MATHÉMATIQUES ET SCIENCES HUMAINES

BRIGITTE LE PÉVÉDIC
DENIS MAUREL

Un dictionnaire électronique évolutif par apprentissage

Mathématiques et sciences humaines, tome 136 (1996), p. 43-49
<http://www.numdam.org/item?id=MSH_1996__136__43_0>

Math. Inf. Sci. hum. (34e année, n°136 *1996*, pp.43-49)

UN DICTIONNAIRE ÉLECTRONIQUE ÉVOLUTIF PAR APPRENTISSAGE

Brigitte LE PÉVÉDIC[1] et Denis MAUREL[2]

RÉSUMÉ — *Dans le cadre du développement d'un système d'aide à la saisie rapide de textes pour des personnes handicapées physiques, nous allons aborder, dans cet article, l'approche lexicale.*

Le principe du système est le suivant : en fonction du début du texte déjà saisi et des premières lettres du mot en cours de saisie nous comptons proposer à l'usager la liste des mots les plus fréquents dans la ou les catégories grammaticales les plus probables, tout en respectant le cadre sémantique.

Il s'agit donc de réaliser un dictionnaire électronique comportant des indications morphologiques, syntaxiques et sémantiques, ainsi qu'un système d'apprentissage permettant une adaptation de ce dictionnaire à l'utilisateur.

SUMMARY — An electric dictionary with a learning system

This paper presents the work to construct a communication-aid software in order to speed up the capture of data for physically handicapped people. In this article, the lexical approach is developed.

The principle is : in function for beginning texts and the first letters of words current capture, we want to suggest to users a list of more frequently used words in a more probable grammatical category with a correct semantical context.

We propose therefore to carry out an electronic dictionary with morphological, syntactical and semantical informations and a learning system which allows adjustment to this dictionary for the user.

INTRODUCTION

L'objet de ce travail est le développement d'une méthodologie d'aide à la communication écrite par des techniques de prédictions morphosyntaxiques et d'évolutivités. Le temps de saisie d'un texte pour des personnes handicapées physiques est un problème majeur. En effet, Boissière 1990 [2] fait la remarque suivante : "supposons que la personne ait pu trouver le matériel le mieux adapté à son handicap lui permettant d'écrire avec le maximum de confort. Alors le handicapé au mieux de ses possibilités atteindra sa vitesse de frappe maximum qui restera quand même largement en dessous des performances d'une personne valide (1 à 6 caractères/minute pour un tétraplégique contre 10 mots/minute pour un élève de 8-9 ans, 25 mots/minute pour une secrétaire)".

---

[1] Institut de Recherche en Informatique de Nantes (IRIN) - 2, rue de la Houssinière, BP 92208 44322 Nantes cedex 03. Tel: (33) 02-40-37-30-37 - Email : lepevedic@irin.univ-nantes.fr

[2] LI/E3I/ Université François-Rabelais - 64, avenue Jean-Portalis 37 200 Tours. Tel: (33) 02-47-36-14-35 - Email : maurel@univ-tours.fr

On peut différencier deux grandes tendances : la première apporte une aide technique en adaptant un appareillage, mais cette méthode ne permet pas forcement une vitesse de frappe rapide. La seconde tendance, celle que nous avons retenue, correspond à une aide logicielle.

Il s'agit d'accélérer la saisie de textes pour des personnes handicapées physiques (ne disposant pas de l'usage normal d'un clavier) en proposant le plus rapidement possible le mots qu'elles souhaitent écrire. Cette technique permet d'économiser la saisie des dernières lettres de chaque mot. Le principe est le suivant (*cf.* figure 1), nous mettons en évidence la liste des lettres les plus probables en fonction du début du mot saisi ainsi que la liste des mots les plus fréquents en fonction du début de la phrase. Un curseur se déplace à l'écran sur ses listes ; l'utilisateur n'a plus qu'à sélectionner une des propositions et celle-ci vient s'inscrire directement dans son texte.

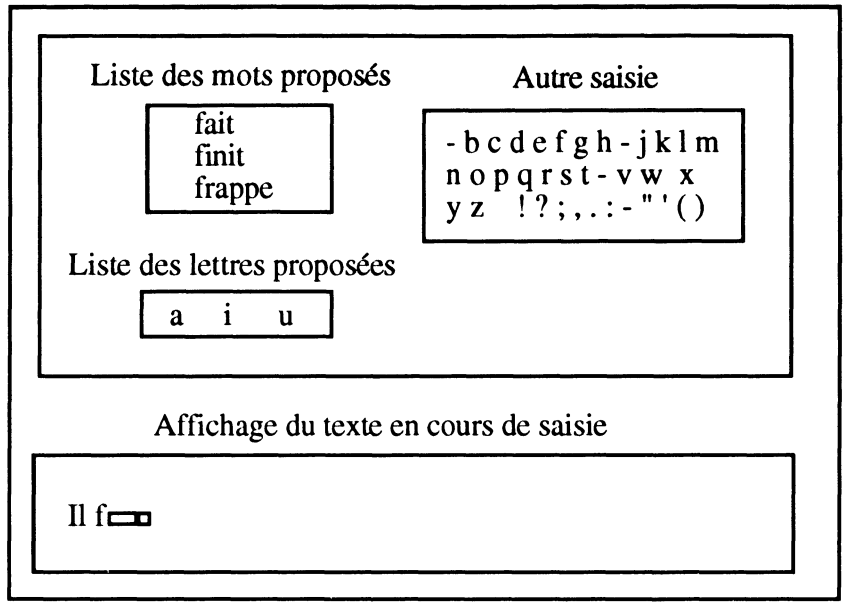

Figure 1 : Interface du logiciel de saisie

Nous avons donc besoin d'un dictionnaire comportant des indications morphologiques, syntaxiques et sémantiques ainsi que d'un système d'apprentissage permettant une adaptation de ce dictionnaire à chaque utilisateur.

Le caractère "évolutif" du dictionnaire permet, d'une part, l'enrichissement du dictionnaire initial (c'est à dire l'introduction éventuelle dans le dictionnaire des mots "nouveaux" apparus lors de la saisie d'un texte) ; et, d'autre part, la modification de la fréquence des mots au fur et à mesure des utilisations. A partir d'une version initiale, le dictionnaire sera progressivement adapté à l'usager.

Il sera également possible de faire appel, lors d'une saisie, à des dictionnaires spécifiques d'un domaine (histoire, géographie, sciences, etc...).

La première partie traitera des problèmes linguistiques rencontrés lors de la constitution de ce dictionnaire ; une deuxième partie des solutions informatiques. Enfin une conclusion définira la suite à donner à ce projet.

1. ÉTUDE LINGUISTIQUE

1.1. La notion de fréquence

La fréquence est le paramètre le plus important du dictionnaire. Elle permet de sélectionner parmi l'ensemble du vocabulaire de l'utilisateur les trois mots et trois lettres à lui proposer. De plus, cette information va permettre d'adapter le dictionnaire à chaque utilisateur, afin de

distinguer son vocabulaire courant de son vocabulaire occasionnel. Il sera ainsi possible de lui proposer en priorité les mots appartenant à son vocabulaire usuel. Le premier dictionnaire que nous avons constitué est issu principalement des études de N. Catach 1984 [2] et A. Juilland 1970 [5].

Déterminons tout d'abord ce que l'on entend par le terme de "fréquence".

1.2. La méthode de calcul proposée par A. Juilland
Le corpus utilisé par A. Juilland est d'une très grande richesse. Il est constitué de 3 078 livres ou articles de 513 auteurs. Chaque texte sélectionné est affecté à l'une des cinq catégories suivantes : Roman, Théâtre, Essais, Presse, Textes scientifiques et techniques (*cf.* figure 2).

Catégorie	Auteur/Titre	Livre/Article
Roman	54	344
Théâtre	105	851
Essais	108	805
Presse	34	435
Sciences	212	643
Total	513	3078

Figure 2 : Composition du corpus de A. Juilland

A. Juilland va ensuite établir entre ces cinq catégories un indice d'apparition, appelé indice de "dispersion" (*D*). Cet indice de dispersion n'intervient que pour les mots sous leurs formes lemmatisées, il sera compris entre 0 et 1.
Il se calcule de la manière suivante :

$$1 - \frac{\sqrt{N \Sigma X_i^2 - F^2}}{2*F}$$

où :

- X_i est le nombre d'apparitions du lemme dans la catégorie *i*,
- *F*, la somme des X_i d'un lemme c'est à dire le nombre total d'apparitions du lemme,
- *N*, le nombre de catégories (dans Juilland *N*=5).

Un coefficient d'usage (*CU*) est calculé pour chaque mot lemmatisé : $CU = (F*D) / 100$. Ce coefficient d'usage repose sur une mise en rapport du nombre d'occurrences (*F*) d'un mot lemmatisé et de son indice de dispersion (*D*). L'indice de dispersion est donc de plus en plus efficace au fur et à mesure que le nombre d'apparitions d'un mot diminue : il est donc peu pertinent pour les mots très utilisés.

Le nombre d'occurrences d'un mot lemmatisé (*F*) est égal au total de toutes les occurrences de toutes ses formes fléchies. Ceci implique que le coefficient de dispersion ne peut être calculé que pour les lemmes. Chaque flexion d'un mot se réfère donc au lemme dont il découle.

Un numéro d'identification du coefficient d'usage (*NCU*) par ordre décroissant est associé au *CU* car deux lemmes peuvent avoir le même *CU*.

Pour résumer, *le mot le plus fréquent est celui qui à le plus grand coefficient d'usage (ou le plus petit NCU) et le nombre d'occurrences le plus élevé. La fréquence intervient au niveau lexical (chaque mot est relié au lemme dont il dépend) mais il existe une répartition morphologique (chaque flexion possède un nombre d'occurrences propre).*

1.3. Le contenu du dictionnaire
A. Juilland a créé un dictionnaire comportant aux alentours de 25000 mots (lemmes et formes fléchies) soit 5083 lemmes. Les 4000 mots sélectionnés par N. Catach, soit 1620 lemmes, sont un sous-ensemble du dictionnaire de A. Juilland et représentent les mots les plus usités de la langue française.

Les formes fléchies de ces mots sont introduites dans le dictionnaire. Ainsi il n'est pas nécessaire de fléchir les mots avant de les proposer à l'utilisateur et par conséquent la vitesse d'exécution du logiciel s'en trouve améliorée. De plus les formes fléchies d'un mot n'ont pas la même probabilité d'apparition.

Les locutions sont également introduits car ils ont une fréquence propre. Par exemple les fréquences des mots "temps", "de" et "en" seront différentes de la fréquence du mot composé "de temps en temps".

Nous avons séparé notre dictionnaire en deux parties : une partie chargée en mémoire et une autre stockée sur disque. Le dictionnaire placé en mémoire sera constitué de l'ensemble des 4000 mots sélectionnés par N. Catach. Quant au dictionnaire placé sur disque, il contiendra l'ensemble des formes répertoriées dans le dictionnaire de A. Juilland dont les formules de calcul sur la fréquence ont été reprises et adaptées.

Le dictionnaire chargé en mémoire est d'une taille modeste, mais l'ensemble lexical qui y est présenté permet de couvrir 90,51% des occurrences d'un texte courant. Un ajout massif supplémentaire n'augmenterait ce pourcentage que d'une façon dérisoire sans jamais atteindre 100%.

Une entrée du dictionnaire contient le mot sous sa forme fléchie, le numéro d'identification du coefficient d'usage (*NCU*) et le nombre d'occurrences du mot. Par exemple, le mot le plus fréquent , "le" aura pour *NCU* 1 et pour nombre d'occurrence 8609; et deux mots ayant le même *NCU* (issus donc d'un même lemme) par exemple "un" et "une", vont différer par leur nombre d'occurrences (*cf.* figure 3).

Forme fléchie	NCU	Occurrences
le	1	8609
un	3	5968
une	3	5076
Beau	541	146
Belles	541	121

Figure 3 : Cinq entrées du dictionnaire

1.4. L'évolution de la fréquence
Comme nous l'avons vu, la "fréquence" d'un mot dépend de deux paramètres, le nombre d'occurrences du mot et le coefficient d'usage.

Sa mise à jour se passe en deux temps. En effet, les deux paramètres ne peuvent évoluer simultanément puisqu'ils ne font pas référence aux mêmes notions.

- Le premier paramètre dépendant du nombre d'occurrences du mot, il suffit, pour le faire évoluer, de l'incrémenter à chaque fois que le mot est utilisé.
- Le second paramètre concerne l'ensemble des formes fléchies du mot lemmatisé. Il faut pour chaque lemme recalculer un coefficient d'usage en utilisant la méthode de A. Juilland. Pour que l'évolution de ce paramètre soit significative, il faut attendre d'avoir saisi un nombre représentatif de mots. Une mise à jour régulière est donc prévue.

2. ÉTUDE INFORMATIQUE

2.1 La représentation du dictionnaire en mémoire

La plupart des travaux utilisant un dictionnaire (par exemple B. Courtois et M. Silberztein 1990 [4]), sont astreints à une concentration maximal des informations (*cf.* D. Revuz 1991 [11] et M. Mohri 1994 [10]). Pour notre application, seul l'ensemble du vocabulaire courant de l'utilisateur (environ 4 500 mots) est chargé en mémoire. De plus la rapidité de la recherche des informations pour la création de la liste des propositions repose sur l'organisation du dictionnaire en mémoire et sur son accès.

Nous avons opté pour une représentation sous la forme de 26 listes de mots chaînées dans l'ordre alphabétique ; chaque liste est, elle même, composée de sous-listes chaînées par catégorie grammaticale ; une table (*cf.* A. Aho et al 1993 [1]) repère le premier mot de chacune des sous-listes. Cette structure permet d'introduire, également, un chaînage grammatical qui trouvera son utilité lors de l'intervention du critère syntaxique dans la recherche (*cf.* figure 4).

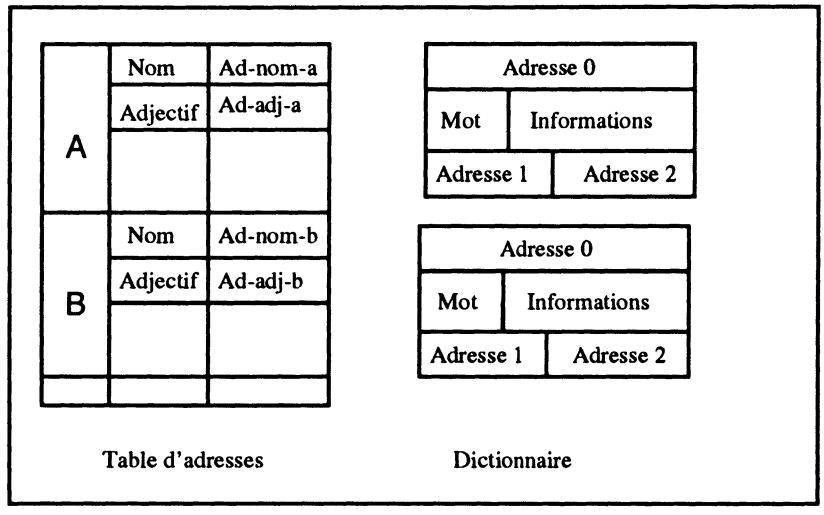

Figure 4 : Structure du dictionnaire en mémoire

Légende :
- Ad-x-y : adresse du premier mot commençant par la lettre y de la catégorie grammaticale x
- Adresse 0 : adresse du mot en mémoire
- Adresse 1 : adresse (s'il elle existe) du prochain mot commençant par la même lettre
- Adresse 2 : adresse (s'il elle existe) du prochain mot ayant la même catégorie grammaticale.

2.2. La recherche d'un mot

Le critère de recherche d'un mot est actuellement lexical : en fonction des premières lettres du mot, le logiciel doit proposer à l'utilisateur les mots et les lettres les plus fréquents correspondant au début de la saisie.

A partir de la première lettre saisie, on recherche dans la table d'adresses la sous-liste à explorer (26 sous-listes possibles). Une fois cette sous-liste détectée, nous allons au fur et à mesure de la saisie du mot la restreindre.

Par exemple :
Si nous saisissons la lettre "m", nous allons travailler sur la 14ème sous-liste. Puis si la saisie se poursuit par la lettre "e", nous allons enlever de la 14ème sous-liste, tous les mots précédent "me" ainsi que l'ensemble des mots suivant "me". Cette restriction est simple à réaliser, il suffit d'utiliser deux pointeurs de délimitation. Cet intervalle diminue à chaque saisie d'une nouvelle lettre, ce qui accélère d'autant plus le temps de réponse du logiciel (*cf.* figure 5).

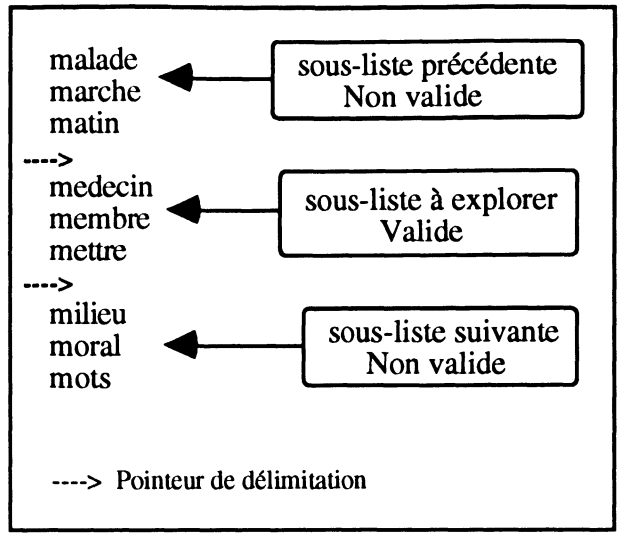

Figure 5 : Recherche d'un mot en mémoire

CONCLUSION

Comme nous l'avons montré, il est possible de proposer rapidement une liste de mots probables lors de la saisie des premières lettres. A condition de baser ce système sur des études linguistique et informatique approfondies. De plus, la prise en compte de l'évolution de la fréquence de chaque mot permet d'obtenir un dictionnaire électronique évolutif et donc adapté à chaque utilisateur.

Cependant, pour mener à bien ce projet, certains points restent à travailler. La création et la gestion d'un dictionnaire évolutif comportant des fréquences d'apparitions ne sont pas suffisantes pour déterminer le mot qu'un utilisateur veut saisir. En effet, la proposition d'un mot ne dépend pas uniquement de sa fréquence. Elle dépend, également, du contexte.

L'introduction de la syntaxe et la sémantique semble indispensable pour un fonctionnement correct.

- L'introduction de la syntaxe va permettre de déterminer la (ou les) catégorie(s) grammaticale(s) la (ou les) plus probable(s) suivant le contexte gauche (*cf.* F. Debili 1982 [5]). L'analyseur ainsi construit va aider à la construction correcte d'une phrase. actuellement cet analyseur est en cours de réalisation (*cf.* B. Le Pévédic, D. Maurel [8] et [9]).

- L'introduction de la sémantique, avec la constitution d'un réseau sémantique (thésaurus) va permettre de sélectionner les mots à présenter dans un cadre sémantique cohérent avec le contexte qui précède. On pourra par exemple consulter à ce sujet J. C. Lejosne *et al.* 1992 [7]. Cette partie fera l'objet de travaux futurs.

BIBLIOGRAPHIE

[1] AHO, A., HOPCROFT, J., ULLMAN, J., *Data structure and algorithms,* Reading, Addisson-Wesley, 1983 ; Paris, Inter édition, 1987 (traduction française).
[2] BOISSIÈRE, P., *Un système auto-organisationnel pour faciliter le dialogue écrit homme-machine*, Thèse d'état, Université de Toulouse IRIT, 1990.
[3] CATACH, N., *Les listes orthographiques de base du français (LOB)*, Paris, Nathan-recherche, 1984.
[4] COURTOIS, B., SILBERZTEIN, M., "Dictionnaire électronique du français", *Langues française*, n°87, Paris, Larousse, septembre 1990, 11-22.
[5] DEBILI, F., *Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales sémantique*, Thèse d'état , Paris, 1982.
[6] JUILLAND, A., BRODIN, D., DAVIDOVITCH, C., *Frequency dictionary of french words*, 5000 mots, Paris, Mouton, 1970.
[7] LEJOSNE, J.-C., LAUVRAY, J., KLEIN, J., ROMARY, L., "Étude de groupes nominaux complexes", *Lexique et Inférence*, Paris, Klincksieck, 1992.
[8] LE PÉVÉDIC, B., MAUREL, D., "La prédiction d'une catégorie grammaticale dans un système d'aide à la saisie pour handicapés", *in Actes TALN'96*, Marseille, 1996.
[9] LE PÉVÉDIC, B., MAUREL, D., "Un logiciel d'aide à la communication pour des personnes handicapées", *in Actes NLP+IA'96*, Moncton (Canada), 1996.
[10] MOHRI, M., *On some applications of finite-state automate theory to natural language processing*, Rapport de recherche de Institut Gaspard Monge, 94-22, Université de Marne-la-Vallée, 1994.
[11] REVUZ, D., *Dictionnaires et lexiques : méthodes et algorithmes*, Thèse de doctorat, Institut Blaise Pascal, Paris, 1991.