s'adapter au vocabulaire et à la syntaxe employés. La saisie d'un mot connu (c'est-à-dire dans le dictionnaire ou dans la liste des mots nouveaux déjà saisis) entraînera l'incrémentation automatique de son nombre d'occurrences. De même, au niveau syntaxique, les fréquences associées à chaque transition des trois automates stochastiques sont, elles aussi, incrémentées avec la saisie.

Précisons maintenant comment sont traités les mots nouveaux[4]. Continuons sur notre exemple, avec la phrase « *Il danse le twist.* » Le mot *twist* n'est pas dans notre dictionnaire d'origine. Il sera donc considéré comme un mot nouveau. Puisque nous avons reconnu le déterminant *le*, le système *HandiAS* suppose que le mot est un nom (masculin et singulier), comme indiqué sur l'automate des chaînes nominales (voir la figure 2, la fréquence Déterminant–Nom étant plus élevée que la fréquence Déterminant–Adjectif). Cette supposition sera conservée en mémoire, mais pourra être modifiée en adjectif (deuxième choix possible sur cette même figure), si cela s'avérait nécessaire. Le nombre d'occurrences

[4] Rappelons que cette partie n'est pas encore entièrement développée dans le prototype existant.

du mot *twist* évoluera ensuite de la même manière que les mots du dictionnaire. Cependant, pendant la session courante, ce mot bénéficiera d'un numéro de coefficient d'usage provisoire peu élevé, de façon à ce qu'un mot rare, déjà utilisé dans un texte, puisse réapparaître dans les listes de suggestions.

À la fin de la session de travail, *HandiAS* proposera ses hypothèses sur le mot *twist* (catégorie, genre et nombre), ainsi que des hypothèses sur le lemme et les flexions associés. Après validation par l'utilisateur, le mot sera intégré au dictionnaire. Si l'utilisateur pense ne jamais le réutiliser, il peut aussi choisir de refuser cette intégration ; cela sera, en particulier, très utile pour des noms propres très occasionnels. Lorsque des mots nouveaux sont ainsi intégrés au dictionnaire, le numéro de coefficient d'usage est recalculé, à partir des occurrences de chaque lemme.

La syntaxe, elle aussi, est susceptible d'évoluer. Prenons la phrase « *Comment se danse le twist ?* ». Celle-ci commence par une rupture, puis se poursuit par une chaîne verbale et une chaîne nominale, avant de se terminer par un point d'interrogation. Après vérification de l'acceptabilité de cette structure syntaxique dans une table, l'automate de phrase (fig. 1) sera modifié, comme le présente la figure 3.

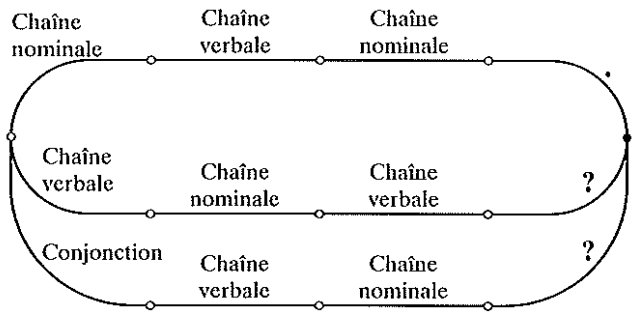

Fig. 3.– Modification de l'automate de phrase

Ces tables d'acceptabilité[5] sont au nombre de deux, une pour chaque chaîne, et indiquent les successions possibles entre les différentes catégories syntaxiques. Dans notre exemple, pour modifier l'automate de phrase comme nous venons de le faire, il faut savoir, d'une part, par quelle catégorie syntaxique une chaîne peut débuter ou finir et, d'autre part,

[5] Sur la notion de table d'acceptabilité, voir [Maurel, 1991].

Tableau 3
Une table d'acceptabilité

	Début de chaîne	Déterminant	Pronom personnel sujet	Nom	Adjectif	Fin de chaîne
Déterminant	+	–	–	+	+	–
Pronom personnel sujet	+	–	–	–	–	+
Nom	–	–	–	+	+	+
Adjectif	–	–	–	+	+	+

quelles sont les successions de catégories autorisées à l'intérieur d'une même chaîne. Le tableau 3 présente une table d'acceptabilité pour les chaînes nominales reconnues par l'automate de la figure 2.

7. Résultats et perspectives

Actuellement, le système *HandiAS* reste un prototype et n'a pas été testé en situation d'aide à la communication des handicapés. Des contacts ont été pris pour cela. Cependant, des tests ont été réalisés. Un corpus, la *Charte de l'environnement Malgache*, nous a servi de corpus d'apprentissage syntaxique et de corpus de test. Le tableau 4 présente ces résultats en donnant : d'une part, le nombre de mots reconnus avant la saisie complète de toutes les lettres et, donc, sélectionnés par l'utilisateur ; d'autre part, le nombre de lettres saisies sur le clavier simulé, avant cette reconnaissance. Ces deux types d'actions représentent l'ensemble des *actions de frappe* de l'utilisateur ; le total de ces actions de frappe doit donc être comparé au nombre total de caractères du texte à saisir, c'est-à-dire à 10 866 actions. Pendant ces trois tests, le module d'adaptation a fonctionné normalement.

Nous constatons que l'apport combiné de la fréquence d'apparition des mots et de la syntaxe (y compris l'adaptation) permet d'obtenir des

Tableau 4

Les résultats du test

		Avec notre système			Sans lui
		Fréquentiel	Syntaxique	*HandiAS*	
Actions	Sélection	3680	3653	3345	0
	Saisie	3290	2953	2791	10866
	TOTAL	6970	6606	6136	10866
% d'actions faites		64,15	60,80	56,47	100
% d'actions économisées		35,85	39,20	43,53	0

résultats bien supérieurs à l'une de ces deux méthodes prise isolément. Ce qui nous encourage à poursuivre nos travaux.

Il faudrait en effet améliorer la constitution de notre base de départ, par un travail sur corpus plus ambitieux que celui que nous avons eu le temps de réaliser jusqu'à présent. Le module syntaxique pourrait être complété (chaînes adjectivales, chaînes adverbiales, concordance des temps...). Nous pensons aussi travailler l'aspect sémantique, en ajoutant peut-être un *thesaurus* sur les lemmes... Enfin, nous comptons aussi adapter notre système à d'autres langues, car les programmes informatiques développés dans *HandiAS* sont indépendants de la langue. Nous cherchons pour cela à établir des collaborations avec des linguistes travaillant sur d'autres langues que le français.

Bibliographie

CATACH (N.) : 1984, *Les listes orthographiques de base du français* (Paris, Nathan).

DEBILI (F.) : 1982, *Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales sémantiques* (Thèse d'état, Université Paris 11).

EL-BÈZE (M.) : 1993, *Les modèles de langue probabiliste : Quelques domaines d'application* (Habilitation à diriger des recherches, LIPN Université Paris-Nord).

HARRIS (Z. S.) : 1968, *Mathematical Structures of Language* (Interscience Publishers, Paris, Dunod, 1971, traduction française).

JUILLAND (A.), BRODIN (D.), DAVIDOVITCH (C.) : 1970, *Frequency dictionary of french words* (La Haye, Mouton & Co).

LE PÉVÉDIC (B.) : 1997, *Prédiction Morphosyntaxique Évolutive dans un système d'aide à la saisie de textes pour des personnes handicapées physiques* (Thèse de doctorat en informatique, Université de Nantes).

MAUREL (D.) : 1991, « Préanalyse des adverbes de date du français », *TA information*, volume 32, n° 2, pp. 5–17.

MÉNÉZO (J.) : 1992, *Désambiguïsation lexicale par filtrage en cascade* (Mémoire CNAM, Grenoble).