Ressources et activités pédagogiques dans un environnement d'aide à l'apprentissage lexical du français langue seconde

Thierry Selva

Numéro d'ordre : 759 Année 1999

THÈSE

Présentée à

L'UFR DES SCIENCES ET TECHNIQUES
DE L'UNIVERSITÉ DE FRANCHE-COMTÉ

Pour obtenir le

GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE FRANCHE-COMTÉ

Spécialité : Automatique et Informatique

Ressources et activités pédagogiques dans un environnement informatique d'aide à l'apprentissage lexical du français langue seconde

Par

Thierry Selva

Soutenue le 29 octobre 1999 devant la Commission d'examen
composée de

Paul BOGAARDS	Rapporteur
Thierry CHANIER	Directeur
Paul SABATIER	Rapporteur
Michel TRÉHEL	Examinateur

Remerciements

Je remercie vivement en premier lieu mon directeur de thèse, le professeur Thierry Chanier, pour toute l'aide précieuse et les nombreux conseils qu'il m'a prodigués tout au long de ce travail.

Par ailleurs, je remercie les personnes à l'origine du projet ALEXIA auquel j'ai participé à partir de la fin 1994 : Thierry Chanier, Colette Colmerauer, Christophe Fouqueré, Anne Abeillé, Françoise Picard et Michael Zock.

Je remercie plus spécifiquement :
- Sylvain Pichon, pour son excellent travail sur l'affichage graphique de réseaux lexicaux et dont le travail de programmation a été intégré dans ALEXIA,
- Christophe Fouqueré, pour m'avoir donné la possibilité d'utiliser son analyseur morphologique indispensable à plusieurs modules d'ALEXIA,
- Fabrice Issac, pour m'avoir aidé dans la programmation des communications entre les langages C et HyperCard,
- Nathalie Cointe, pour son travail linguistique qui a guidé en partie mes descriptions lexicales et pour son apport dans la constitution du corpus,
- Lise Duquette, pour avoir gentiment accepté de relire le début de cette thèse.
- Les membres du Laboratoire de Recherche sur le Langage, Université Clermont2, dans lequel cette thèse a débuté, ainsi que ceux du Laboratoire d'Informatique de Besançon, dans lequel elle s'est continuée.

Je remercie également ma famille et mon grand-père de Los Masos, parti vers un autre monde au cours de cette thèse et à qui je dédie ce travail, pour leur soutien financier.

Enfin, je remercie Isabelle pour avoir eu la patience et le courage de supporter mes nombreuses absences.

Table des matières

Liste des abréviations

Dictionnaires

CECC	COBUILD English Collocations on CD-ROM
CIDE	Cambridge International Dictionary of English (Procter 1995)
COBUILD	Collins Cobuild (Sinclair 1995)
COBUILDCD	Collins Cobuild on CD-Rom (1995)
DEC	Dictionnaire Explicatif et Combinatoire du français contemporain (Mel'cuk 1992)
DFLE	Dictionnaire du Français Langue Étrangère (Dubois 1986)
E-Dict	E-Dict Dictionay (Collins Cobuild 1998)
GLCD	Grand Larousse bilingue français-anglais sur cédérom (1996)
HO	Hachette-Oxford (1994)
HOCD	Hachette-Oxford sur cédérom (1996)
LDOCE	Longman Dictionary Of Contemporary English (Summers 1995)
LIED	Longman Interactive English Dictionary (1993)
LLA	Longman Language Activator (Summers 1993)
MRP	Micro Robert Poche (Rey 1987)
OALD	Oxford Advanced Learner's Dictionary (Crowther 1995)
OWPD	Electronic Oxford Wordpower Dictionary (1994)
PLCD	Petit Larousse sur cédérom (1996)
PR	Petit Robert (Rey 1996)
PRCD	Petit Robert sur cédérom (Rey 1996)
RE	Robert Électronique (1994)
RJI	Robert Junior Illustré (de Bellefonds 1993)

Autres abréviations

ALAO	Apprentissage des Langues Assisté par Ordinateur
LIPN	Laboratoire d'Informatique de Paris-Nord
LRL	Laboratoire de Recherche sur le Langage (Université Clermont2)

INTRODUCTION

Ce mémoire de thèse concerne l'élaboration d'un environnement informatique d'aide à l'apprentissage lexical du français langue étrangère ou seconde, ALEXIA. Au-delà de l'environnement lui-même, dont la conception s'inspire de travaux antérieurs, l'étude porte particulièrement sur les ressources lexicales qui le composent, corpus de textes et dictionnaire électronique, et leur utilisation pour générer des activités pédagogiques dans le cadre de l'apprentissage lexical sur ordinateur.

Cette thèse est issue des travaux de l'atelier « Modélisation de l'acquisition lexicale en langue seconde » du GDR sciences cognitives de Paris. Initialement constitué de Thierry Chanier, Colette Colmerauer, Christophe Fouqueré, Anne Abeillé, Françoise Picard et Michael Zock de 1991 à 1995, cet atelier a débouché sur le projet ALEXIA en 1995, sous la responsabilité de T. Chanier et C. Fouqueré. Il a donné lieu à la soutenance de la thèse de Fabrice Issac, dirigée par C. Fouqueré, « Analyse syntaxique et apprentissage des langues » (Issac, 1997), et du mémoire de DEA de Sylvain Pichon, dirigé par T. Chanier, « Les relations sémantiques dans l'apprentissage lexical en langue seconde : Visualisation graphique et interaction dans un dictionnaire électronique » (Pichon, 1996). Par ailleurs, Nathalie Cointe a travaillé sur les grandes lignes du projet, la description lexicale et la constitution du corpus de 1993 à 1994 en tant que doctorante. Commencée initialement au sein du Laboratoire de Recherche sur le Langage, Université Clermont2, à la fin 1994, cette thèse s'est ensuite déroulée au Laboratoire d'Informatique de Besançon à partir de fin 1996. La liste des publications concernant le projet ALEXIA figure à part à la suite de la bibliographie générale.

La nature même du média informatique tend à privilégier l'autonomie de l'apprenant lors de ses tâches d'apprentissage d'une langue étrangère. Il peut en effet s'adapter au rythme et aux possibilités de l'apprenant. L'examen des environnements lexicaux existants (Mayday, Sussex *et al.*, 1994 ; Lexica, Goodfellow, 1994) montre que l'autonomie passe par l'incorporation de ressources lexicales propres à fournir à l'utilisateur les matériaux d'apprentissage et les outils d'aide. L'environnement ne doit cependant pas se résumer à ces ressources, mais doit s'articuler autour d'un programme pédagogique pertinent. C'est uniquement dans ce sens que l'on peut parler d'aide à l'apprentissage.

Dès lors, la partie centrale de ce travail a été de concevoir, à partir des théories psycholinguistiques sur le lexique mental et de la modélisation de l'apprentissage lexical, les ressources adaptées à un apprentissage en autonomie et la manière dont celles-ci s'articulent.

Notre environnement fonctionne autour d'un schéma d'apprentissage en trois étapes (qui s'applique plus généralement à l'acquisition lexicale) : exposition/compréhension, mémorisation et production/maîtrise du lexique.

Pour traiter le volet de l'exposition/compréhension, un des premiers objectifs a été de circonscrire notre cadre d'étude. Veut-on travailler sur la langue française entière ? Est-ce vraiment un objectif pédagogique réalisable ? Ou doit-on plutôt se concentrer sur une partie spécifique de la langue mais suffisamment générale et s'adressant au plus grand nombre pour pouvoir ensuite être transposée ou étendue ? La couverture de la langue entière impose la constitution de ressources en conséquence : un corpus de textes de très grosse taille (plusieurs dizaines de millions de mots, voire de centaines) contenant l'ensemble des mots de la langue et de leur sens, ainsi qu'un dictionnaire de plusieurs dizaines de milliers d'entrées, comme c'est le cas pour les dictionnaires de langue. Contrairement à l'anglais, ces ressources ne sont pas disponibles directement et de toute manière ne sont pas adaptées à l'apprentissage du français contemporain tel que nous le concevons (par exemple FRANTEXT, corpus portant sur la littérature française du 16^{e} au 20^{e} siècle, ou corpus Le Monde, assemblage hétéroclite d'articles rédigés dans un style défini ; quant aux dictionnaires d'apprentissage, il n'en existe aucun qui soit vraiment satisfaisant pour le français comme nous le verrons dans le chapitre 3). Devant la nécessité de la constitution de ressources propres, les initiateurs du projet au LRL ont opté pour la description d'un champ du français courant, le thème *emploi*, *travail*, *chômage*, suffisamment commun pour pouvoir s'adresser à tous. Mais ce cadre étant défini, que signifie un corpus représentatif ? Quelles doivent en être ses caractéristiques ? Et d'autre part, à quoi va-t-il servir et en quoi son utilisation à d'autres buts que la lecture va en décider sa structure et les informations qu'il porte ?

La problématique des dictionnaires et, plus largement, des bases de données lexicales, possède quelques points communs. Nous verrons que les problèmes d'accès lexicaux et d'aide à la compréhension sont fondamentaux à tout dictionnaire et en conditionnent l'utilité même. Dès lors, à partir de l'examen de dictionnaires existants, nous tenterons de répondre aux questions suivantes : qu'est-ce que l'accès lexical, quels sont les problèmes posés par l'accès lexical et quelles sont les stratégies employées par les apprenants pour retirer l'information ? D'autre part, il faut s'interroger sur la relation entre le dictionnaire papier et le dictionnaire électronique. En quoi sont-ils différents ? Les informations présentées seront-elles les mêmes ? Qu'apporte le support électronique ? Quelles limites liées au support papier permet-il de franchir ? Y a-t-il des moyens différents de présenter l'information et si oui, permet-il d'autres types de présentation de l'information ?

La modélisation même de cette base de données informatique tiendra bien sûr compte de l'accès lexical et des problèmes d'aide à la compréhension. Mais de même que les corpus, il faut aussi s'interroger sur l'utilité d'une telle base lexicale dans un environnement informatique d'aide à l'apprentissage lexical. Servira-t-elle uniquement pour la consultation et la compréhension ?

Le deuxième volet de notre schéma concerne la mémorisation du vocabulaire. Sur ce point, nous nous en tiendrons aux travaux de Goodfellow sur l'utilité d'organiser son propre vocabulaire dans un dictionnaire personnalisé. Le fait de devoir structurer son propre vocabulaire implique une réflexion plus profonde sur celui-ci. Comme l'affirment les modélisations de l'apprentissage, l'effort mental ainsi généré est bénéfique pour l'incorporation de nouvelles connaissances au sein des anciennes et favorise de ce fait la rétention du vocabulaire.

Mais c'est le troisième volet, celui de la maîtrise du vocabulaire qui va guider le plus fortement la conception de notre environnement. Outre la structuration et la réflexion sur le vocabulaire, la maîtrise lexicale passe aussi par l'entraînement et l'exécution d'activités lexicales qui favorisent l'inférence, une des stratégies les plus couramment employées pour l'apprentissage lexical. Parmi toutes celles existantes, il convient d'en isoler certaines, que l'on choisira parmi les plus pertinentes, en ayant en tête les deux questions suivantes : en quoi l'informatique va permettre de générer automatiquement des activités d'un type nouveau, différent du support papier, et en quoi ces activités tiendront-elles compte des caractéristiques des ressources de l'environnement (corpus, dictionnaire mais aussi dictionnaire personnel) ? Dès lors, la conception des ressources et des activités sera fortement liée : quelles ressources pour quelles activités ? Il nous revient donc de définir les caractéristiques de ces activités en fonction des matériaux disponibles.

Ce travail sera composé de neuf chapitres au cours desquels l'environnement sera peu à peu dessiné. Après quatre premiers chapitres plus théoriques analysant l'existant, nous verrons en quoi les différentes parties de l'environnement tentent de répondre aux questions posées plus haut.

Le premier chapitre traitera de l'acquisition lexicale. Après avoir examiné le résultat des travaux en psycholinguistique sur le lexique mental, nous étudierons la modélisation de l'acquisition lexicale qui se définit par l'intégration de nouvelles connaissances au sein des anciennes. L'apprentissage lexical en langue seconde étant un processus lent et graduel, nous verrons le rôle important de la révision et du renforcement des traces mémorielles. Pour finir, nous étudierons l'importance de la compréhension de mots nouveaux dans les textes grâce aux stratégies d'inférence.

Le deuxième chapitre traitera des environnements informatiques d'aide à l'apprentissage lexical existants. Nous verrons qu'après les programmes de première génération limités par la technologie mais dont certains principes de conception étaient porteurs d'idées pertinentes, sont apparus des systèmes plus performants, intégrant des ressources lexicales privilégiant l'autonomie, et disposant de véritables programmes pédagogiques. Nous étudierons notamment le système Lexica (Goodfellow 1995), dont certaines idées seront reprises dans ALEXIA.

Le troisième chapitre se focalisera sur les rapports qu'entretiennent les apprenants avec les dictionnaires et sur l'utilité de ces derniers pour l'apprentissage. Nous verrons la manière dont les dictionnaires anglais d'apprentissage sur support papier tentent de résoudre les problèmes d'accès lexical et d'aide à la compréhension et à la production, ainsi que les caractéristiques

de leur équivalents électroniques qui sont loin de tenir compte des possibilités actuelles de l'informatique.

Le quatrième chapitre évoquera les activités lexicales que nous comptons intégrer dans ALEXIA. Il s'agira de sélectionner parmi celles qui existent déjà les activités à même de tirer parti des ressources dont nous disposons. Le choix sera donc guidé à la fois par des questions d'ordre pédagogique et par des questions d'ordre computationnel : une des originalités de ce travail étant de pouvoir générer automatiquement des activités qui n'existent pour l'instant que sur le papier.

Après ces quatre premiers chapitres, nous décrirons peu à peu l'environnement ALEXIA. Ainsi, le cinquième chapitre traitera d'une part de la modélisation de la base de données lexicales du système, organisée sous forme de réseaux lexicaux et tentant de ce fait de reproduire l'organisation du lexique mental, et d'autre part, de la présentation des connaissances lexicales ainsi que de leur accès et leur compréhension. Nous verrons en quoi l'informatique se différencie du support papier.

La rupture avec le papier sera consommée dans le sixième chapitre au cours duquel, après avoir discuté de l'influence du multimédia sur l'apprentissage lexical et des représentations visuelles existantes, nous exposerons la génération automatique de réseaux lexicaux graphiques (ou carte sémantique). Nous décrirons les spécificités de ces représentations et de l'interaction avec l'apprenant.

Les septième et huitième chapitres traiteront de l'implémentation des activités lexicales. Tandis que le premier insistera sur la préparation du corpus et des informations qu'il doit recevoir (étiquetage morphologique, indexation) pour la génération de concordances (activités de recontextualisation), le deuxième abordera les problèmes liés à la réalisation des activités (sélection des mots d'après les informations du dictionnaire personnel, lequel sera à ce moment décrit, gestion du contexte pour l'inférence, aides, etc.).

Enfin, le dernier chapitre récapitulera l'ensemble du système et décrira deux dernières parties : le corpus utilisé et le modèle de l'apprenant.

Avant d'aborder le premier chapitre, nous précisons la terminologie, empruntée en partie à Mel'cuk (1995), que nous allons employer dès à présent qui concerne la notion de mot, concept flou présentant plusieurs visages selon les cas. En effet, outre le problème de la segmentation d'une suite de lettres en mot (qui peut comprendre un ou plusieurs blancs : collocations, expressions semi-figées), l'incertitude est aussi sémantique. Lorsqu'on évoque un mot, parle-t-on d'une suite de lettres, d'homonymes, d'une acception ? Aussi, nous emploierons la terminologie suivante :

- La graphie, qui désigne toute suite de lettres segmentée en une unité lexicale, soit un mot tel qu'on le perçoit intuitivement sans plus de détail.
- Le vocable, qui désigne le mot polysémique (ou monosémique) que l'on peut considérer comme la somme de ses acceptions.
- La lexie, qui désigne une acception ou un sens particulier d'un vocable.

CHAPITRE 1

L’acquisition lexicale

Avant d’étudier l’acquisition lexicale proprement dite, nous allons examiner l’organisation et la structuration du lexique mental à partir des résultats qu’ont livrés les recherches en psycholinguistique. Il est nécessaire de savoir la manière dont sont stockés les vocables dans la mémoire afin de déterminer comment l’incorporation de nouveaux vocables, c’est-à-dire l’acquisition, s’effectue.

Nous passerons ensuite en revue les principales caractéristiques du processus d’acquisition lexicale avant de s’attarder sur l’apprentissage à partir d’un contexte écrit, c’est-à-dire lors de la lecture d’un texte en langue étrangère. Ces principes guideront l’élaboration de l’environnement ALEXIA.

1 Le lexique mental

Les recherches menées en psycholinguistique nous éclairent maintenant un peu mieux sur la manière dont les vocables pourraient être stockés dans la mémoire de chaque individu.

1.1 Les performances du lexique mental

Comment les vocables sont-ils stockés dans notre mémoire ? Sont-ils disposés en vrac, au hasard, sans aucun lien entre eux ? Certainement pas, et ceci pour deux raisons.

La première tient au nombre de vocables dont dispose chaque être humain. Les estimations divergent, car il est difficile d’évaluer la quantité de vocables que connaît une personne. Bogaards (1994, pp. 69) estime qu’un locuteur moyen pourrait connaître près de 35 000 vocables; les études relevées par Aitchison (1987, pp. 7) situent le vocabulaire d’un adulte éduqué entre 50 000 et 250 000 vocables. Nous n’entrerons pas dans le débat mais nous retiendrons que, dans tous les cas, la quantité de vocables connus est considérable et suppose un classement performant et systématique des éléments du lexique mental. Un empilement en vrac dans la mémoire ne pourrait pas expliquer les performances étonnantes de chacun en matière de vitesse de reconnaissance et de production des vocables.

En effet, et ceci est la deuxième raison, d'après Marslen-Wilson et Tyler (1980 et 1981), il ne faut qu'une fraction de seconde (200 ms) en langue maternelle pour reconnaître un vocable après l'avoir entendu. Dans beaucoup de cas, il est reconnu avant la fin de sa prononciation. De fait, quand on parle, on prononce environ de 150 à 300 vocables à la minute suivant le rythme d'élocution, soit un toutes les 200 à 400 ms. De plus, les erreurs lexicales sont très faibles (une sur 2 000 vocable prononcés, Levelt, 1989). Lors d'une lecture, le rythme de reconnaissance peut être encore plus élevé.

Mais le plus étonnant est la rapidité avec laquelle on arrive à reconnaître les vocables inexistants. D'après Marslen-Wilson et Tyler (1980 et 1981), une personne est capable de rejeter une séquence de son ne représentant aucun vocable en 450 ms, c'est-à-dire qu'il nous faut moins d'une demi-seconde pour décider si ce qu'on entend nous est connu ou pas. Pareille efficacité ne peut reposer que sur des procédures de recherche particulièrement performantes et sur une organisation très élaborée du lexique mental. En effet, pour affirmer sans erreur que l'on ne connaît pas tel vocable, il faut soit parcourir d'une manière ou d'une autre l'ensemble du lexique mental, soit utiliser ses structures pour ne pas tout parcourir tout en parvenant au même résultat.

1.2 Lexique mental et dictionnaire

Les êtres humains ont constitué depuis longtemps des ensembles de vocables destinés à être consultés : les dictionnaires. Leur organisation repose pratiquement toujours sur le classement par ordre alphabétique de ses éléments. Même si cette organisation nous semble *a priori* simple et les procédures de recherche pratiques, peut-on supposer une structure comparable du lexique mental ? La réponse se trouve dans un des moyens d'investigation des processus mentaux touchant au lexique : les anomalies ou lapsus. Si le rangement des items lexicaux était alphabétique, on s'attendrait à ce que les lapsus fassent apparaître des vocables proches alphabétiquement de ceux qui devraient normalement être produits. Or ce n'est pratiquement jamais le cas. On ne prononce jamais opinion à la place de, par exemple, opération, opérer, opportun, opposé, oppresser ou opticien.

Mais la différence entre dictionnaire et lexique mental ne s'arrête pas là. Elle est bien plus profonde.

En effet, on peut constater que les quantités d'informations de part et d'autre ne sont pas comparables. Le lexique mental contient de loin bien plus d'information que tout dictionnaire. Une foule de détails ne sont pas considérés car les dictionnaires sont inévitablement limités et ne peuvent pas contenir tous les détails possibles sur chaque vocable. Hudson (1984) remarque : « Il n'y a pas de limite à la quantité d'information détaillée… qui peut être associée à un item lexical. Les dictionnaires existants, même les plus gros, ne peuvent spécifier les items lexicaux que de manière incomplète. » Par exemple, on lit dans le PR que peindre signifie : « couvrir, colorer avec de la peinture. » Mais peint-on sa main lorsqu'on la pose sur une barrière fraîchement peinte ? Et même si la définition de peindre était de « couvrir intentionnellement de peinture », peint-on les poils du pinceau lorsqu'on les plonge dans le pot de peinture ? Car si l'artiste plonge son pinceau dans le pot, c'est bien pour le couvrir de

peinture qui plus tard lui servira à composer son tableau. Pourtant, personne ne dira que l'artiste peint son pinceau (Fodor, 1981). Les informations sur les collocations sont très faibles. Peut-on dire « un troupeau ou une harde de loups » ? Et qu'en est-il du lait « rance » ? De même pour les informations syntaxiques : large et principal sont tous deux adjectifs. Mais si l'on peut dire « la route est large », « la route est principale » n'est pas correct (Aitchison, 1987, pp. 13). De même pour la phonétique qui ne tient pas compte des accents dans différentes régions d'un même pays, etc., etc.

Une dernière différence essentielle est que le lexique mental n'est pas figé mais en constante évolution. Non seulement une personne peut inventer des vocables pendant qu'elle est en train de parler, mais peut aussi donner de nouveaux sens à une lexie. Ce qui n'empêchera pas pour autant la compréhension.

1.3 L'organisation du lexique mental

Si la disposition des vocables du lexique mental ne suit pas l'ordre alphabétique, quelle est son organisation ?

On observe en premier lieu qu'ils ne sont pas indépendants les uns des autres. Il est parfois difficile de se représenter un vocable sans penser à ceux qui l'entourent : peut-on penser à tiède sans penser à froid ou chaud ? Mais dans ce cas, comment sont-ils liés les uns aux autres ?

Il ne faut pas penser que les vocables d'une langue couvrent la réalité d'une manière régulière, bien que différente d'une langue à l'autre, telles les pièces d'un « puzzle » qui s'emboîtent les unes dans les autres et qui se conçoivent les unes par rapport aux autres. Les choses ne sont pas si simples car il peut y avoir plusieurs vocables pour exprimer une notion (léopard et panthère) tandis que d'autres concepts ne sont pas lexicalisés (comment appelle-t-on une plante morte ?) Il y a parfois recouvrement de sens lorsque plusieurs vocables ont un ou plusieurs traits en commun (mère, brebis, poule ou princesse sont des êtres de sexe féminin).

Beaucoup de modèles essayant d'expliquer ces cohabitations des vocables dans le lexique mental ont été proposés mais l'ensemble converge vers deux grands types de théories. Il y a d'une part les « atomic globule theories » et d'autre part les « cobweb theories ». Les premières affirment que les vocables sont construits à partir d'un ensemble commun d'« atomes de sens » (en fait de primitives sémantiques) et que les vocables reliés possèdent plusieurs atomes en commun. Les secondes pensent que si les vocables sont reliés entre eux, c'est à cause de l'existence de liens créés par les locuteurs. D'un côté, les vocables sont vus comme un assemblage de morceaux élémentaires, de l'autre ils sont considérés à part entière avec leurs caractéristiques et formant un réseau (théories des toiles verbales). Même si le consensus n'est pas total, les chercheurs se tournent désormais davantage vers la deuxième type de théories, car l'association de vocables dans la mémoire a pu être mise en évidence tandis qu'aucune expérimentation n'a montré de façon concluante l'existence des primitives sémantiques. Voyons donc ce que nous enseignent les théories des toiles verbales (Aitchison, 1987, pp. 72-85).

1.4 Les théories des toiles verbales

Dans cette hypothèse, le lexique mental est un vaste réseau, une toile verbale, dans lequel les nœuds sont les items lexicaux reliés entre eux par des chemins.

Les premiers travaux sur les réseaux sémantiques ont montré que les liens entre les vocables se formaient par l'habitude. Lors des expériences sur les associations des vocables, il apparaît que les réponses sont fortement conventionnelles : ainsi le premier vocable évoqué par marteau est très souvent clou (plus de la moitié des réponses), haut est évoqué par bas et noir par blanc. Les vocables ainsi obtenus appartiennent presque toujours au même champ sémantique. Ainsi, aiguille ne suggère jamais clou bien que tous deux soient des objets pointus, mais plutôt des vocables reliés à la couture (fil, épingle, trou et coudre). De même, on obtient assez souvent le deuxième membre d'une paire (mari et femme) ou de vocables en opposition (grand et petit). Malheureusement, on ne peut pas dresser ainsi la carte du lexique mental, car, d'une part, l'évocation explicite d'un vocable par un autre n'est pas une manière naturelle de fonctionner et d'autre part, les évocations sont fortement liées au contexte. Lune tout seul évoque soleil, nuit ou étoile mais en compagnie d'éléphant ou baleine évoque plutôt grand (Coleman, 1964). Il est donc illusoire de vouloir dresser une carte fixée des liens entre les vocables lorsqu'ils sont si influençables par le contexte.

En fait, à partir des réponses données aux tests d'associations, on établit que les liens peuvent être principalement de quatre types (classés par fréquence de réponse, les plus courants en premier) :

- entre vocables co-occurrents, qui sont les vocables apparaissant le plus souvent dans les réponses aux tests d'association. Ils appartiennent aux mêmes champs sémantiques avec le même niveau de détail (sel et poivre ; papillon et mite ; rouge, blanc, bleu, etc.),
- entre les membres d'une collocation, apparaissant souvent ensemble dans des expressions plus ou moins figées (eau et salé ; bleu et marine ; etc.),
- entre hyponyme et hyperonyme (papillon et insecte ; rouge et couleur ; etc.),
- entre synonymes, plus rarement (léopard et panthère).

Bogaards (1994, pp. 71), quant à lui, classe ces liens en deux groupes : d'une part les relations intrinsèques (ou catégorielles) au lexique (hyponymie et synonymie) et d'autre part les relations associatives (co-occurrence et collocation). Par ailleurs, il fait remarquer justement que ce ne sont pas véritablement les vocables qui sont liés entre eux mais leurs lexies, c'est-à-dire des éléments ayant une unité certaine au niveau sémantique. « Ainsi, ce n'est pas le mot (vocable) rayon qui fait partie d'un réseau lexical donné, en raison même de ses sens multiples. Ce sont plutôt les différentes unités lexicales (lexies) ayant la forme rayon qui entretiennent des relations multiples et participent à autant de réseaux différents : les toiles verbales sont organisées selon des critères exclusivement sémantiques. Par conséquent, le rayon du soleil appartiendra à un autre réseau que le rayon d'une bibliothèque ou le rayon d'un grand magasin, mais la forme rayon donne accès à tous ces réseaux » (Bogaards, 1994, pp. 74).

A partir de ces observations, peut-on connaître la manière dont ces différents liens coexistent dans le lexique mental ? Certains liens sont-ils plus forts ou durent-ils davantage que d'autres ?

D'après certains lapsus (Aitchison, 1987, pp. 76),

He's a born *sailure* (success + failure)[1]

il apparaît que les vocables sont bien stockés par champs sémantiques et que les co-occurrents sont fortement associés. Ils sont si liés que les aphasiques (personnes ayant des troubles de langages) ont du mal à les dissocier. Par ailleurs, toujours chez les aphasiques, ce sont les liens qui résistent le mieux à des lésions du cerveau. Même si, d'après Hotopf (1980) on peut classer les relations entre co-occurrents dans trois catégories (contrastifs : pomme et poire ; opposés : haut et bas ; cousins sémantiques : dimanche et janvier, qui appartiennent au champ des dates mais dans différents types), il est cependant difficile d'être précis sur l'organisation détaillée des co-occurrents dans le lexique mental, car la structure des groupes est fortement dépendante des vocables eux-mêmes et non généralisable.

Les liens parmi les membres des collocations sont aussi assez forts. Ils persistent avec l'âge (Howard *et al*., 1981) et en dépit de lésions du cerveau chez les aphasiques (Goodglass et Baker, 1976). Les collocations, expressions plus ou moins figées et plus ou moins transparentes quant à leur sens, semblent être traitées en bloc, comme tout autre item lexical (Swinney et Cutler, 1979 ; Cutler, 1983 ; Gibbs et Gonzales, 1985 ; Schweigert et Moates, 1988). Elles ne sont donc pas décomposées ou recomposées lors d'une tâche de compréhension ou de production. Ceci est un point important concernant leur statut dans les dictionnaires et dans la manière de les apprendre. Elles sont très nombreuses (entre 300 et 400 000 selon le LADL, principalement des noms composés).

En revanche, les liens entre hyponymes et hyperonymes semblent plus faibles. Ils apparaissent peu souvent dans les lapsus (Hotopf, 1980) même si les amalgames sont plus fréquents chez les aphasiques. D'autre part, ils sont relatifs à la notion de prototypie. Le lien sera plus fort avec un hyponyme qui correspond bien au prototype de l'hyperonyme. En effet, on dira plus facilement qu'un moineau est un oiseau plutôt qu'un pélican est un oiseau (pour beaucoup de monde le moineau est l'oiseau le plus représentatif).

Vers la fin des années 60, de nombreux chercheurs pensaient que le lexique mental était organisé par une structure hiérarchique (par exemple Collins et Quillian, 1969). Ainsi pour les animaux, on avait la structure suivante : animal – oiseaux, insectes, poissons, etc. – différents types d'oiseaux (moineau, canari, hirondelle, cigogne, chouette, etc.), différents types d'insectes, de poissons, etc. et ainsi de suite. Cette hypothèse était confirmée par les temps de réponses plus courts à la question « un canari est-il un oiseau » ? qu'à la question « un canari est-il un animal » ? (il y a en effet un niveau de plus à parcourir pour cette dernière question).

Cependant, ce résultat n'est pas vraiment concluant sur l'organisation hiérarchique du lexique mental car il aurait pu être obtenu pour d'autres raisons. En effet, il pourrait s'agir tout simplement d'une association plus forte entre canari et oiseau qu'entre canari et animal, car plus fréquente dans la langue. D'autre part, il y a bien plus d'animaux possibles que d'oiseaux

[1] La première lettre de failure est remplacée par la première de success, failure et success étant fortement reliés.

possibles et la multiplicité des choix pourrait expliquer un temps de réponses plus long. Mais la preuve la plus convaincante surgit lorsqu'on fixe non pas l'hyponyme le plus bas (le canari) mais l'hyperonyme le plus haut (l'animal). Ainsi aucune différence dans la vitesse de traitement n'est constatée lorsqu'on demande si « un chien est un animal » et lorsqu'on demande si « un caniche est un animal ». Ce résultat suggère que même si les gens organisent les vocables dans des bouquets, il est peu probable qu'ils montent et descendent dans les arbres lexicaux comme s'il s'agissait d'échafaudages (Aitchison, 1987, pp. 81).

Pour finir sur les différents types d'associations, les liens entre synonymes semblent eux aussi plus faibles que les deux premiers types vus plus haut. Ils se retrouvent dans les amalgames et surgissent dans des cas bien précis, par exemple lorsque les deux vocables auraient pu convenir (that's torrible, terrible + horrible) car signifiant, dans ce contexte, la même chose. Les lapsus entre vocables partiellement synonymes (ce qui est le cas le plus fréquent) sont plus rares et ne se produisent véritablement que lorsque les deux vocables appartiennent à des champs sémantiques différents (comme dans « the inhabitants of the car were unhurt » (inhabitants est mis à la place d'occupants)). Les lapsus résultent plutôt du choix du vocable de remplacement lorsque la recherche du vocable convenable a échoué.

Pour résumer, nous dirons donc que l'examen des associations obtenues dans les tests nous amène à penser que les vocables sont principalement rangés en champs sémantiques et liés entre eux par des relations plus ou moins fortes suivant leur nature. Ces champs seraient d'ailleurs plus ou moins indépendants les uns des autres car il est apparu, au cours d'aphasies, que certains pouvaient être atteints mais en laissant les autres intacts. Un autre argument en faveur des champs sémantiques se constate lors de l'apprentissage d'une liste de vocables. Si, parmi ces vocables, certains peuvent être regroupés au sein d'un même champ, la mémoire les traitera de la même manière : soit le sujet retient tous les vocables de ce groupe, soit il les oublie tous (Bogaards, 1994, pp. 70).

Pour finir avec l'étude du lexique mental, nous allons examiner l'influence des catégories grammaticales dans le stockage des vocables ainsi que le cas des dérivés syntaxiques.

1.5 Les catégories grammaticales

L'étude des lapsus montre que très fréquemment un vocable est remplacé par un autre de même catégorie grammaticale. Ainsi 90 % des noms et verbes et 60 % des adjectifs retiennent leur catégorie grammaticale (Browman, 1978). Ce résultat est confirmé par les réponses aux tests d'associations dans lesquels 80 % des noms évoquent des noms tandis que les verbes et adjectifs le font une fois sur deux (Deese, 1965). D'autre part, les aphasiques ont tendance à oublier surtout les verbes (Allport et Funnell, 1981 ; Hand *et al.*, 1979). Cela peut s'expliquer par le fait que les verbes contiennent beaucoup plus d'informations d'ordre syntaxique (nature du sujet, éventuellement de l'objet, de prépositions, etc.) que les noms et adjectifs. Ces constatations tendent à prouver que les vocables sont stockés de manière hermétique en fonction de leur catégorie grammaticale et cela même en cas d'homonymie (comme pour boucher, à la fois nom et verbe).

Sémantique et syntaxe sont donc quelque peu liées. En effet, tandis que la majorité des noms fait surtout référence à des choses (abstraites ou concrètes) et des personnes, les verbes désignent surtout des actions ou des états. Les vocables doivent donc être considérés comme des pièces de monnaie dont une face contient les informations phonétiques et l'autre le sens et la catégorie grammaticale qui seront, dès lors, indissociables.

Il nous reste à examiner un dernier aspect du lexique mental : comment le lexique mental traite l'architecture interne des vocables, suivant qu'ils comportent oui ou non des préfixes ou des suffixes.

1.6 L'architecture interne des vocables

On peut distinguer deux sortes de vocables : les indécomposables (éviter, forcer, perdre) et les dérivés formés à l'aide de préfixes et suffixes (in-évit-able, forcé-ment, perdi-tion). La question se pose alors de savoir comment le lexique mental organise ces deux types de vocables. Sont-ils stockés en bloc, non décomposés, ce qui augmente alors la masse de vocables dans la mémoire, ou alors sont-ils stockés fragmentés et recomposés lorsque nécessaire à l'aide de règles simples, ce qui augmente alors la charge cognitive ? Il ne faut pas oublier les marques de flexion qui ajoutent à la complexité du vocable formé (maison-s, chante-rai).

Concernant la flexion, l'étude des lapsus montre clairement que les marques de flexion ne sont pas stockées avec le vocable mais ajoutées dans le feu du discours (He go backs au lieu de He goes back ; They point outed au lieu de They pointed out). Cette conclusion est renforcée par d'autres observations : les lapsus conduisent à la formation de vocables inexistants (outed), certains verbes irréguliers sont conjugués comme s'ils étaient réguliers (fighted au lieu de fought), lors d'ambiguïtés sur la marque du pluriel certains oscillent entre les deux formes possibles (mothers-in-law ou mother-in-laws), etc. D'autre part, il semblerait que l'exposition préalable à des formes fléchies pour des verbes réguliers (jumps pour jump) accélèrerait les temps de reconnaissance du vocable de la même manière qu'une exposition préalable du vocable lui-même non fléchi (Murrell et Morton, 1974 ; Stanners *et al*., 1979). Ce qui n'est pas le cas pour les formes fléchies irrégulières (hung et hang). La conclusion est donc que les marques de flexion sont ajoutées lorsqu'on parle même si, l'usage aidant, certaines formes fléchies sont stockées directement dans la mémoire (lèvres, courses).

Les vocables comportant des préfixes sembleraient, si l'on se fie à l'expérience de Taft et Forster (1975), être stockés sans leur préfixe. Mais cette expérience est plutôt infirmée par les objections des partisans des vocables « prêts à l'emploi ». En effet, il apparaît que dans beaucoup de cas, les préfixes et les racines ne sont pas cohérents. Quel est le lien sémantique commun dans le préfixe con- attaché à consumer, conférer, concevoir, condamner, etc. et celui de la racine –férer dans des verbes tels que conférer, déférer, inférer, préférer, référer, etc. à moins de connaître le latin, ce qui n'est pas inné ? Si les vocables étaient stockés de manière fragmentée, cela serait source de difficulté et de confusion pour pouvoir les recomposer d'après leur préfixe et leur racine. Des arguments supplémentaires proviennent des lapsus (Aitchison, 1983). Les vocables préfixés préservent leur début aussi bien que ceux qui ne le sont pas et d'autre part, ils sont interchangés avec ceux qui ne le sont pas, preuve qu'ils ne

forment pas une catégorie spéciale. Il semble donc qu'à la lueur de ces constatations les vocables préfixés soient bien stockés dans la mémoire en tant que tels et non décomposés en préfixe et racine.

Des résultats similaires sont trouvés pour les suffixes. Même si une certaine régularité apparaît dans la dérivation de vocables tels que collision (collide), division (divide), interference (interfer), l'ajout de suffixes connaît aussi une certaine part d'imprévisibilité : certains vocables sont sans base (perdition, conflagration, probity), la dérivation est irrégulière (comprehend/comprehensive ; revolve/revolution ; succeed/succession) quand elle est lexicalisée (induce se dérive en inducement et induction, mais si on a production à partir de produce, *producement n'existe pas). En français, *allocateur n'existe pas bien que son sens soit compréhensible. Les lapsus conduisent aux mêmes conclusions : par exemple les suffixes sont conservés même si la base est erronée (provincial à la place de provisional). Donc, tout comme les vocables préfixés, les vocables suffixés sont conservés et organisés dans la mémoire de manière non parcellaire.

Ce constat n'exclut pas la possibilité de décomposition en morphèmes. En effet, il y a, derrière le lexique central, une mémoire auxiliaire qui contient les formes analysables (dé-voile-ment, re-model-er). En outre, il existe aussi une boîte à outils lexicaux, un ensemble de règles morphologiques qui permet aussi bien d'analyser des formations nouvelles que de former soi-même des vocables nouveaux (Aitchison, 1987, pp. 161).

Voyons maintenant les conséquences de ces résultats dans le processus de l'apprentissage lexical.

2 L'apprentissage lexical

2.1 Processus d'apprentissage

Selon Tréville et Duquette (1996), le vocabulaire comporte deux aspects, qui correspondent à deux niveaux de traitement. Le premier concerne la forme des vocables (composante phonétique et graphique) tandis que le deuxième a rapport à leur sens (aspect sémantique). Ces aspects ne mettent pas en jeu les mêmes capacités cognitives. Le premier est en effet plus superficiel et intervient en premier dans l'acquisition d'une langue. Ce sont les indices phonétiques et graphiques qui déclenchent le processus d'acquisition des vocables en langue maternelle. Ils sont surtout de type perceptuel chez l'enfant. Par contre, chez l'adulte, l'acquisition se produit par l'intégration d'indices sémantiques dans les réseaux du lexique mental. Lors de l'apprentissage d'une langue seconde, les deux aspects sont présents dès le départ, mais si l'intégration des vocables nouveaux se produit plutôt suivant des critères de forme au début, elle laisse place peu à peu à des associations plus profondes de nature sémantique au fur et à mesure que la compétence se développe (Singleton, 1994).

C'est la tâche d'acquisition qui fixe le niveau de traitement. Une tâche de répétition, qui ne fait intervenir que la forme des vocables ne produira qu'une fixation superficielle dans la mémoire à long terme. Au contraire, une tâche de raisonnement ou de comparaison détaillée agira plus profondément et impliquera l'intégration du vocable dans divers réseaux mentaux de l'apprenant (syntagmatique, paradigmatique, hyponymique, etc.). L'acquisition se produit

lorsque les vocables nouveaux sont incorporés aux réseaux concernant la forme des vocables et aux réseaux sémantiques, lorsque les connaissances nouvelles véhiculées se fondent et s'associent aux anciennes, et cela principalement à un niveau sémantique.

La difficulté des tâches est aussi un facteur décidant de la qualité du processus d'acquisition. Ainsi, Jacoby *et al.* (1979), faisant effectuer des tâches de difficultés diverses à des apprenants (il leur était demandé soit de copier des vocables, soit de contrôler et si nécessaire de corriger l'orthographe ; dans une autre tâche, ils devaient indiquer lequel de deux vocables était associé le plus étroitement à un vocable donné), parviennent à la conclusion que les tâches difficiles mènent à des traces mémorielles mieux établies que les tâches faciles. Plus la tâche initiale est complexe, plus l'enregistrement dans la mémoire qui en découle sera riche, détaillé, et précis. L'enregistrement d'un vocable n'est pas un phénomène ponctuel et définitif, mis en place une fois pour toute. Il doit être réactualisé pour subsister. Or plus la trace mémorielle est riche et précise, plus elle a de chances d'être retrouvée, réutilisée et, par ce fait même, renforcée (Bogaards, 1994, pp. 93).

Du reste, dans cette même expérience, Jacoby *et al.* notent que, bien que les vocables n'aient pas été appris explicitement (les étudiants n'avaient pas été prévenus qu'ils allaient être réinterrogés), ils étaient capables soit de se les rappeler, soit de les reconnaître dans diverses listes de vocabulaire, mélangés à d'autres vocables. On peut donc parler d'apprentissage fortuit et non-intentionnel. Une autre expérience, menée par Wilson et Bransford et rapportée par Gairns et Redman (1986), montre que l'intention d'apprendre ne mène pas forcément au meilleur résultat et que les tâches significatives, c'est-à-dire qui signifient quelque chose pour l'apprenant et où celui-ci est impliqué personnellement, provoquent un apprentissage bien plus efficace. Le sens est un élément prépondérant dans l'acquisition, car l'esprit humain fonctionne avec des catégories significatives, avec des symboles ayant un contenu sémantique et non pas avec des signaux dépourvus de sens (Lakoff, 1987).

L'acquisition lexicale est un processus graduel et lent. C'est sous l'effet de la répétition et de la manipulation mentale du vocabulaire que les associations se mettent en place à des rythmes divers. Pour Meara (1989), « les vocables connus correctement une semaine peuvent être complètement oubliés la semaine suivante ; les vocables entraînant une forme particulière d'erreur une semaine donnée peuvent causer un type d'erreur totalement différent la semaine d'après. » Pour fixer l'item lexical dans la mémoire, certains chercheurs (Oxford et Crookall, 1990) préconisent la révision structurée. Il s'agit de se doter d'un « planning » de révision, sachant qu'un vocable nouveau doit être vu 6 à 10 fois avant d'être mémorisé (les vocables sont revus dans le temps à intervalles de durée croissante). Toutefois, l'acquisition lexicale n'est pas uniquement affaire de répétition, même si celle-ci finit toujours par produire un effet. D'autres facteurs entrent en jeu tels que la motivation personnelle et les besoins individuels. En outre, chaque apprenant dispose dans son propre lexique mental de divers réseaux, qui ne sont pas tous structurés et employés de la même manière. Un item lexical ne sera pas intégré dans le même environnement mental suivant le locuteur. La création d'associations ou la constitution de « toiles verbales » est donc une entreprise hautement individualisée. « Chacun construit, au cours de son histoire personnelle et avec ses accents

individuels, le lexique qui lui convient, avec les connotations et les images qui sont propres à chaque individu et au contexte socioculturel où il vit » (Bogaards, 1994, pp. 97).

2.2 Compréhension et acquisition

L'un des moyens naturels d'exposition à de nouveaux vocables est la lecture et l'écoute (textes écrits ou oraux). À ce titre, la compréhension joue un rôle capital. Sans compréhension, comme nous l'avons vu plus haut, l'acquisition est fortement limitée. Selon Tréville et Duquette (1996), « le processus de la compréhension peut se définir comme l'interaction entre les connaissances antérieures et les connaissances nouvelles. Il y a compréhension quand l'individu peut rendre significatif l'apport langagier (connaissances nouvelles), c'est-à-dire quand il peut établir un lien entre l'acquis récent (vocabulaire et règles lexicales par exemple) et l'acquis déjà ancré dans la mémoire à long terme (mémoire sémantique). »

De ce fait, lors de la confrontation au texte, la compréhension n'est possible que si une certaine proportion d'éléments lexicaux est connue. D'autres éléments que le vocabulaire rentrent en compte tels que, pour les textes écrits, deviner des vocables inconnus, faire des inférences, reconnaître le type du texte et sa structure, trouver l'idée principale d'un paragraphe. Pourtant, il a été démontré que la compréhension écrite était fortement tributaire du vocabulaire, bien plus que des autres facteurs ci-dessus (Anderson et Freebody, 1981 ; Beck, Perfetti et McKeown, 1982 ; Kameenui, Carnine et Freschi, 1982 et Stahl, 1983). Par ailleurs, Laufer (1991) a trouvé des corrélations significatives entre deux tests de vocabulaire (*Vocabulary Levels Test* par Nation (1983a) et *Eurocentres Vocabulary Test* par Meara et Jones (1989)) et les performances de lecture d'apprenants d'une langue seconde.

En revanche, la difficulté syntaxique n'affecte pas la compréhension écrite (Ulijn et Strother, 1990). De même les stratégies de lecture, même si celles-ci sont performantes et transposées à partir de la langue maternelle, ne sont véritablement utiles qu'au-delà d'un certain seuil de connaissance de la langue cible (Alderson, 1984 ; Perkins, Brutten et Pohlmann, 1989).

Comme nous venons de le voir plus haut, la nature de ce seuil est essentiellement lexicale. D'après Haynes et Baker (1993), le handicap le plus significatif n'est pas le manque de stratégie de lecture mais un vocabulaire insuffisant. Des recherches ont été menées afin de déterminer quantitativement la nature du seuil lexical de compréhension. En d'autres termes, combien faut-il connaître de vocables pour pouvoir lire et comprendre un texte authentique, c'est-à-dire pour appliquer les stratégies de lecture efficaces avec succès ? Par vocable, nous adoptons ici la notion de famille de mots (Nation 1983b), c'est-à-dire le vocable lui-même, accompagné de ses dérivés syntaxiques (travail, travailler, travailleur). D'autre part, nous supposons ici qu'un vocable est connu, du moins pour une tâche de compréhension, lorsqu'une personne est capable de reconnaître ses formes et ses significations les plus courantes de manière automatique, sans réel effort et indépendamment du contexte. D'après Laufer (1991), les apprenants doivent connaître environ 3 000 familles lexicales (ou 3 000*1,6 = 5 000 vocables, Nation 1983a) pour réussir avec le minimum requis (un résultat de 56 %) au test de lecture de

leur institution. On obtient ensuite une progression linéaire de 7 % pour chaque millier de familles de mots supplémentaires connues, c'est-à-dire qu'en connaître 4 000 (6 400 vocables) amène un résultat de 63 %, 5 000, 70 %, etc. jusqu'à un niveau où la progression s'estompe peu à peu. Même si d'autres études avec d'autres tests mènent à des résultats différents, la barre des 3 000 familles semble déterminante. En deçà de cette quantité, les performances aux tests de lecture sont plutôt faibles. En résumé, il faut donc connaître 5 000 vocables pour pouvoir transférer avec succès les stratégies de lecture de sa langue maternelle à la langue cible. En termes de fréquence et de couverture de textes, 5 000 vocables couvrent 90 à 95 % de n'importe quel texte (Nation, 1983a et 1990).

Pour autant, les problèmes lexicaux en compréhension ne se résument pas à une non-familiarité avec les vocables du texte. Il faut bien noter que les vocables ne sont pas égaux par rapport à leur apprentissage, c'est-à-dire que certains sont plus faciles à apprendre et à retenir que d'autres. Laufer (1997) précise qu'à côté des vocables que l'on ne connaît pas, il y a ceux que l'on pense connaître mais qui sont en fait mal ou pas compris, tels les faux amis.

2.3 Vocables faciles et vocables difficiles

Parmi les vocables plus faciles à apprendre, il y a ceux qui sont apparentés d'une langue à une autre et que l'on appelle les congénères (*cognates* en anglais). Ils sont généralement définis comme des vocables de langues différentes ayant (à peu près) les mêmes formes et (à peu près) les mêmes sens (surtout les mêmes sens dans l'utilisation courante), comme régulier français et regular anglais (Bogaards, 1994, pp. 153). Les congénères jouent un rôle particulier dans l'apprentissage du vocabulaire car l'apprenant peut reconnaître des éléments qui semblent avoir un visage connu :

- connu par la forme (les lecteurs anglais d'un texte en français n'auront pas de mal pour reconnaître et attribuer un sens correct à obligation, piano et terrible), même s'il n'est pas aisé de savoir jusqu'à quelle limite un vocable étranger est familier : parfois une seule lettre suffit à obscurcir les liens de parenté entre deux vocables (titre/title) ; parfois, des divergences plus marquées n'empêche pas d'établir des liens comme pour avis/advice (Van Roey, 1990).
- connu lorsque l'apprenant est au courant de certaines correspondances plus ou moins régulières comme, par exemple, la terminaison des vocables. Savoir qu'aux terminaisons françaises -(i) té, -eur ou –ie correspondent les terminaisons anglaises –(i) ty, -or et –y élargit considérablement le champ des vocables reconnaissables (Hammer et Giauque, 1989).

Les congénères sont donc utiles pour comprendre des textes en langue étrangère. Cependant ils peuvent être aussi dangereux lorsque le lien d'appariement concerne la forme mais pas le sens. Il y a tout d'abord les vocables qui présentent des sens très différents comme coin en anglais et en français. Ceux-ci ne sont pas toutefois les plus dangereux car du fait de leur différence sémantique, ils ne se rencontreront pas dans les mêmes contextes, ce qui diminue les risques de confusion. Par contre, on ne peut en dire autant de ceux que Laufer (1997) appelle les vocables « faussement transparents » ou les faux amis.

Il s'agit de vocables dont la forme semble donner des indications sur le sens alors qu'il n'en est rien. Lorsque deux langues sont concernées, il s'agit de vocables dont les formes sont proches mais dont les sens ne sont pas complètement dissemblables. C'est le cas par exemple de nouvelle/novel (roman) pour l'anglais ou régaler/regalar (offrir) pour l'espagnol. Le fait qu'on puisse les rencontrer dans les mêmes contextes augmente les risques de confusion et de mauvaise interprétation.

D'autre part, il y a les pseudo-composés (structure morphologique trompeuse) comme outline en anglais ou recrue en français, les expressions idiomatiques dont le sens non transparent induit en erreur ou déroute le lecteur (bosser comme un fou, travailler au noir), les « synformes », vocables qui, dans la langue étrangère, ont des formes lexicales similaires (cute/acute, available/valuable, conceal/cancel, economic/economical pour l'anglais ou industriel/industrieux pour le français) et les vocables polysémiques dont un sens seulement présente une similarité avec un vocable de la langue étrangère (comme abstract, qui correspond à abstrait mais qui veut dire aussi résumé). Dans ce dernier cas, l'apprenant peut être induit en erreur s'il s'accroche, malgré le contexte non approprié, à l'unique sens qu'il connaît.

Laufer (1989) a étudié l'influence des vocables faussement transparents (FT) sur la compréhension des apprenants au cours d'une phase de lecture. Il en ressort que : 1) les erreurs de mauvaise compréhension provenaient plus fréquemment des vocables FT que d'autres vocables non-FT ; 2) les apprenants étaient moins au courant de leur mauvaise compréhension avec des vocables FT qu'avec d'autres vocables non-FT ; 3) il y avait une corrélation significative entre la compréhension écrite des apprenants et leur connaissance des vocables FT.

Lorsqu'un apprenant se heurte à un vocable inconnu dans un texte, il peut soit l'ignorer pour éviter l'interruption de la lecture (s'il estime qu'il n'est pas important et qu'il ne gêne pas la compréhension de la phrase ou du paragraphe), soit regarder dans un dictionnaire ou demander de l'aide à quelqu'un. Mais il peut aussi essayer de deviner, d'inférer son sens à partir du contexte. Or pour cela, il faut qu'il estime ne pas le connaître, d'où le rôle négatif des vocables faussement transparents.

Pour terminer cette partie sur les vocables faciles et difficiles, nous mentionnons le cas particulier, mais très fréquent, des unités polylexicales, groupes de mots possédant un sens propre (ex : petit boulot, bosser comme un fou). Ces unités posent problème aux apprenants, d'une part justement parce qu'elles sont constituées de plusieurs éléments, ce qui rend plus délicat leur repérage dans un texte, et d'autre part parce que leur sens est plus ou moins opaque suivant les cas. Enfin, elles ont un comportement syntaxique propre qui autorise ou pas, de manière imprévisible, des variations, des insertions ou tout autre transformation grammaticale. Elles peuvent donc être difficiles à comprendre et, du fait de leur idiomaticité, difficilement prédictibles en cas de production.

Nous appellerons co-occurrence un groupe de mots apparaissant fréquemment ensemble (ex : verser un salaire). En général, on peut faire varier au moins un des constituants sur l'axe paradigmatique (toucher, percevoir, recevoir un salaire ; toucher un salaire, une allocation, des revenus). Toutefois, cela n'est pas suffisant pour les considérer comme des unités polylexicales, car ces associations ne possèdent pas de sens propre. Néanmoins, elles doivent être considérées dans

l'apprentissage d'une langue seconde car elles ne se construisent pas au hasard mais sont dictées par la langue : ainsi on ne dira pas grand salaire mais gros salaire ou salaire élevé. La maîtrise lexicale d'une langue passe par la connaissance de la correction (ou non) de ces associations. Une collocation est une co-occurrence qui n'admet pas cette variation (sur l'axe paradigmatique) et qui est en quelque sorte consacrée par la langue (petit boulot est une collocation car ni boulot ni petit ne peuvent varier; petit travail, petit job ne sont pas des co-occurrences). Elle possède dès lors un sens propre. Dans les collocations, nous englobons les expressions semi-figées, ou expressions idiomatiques (bosser comme un fou, travailler au noir), dont le sens ne peut être déduit à partir des mots les constituant. Dans la suite de ce travail, pour faciliter la lecture, nous emploierons le terme de "collocation" pour désigner les collocations elles-mêmes, dont le sens est en général plutôt transparent, et les expressions semi-figées dont le sens est plus opaque. Les expression semi-figées se distinguent par le fait qu'on ne peut modifier certains constituants (elles admettent parfois des variations restreintes, d'où le terme de semi-figé) sans altérer leur signification ou leur fonction conventionnelle (Chanier *et al*., 1992). Ainsi prendre le taureau par les cornes n'a pas du tout le même sens que prendre la vache par les cornes.

L'intérêt des collocations dans l'apprentissage lexical d'une langue seconde est grand. En effet, elles sont fortement utilisées par les natifs à l'oral comme à l'écrit, et ne pas les employer revient à appauvrir de façon notable le contenu de son discours ainsi que la manière de transmettre ce discours (importance de la situation de communication). Leur maîtrise doit permettre à l'apprenant de ne pas violer certaines restrictions lexicales (ces expressions respectent la syntaxe générale de la langue, mais n'apparaissent que dans un nombre restreint de formes syntaxiques) et de lui éviter de commettre des erreurs de registre (ou lui donner accès à des registres plus variés) : on ne pourra pas employer bosser comme un fou dans n'importe quel contexte, mais plutôt dans un registre familier.

2.4 L'inférence

L'inférence est une stratégie pertinente et importante lors de la lecture d'un texte dans une langue étrangère. Elle s'effectue notamment à partir du contexte. Pour Sternberg (1987), « la plus grande partie du vocabulaire s'apprend par le contexte » et, pour Clarke et Nation (1980), « deviner les vocables à partir du contexte est une partie centrale du processus de lecture ». Haastrup (1989) affirme que les vocables appris par inférence sont mieux retenus parce qu'ils sont mieux insérés dans un réseau sémantique.

L'inférence consiste, à partir d'un vocable jugé inconnu, de trouver un sens approximatif en fonction du contexte et de vérifier son hypothèse à chaque nouvelle occurrence de ce vocable pour ajuster le sens en fonction des nouveaux éléments. Enfin, comme le note Hulstijn (1993), il est important à un moment donné de consulter une source de référence, tel un dictionnaire, pour s'assurer de la justesse de ses déductions. Cela permet en effet de valider, ou au contraire d'infirmer, tel ou tel type de raisonnement pour pouvoir l'appliquer (ou non) par la suite.

On ne sait pas trop quels sont les raisonnements qui amènent les apprenants à deviner le sens du vocable. Les chercheurs avancent des hypothèses sur la qualité du contexte. Il doit être simple et pertinent, et éveiller le réseau lexical du locuteur natif (Beheydt, 1987), il ne doit pas être trompeur, notamment par l'emploi de métaphores ou de vocables faussement transparents, il doit contenir des indices et ne pas trop comporter de vocables inconnus (Schouten-van Parreren, 1986 (dans Bogaards, 1994, pp. 174) ; Laufer, 1997). Pour Tréville et Duquette (1996), l'inférence lexicale est favorisée par quatre facteurs : la maturité langagière (l'étendue des connaissances déjà en place), la connaissance conceptuelle des vocables (la connaissance du vocable dans la langue maternelle facilitera son apprentissage), l'aptitude à classer les vocables suivant leur morphologie et leurs fonctions grammaticales et enfin l'exposition répétée aux vocables qui attire l'attention sur ceux-ci et en facilite la rétention. Par ailleurs, l'indice de synonymie (le fait qu'un vocable inconnu soit décelé comme synonyme d'un autre présent dans le contexte comme dans la phrase « Le X que j'ai vu, comme les autres loups d'ailleurs, était féroce ») serait le plus facilitateur pour le devinement. D'autre part, il semble surtout fructueux avec les vocables pleins, et probablement plus avec les noms qu'avec les autres classes de vocables.

En fonction de ces paramètres, il n'est pas évident que l'inférence soit toujours possible. Selon Laufer (1997), il est même optimiste de penser que tous les textes contiennent des indices contextuels. D'autre part, l'inférence est liée fortement, comme nous l'avons vu plus haut, à la densité de vocables inconnus. Une insuffisance de connaissance nuit aux déductions ou en génère de mauvaises. Même si l'inférence offre des perspectives intéressantes, elle est sujette aussi à des limitations et des restrictions.

3 Conclusion

Nous avons étudié dans ce chapitre les mécanismes de l'apprentissage lexical en nous intéressant en premier lieu à la structure du lexique mental. Nous avons vu que les performances étonnantes de notre mémoire n'étaient pas le fait du hasard mais d'une organisation particulièrement performante et souple. D'après les travaux en psycholinguistique, il ressort que le lexique mental semble composé de lexies reliées entre elles par des liens de nature sémantique et contextuelle. De plus, il apparaît que le sens et la catégorie grammaticale d'une lexie sont indissociables. Pour finir, un dernier résultat important concernant la morphologie et la dérivation est qu'il semble que les vocables soient stockés comme un tout à part entière et non pas décomposés en affixes et racines et recomposés lors de la compréhension ou de la production du discours.

L'apprentissage lexical peut se définir comme l'incorporation de nouvelles informations lexicales dans les anciennes. Cette incorporation est fonction du niveau de traitement du vocabulaire, un traitement en profondeur sur le sens des vocables favorisant l'apprentissage. Ce dernier n'est pas instantané mais se déroule dans le temps dans lequel les facteurs répétition et révision du vocabulaire jouent un rôle important. L'inférence du sens des vocables à partir du contexte environnant est un des moyens naturels pour apprendre le vocabulaire. Elle est facilitée ou au contraire empêchée par différents critères comme par

exemple la proportion de vocables inconnus dans le texte, la fréquence ou la présence de synonymes.

Nous verrons au chapitre 4 l'intérêt des activités lexicales pour l'acquisition.

Nous allons voir maintenant comment ces résultats ont été considérés dans les environnements informatiques d'aide à l'apprentissage lexical.

CHAPITRE 2

Les environnements informatiques d'aide à l'apprentissage de vocabulaire

D'après Goodfellow (1995), il y a deux générations de programmes d'aide à l'apprentissage du vocabulaire, qui correspondent en gros à ce qui existait avant et après l'apparition des interfaces graphiques. Les programmes ont bénéficié de la facilité d'utilisation que procurent ces interfaces ce qui a amélioré leur interactivité (manipulation plus pratique et plus intuitive) et les a rendus plus attrayants.

L'ajout de ressources linguistiques informatiques en ligne telles que des dictionnaires, des corpus, a permis de plus grandes possibilités et a permis de séparer le rôle de traitement de l'ordinateur (la manière dont il détermine ce que peut vraiment faire l'utilisateur) de ses ressources (ce qu'il offre à l'utilisateur pour l'aider à réaliser ses objectifs). Depuis, les améliorations techniques ont suscité l'intérêt des pédagogues.

1 Les programmes de première génération

Beaucoup de ces programmes ont prétendu enseigner le vocabulaire. En réalité, la plupart étaient des versions informatiques de jeux faisant intervenir des vocables tels que les anagrammes ou les mots-croisés. Ils permettaient de développer le vocabulaire de l'apprenant mais de façon fortuite (Kenning, 1990). En raison de l'interactivité limitée de ces programmes (pas d'interface graphique, il fallait taper les vocables en entier) et de la limitation des algorithmes (l'ordinateur stockait toutes les solutions), l'accent a surtout été mis sur les vocables courts et sur l'orthographe et non le sens. Ce qui semait le doute sur leur supériorité par rapport aux exercices papier. Par la suite, beaucoup a été fait pour rendre ces programmes plus attrayants et plus motivants (ajout de couleur, de sons, de graphiques, mise en place de systèmes de notes) mais, comme il a été noté (Clarke, 1992), ces programmes mettaient plus l'accent sur l'évaluation par des tests de ce qui avait déjà été appris et non pas sur le processus d'apprentissage lui-même.

Malgré les limitations de ces programmes de première génération, il y a eu quelques tentatives d'incorporer des principes pédagogiques s'attachant à la sélection du vocable-cible et au type de l'activité. Ces tentatives se sont appuyées sur la pédagogie de la langue, la psychologie et les études en acquisition d'une langue seconde.

1.1 Sélection des vocables-cible

Certains programmes se sont efforcés d'individualiser le contenu lexical à apprendre en construisant de grandes bases de données lexicales à partir desquelles les apprenants (ou les enseignants) peuvent sélectionner des blocs en fonction de la fréquence ou du sujet. Dans Wordchip (Decoo, 1993), l'ordinateur aide à constituer des blocs de vocabulaire qui sont destinés à être appris en plusieurs années lors de cours de langues non-intensifs. Les blocs peuvent être confectionnés suivant plusieurs critères : la fréquence, le regroupement sémantique, les catégories grammaticales, la longueur des vocables, la distance interlangue (vocables plus ou moins différents dans la langue maternelle et seconde), etc. Ces blocs peuvent être à leur tour subdivisés pour être appris sur plusieurs années. Adam & Eve, de l'Université de Louvain, permet aussi de construire des listes de vocables-cible selon la fréquence de leur occurrence dans des corpus de textes. Des tests de closure peuvent être générés automatiquement par le système à partir de textes sélectionnés par l'apprenant.

Ces programmes permettent d'individualiser le contenu mais pas les stratégies d'apprentissage. D'autre part, là encore, il ne s'agit que de valider des connaissances déjà acquises. Aussi, ces programmes doivent plutôt être considérés comme des aides à l'enseignant plutôt qu'à l'apprenant puisqu'il lui permet de choisir le matériau sur lequel ses apprenants doivent travailler.

1.2 Principes psychologiques

Certains programmes ont été conçus en tirant parti des études en psycholinguistique, se fondant sur la manière dont fonctionne le lexique mental. C'est le cas par exemple de la technique de la répétition accrue dans laquelle la quantité de vocables à apprendre est augmentée progressivement à chaque nouvelle présentation de vocabulaire. Cette technique, selon Siegel & Misselt (1984) donne de meilleurs résultats que les techniques de révisions plus traditionnelles. D'autres programmes associent des vocables avec des images visuelles (travaux d'Atkinson et Raught, 1975). D'autres demandent à l'apprenant de retrouver des vocables à l'aide d'indices ou à l'aide de relations sémantiques pertinentes telles que l'hyponymie ou la synonymie (Catt, 1992).

Ces programmes, bien que s'appuyant incontestablement sur des principes psycholinguistiques pertinents et favorisant la mémorisation ou la production, ont cependant l'inconvénient de ne proposer qu'un nombre fixé de vocables, ceux prévus par les concepteurs. Il n'est pas possible d'incorporer de nouveaux vocables, venant par exemple de lectures. D'autre part, les associations sont aussi déterminées par les concepteurs et sont indépendantes de l'apprenant. Enfin, ces programmes nécessitent beaucoup de préparation pour être étendus à une quantité de vocabulaire plus large.

1.3 Programmes d'acquisition fortuite

Certains programmes tentent de faire acquérir le vocabulaire de manière fortuite en faisant effectuer à l'apprenant des tests de haut niveau tels que la reconstruction de texte (Storyboard, Higgins et Johns, 1984). Ces tests, qui ne peuvent être effectués que sur ordinateur de par l'interactivité qu'ils nécessitent, sont une extension des tests de closure : il faut reconstituer un texte dont tous les vocables manquent, ceux-ci étant signalés par des blancs soulignés. Bien que cette méthode puisse poser de nombreux problèmes aux apprenants (car ils doivent connaître déjà les vocables sinon ils ont peu de chance de pouvoir remplir les trous), elle fait intervenir un certain nombre de processus en jeu dans l'apprentissage lexical en encourageant l'utilisateur à deviner et à inférer les vocables à l'aide de divers indices (lettres, préfixes, suffixes, etc.). Ainsi les vocables (grâce à certaines lettres trouvées), les collocations (un des constituants peut être deviné grâce à la présence d'un autre), les structures grammaticales (découverte de la bonne préposition) peuvent théoriquement être apprises de cette manière. Storyboard met en évidence les relations complexes entre la connaissance lexicale, l'habileté de lecture et les capacités d'inférence et de production. De ce point de vue, ce logiciel est bien plus intéressant que les simples programmes qui testent la mémoire vus plus haut.

En conclusion pour les logiciels de première génération d'aide à l'apprentissage lexical, on peut noter qu'ils visent surtout à tester ce qui est déjà appris (ou non). Les limitations techniques font que les échanges sont souvent pauvres : d'une part, l'apprenant est obligé de taper les vocables, d'autre part, les retours de l'ordinateur sont sommaires. Ils se résument le plus souvent à un « correct ou incorrect » et ne donnent aucune indication sur l'erreur des apprenants. Ils ne varient pas au cours du temps et ne s'adaptent pas à l'apprenant et à ce qu'il a déjà répondu. Ils ne considèrent pas non plus les réponses partiellement justes, ce qui est démotivant pour les apprenants.

Les environnements informatiques doivent permettre à l'apprenant de personnaliser son apprentissage et donc de pouvoir choisir leur propre contenu à partir de ressources linguistiques. On se tourne alors vers les logiciels de deuxième génération qui s'appuient sur le modèle développé par Kukulska-Hulme (1988) (figure 2.1) qui propose que les deux étapes importantes sont de trouver le sens d'un item lexical dans un certain contexte à l'aide d'outils lexicaux et de pouvoir se le remémorer à partir d'un enregistrement écrit afin de l'utiliser en production dans un nouveau contexte. Le rôle de l'ordinateur est de pouvoir organiser les enregistrements écrits à partir de critères sémantiques ou orthographiques.

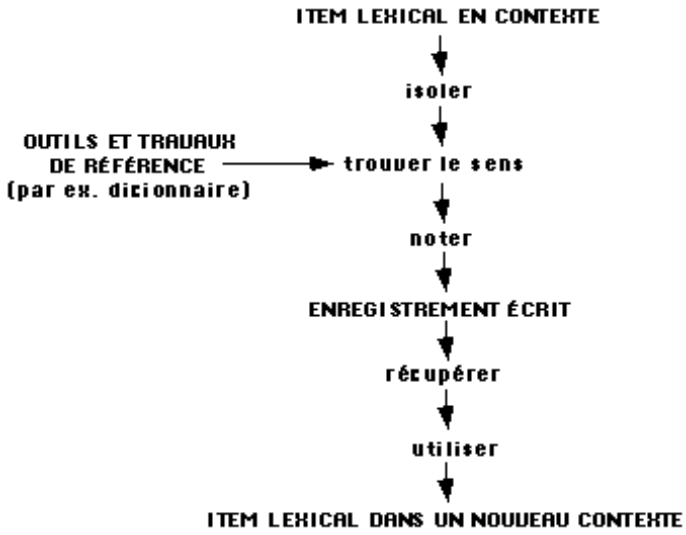


Figure 2.1 : le modèle de Kukulska-Hulme, « voyage d'un item lexical »

Ce modèle, formulé en premier lieu avec les limites des logiciels de première génération (outils lexicaux et sources linguistiques hors-ligne, système ressemblant plutôt à une classique base de données) est pourtant annonciateur de ce qui va pouvoir être fait avec les logiciels de deuxième génération avec les apports des interfaces graphiques et des ressources lexicales et linguistiques en ligne. Le modèle de Kukulska-Hulme se concentre donc sur le processus central de l'apprentissage de vocabulaire en assignant à l'ordinateur un rôle d'organisation des informations cherchées par l'apprenant.

2 Les programmes de deuxième génération

Grâce à l'amélioration des performances des ordinateurs (traitements de texte, interface graphique, hypertexte, ressources en ligne), le modèle de Kukulska-Hulme a pu être concrétisé en de meilleurs termes. De plus, des programmes de traces plus évolués permettent de mieux interpréter les intentions de l'utilisateur. Gillespie et Gray (1992) proposent un système hypertexte permettant à l'apprenant de sélectionner des mots et expressions directement à partir d'un corpus de textes et de les stocker dans un réseau de cartes HyperCard. L'utilisateur développe ainsi sa connaissance de nouveaux vocables et sa capacité à les classer. D'autres programmes (Lyman-Hager *et al.*, 1993) permettent d'associer des informations (qui peuvent être aussi des images ou du son) à des parties de textes, informations qui seront accessibles à l'apprenant en sélectionnant telle ou telle zone et qui lui permettront d'avoir une meilleure compréhension du texte et de classer en fonction d'indices (sémantiques ou autre) plus riches. Néanmoins, ce genre de système demande beaucoup de préparation à cause des informations à incorporer et l'investissement peut s'avérer peu rentable par rapport au vocabulaire véritablement acquis. Il semble donc inévitable d'incorporer aux systèmes d'apprentissage des outils lexicaux ou des ressources permettant de distiller à la demande telle ou telle information sur un vocable donné sélectionné dans le texte.

2.1 Les outils lexicaux

Un grand nombre d'outils lexicaux à base de traitement automatique du langage naturel sont maintenant disponibles. Il s'agit notamment de dictionnaires en ligne de diverses formes (monolingues, bilingues, spécialisés), de concordanceurs, d'analyseurs morphologiques, d'analyseurs syntaxiques et de vérificateurs d'orthographe et de grammaire. L'outil le plus fréquemment rencontré est le dictionnaire électronique (nous y reviendrons au chapitre 5) même si ceux-ci, pour l'instant, n'ont rien de spécifiquement informatique mais sont des adaptations sur ordinateur de la version papier.

Mayday (Sussex, Cumming et Cropp, 1994) est un environnement ouvert d'aide à l'apprentissage lexical qui repose, d'une part, sur un jeu d'activités lexicales et, d'autre part, sur plusieurs outils lexicaux complémentaires. Le but du système est de familiariser l'apprenant au phénomène de la composition de vocables en anglais à l'aide d'affixes, de préfixes et de suffixes. Pour cela, plusieurs activités lexicales sont proposées telles que des exercices sur les affixes des verbes (figure 2.2), des mots-croisés, des correspondances de sens entre une série de définitions et une autre de vocables composés à l'aide d'affixes, de préfixes et de suffixe ou bien des exercices à trous (l'apprenant doit trouver le vocable manquant à l'aide de racines et de suffixes proposés dans une liste).

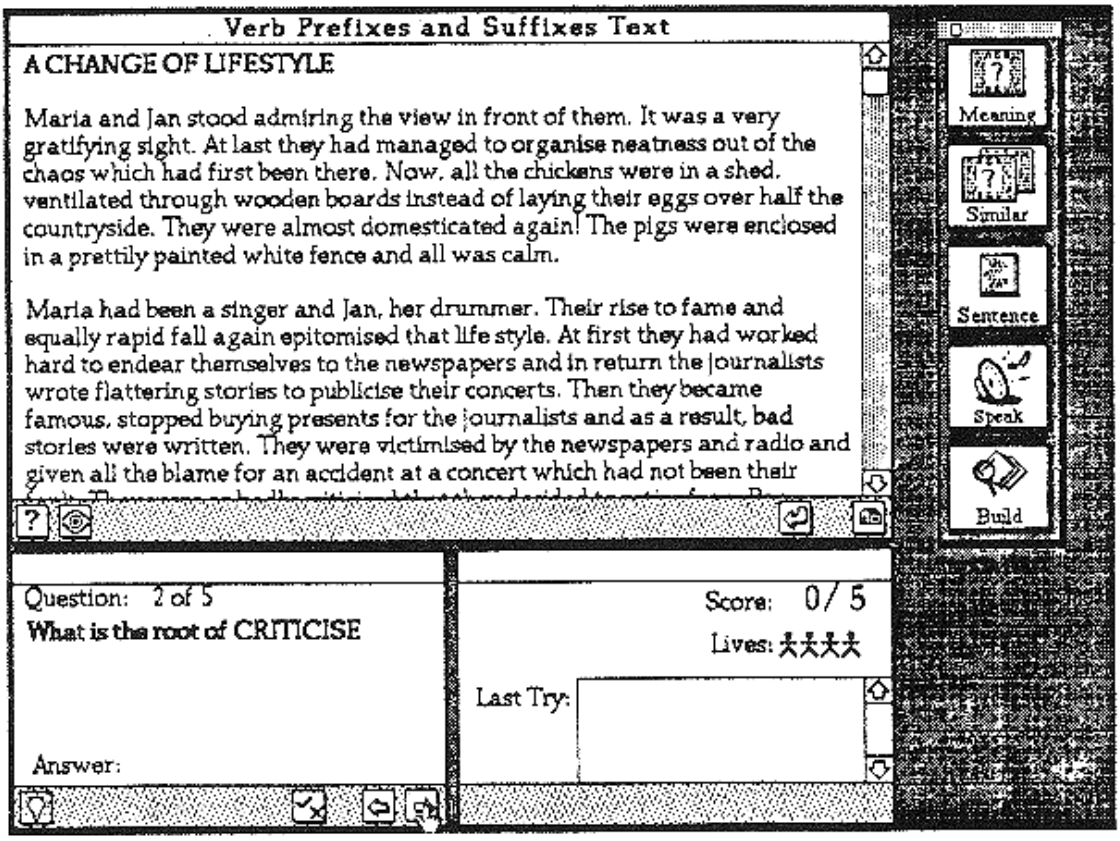


Figure 2.2 : Mayday, exercice sur les affixes des verbes

Pour toutes ces activités, qui sont indépendantes les unes des autres, qui peuvent être faites dans n'importe quel ordre et à tout moment, l'apprenant dispose de nombreux outils lexicaux variés. L'environnement comprend un dictionnaire pour apprenant (*Oxford Advanced Learner's Dictionary*), un dictionnaire de synonymes, un corpus d'exemples, une présentation sonore des vocables pour indiquer la prononciation et un module de construction de vocables composés à l'aide d'une racine. Le but de ce système étant justement d'étudier la composition des vocables, les entrées des dictionnaires sont les racines et les vocables composés doivent être consultés à partir du module de construction. Les définitions du dictionnaire sont numérotées et séparées du reste des informations (phonétique, exemples, etc.) de manière à pouvoir comparer les sens sans avoir à lire d'autres types d'informations. Celles-ci, dans les autres ressources, sont toutefois reliées au sens en question dans le dictionnaire principal par le même numéro, ce qui évite de se perdre. Par ailleurs, le système respecte les principes d'ergonomie et d'interaction homme-machine : la disposition des différents éléments à l'écran doit être la plus simple et la plus intuitive possible. Les outils, l'interaction avec l'apprenant, ses réponses et le retour de l'ordinateur sont dans des zones séparées qui conservent le même aspect et qui restent aux mêmes endroits dans toutes les activités. D'après les auteurs, l'utilisation du système n'a posé aucun problème même pour les étudiants qui éprouvaient des réticences vis-à-vis de l'ordinateur.

Mayday est donc un environnement ouvert d'exploration lexicale pour l'apprentissage de la langue dans le sens où l'apprenant a à sa disposition plusieurs sources d'informations et la possibilité d'appliquer plusieurs stratégies possibles pour parvenir à l'information recherchée. L'enseignant a aussi un rôle important à jouer dans la mesure où il peut aider et conseiller l'étudiant dans son travail. Le but de cette recherche est d'établir la quantité de connaissances que l'enseignant doit avoir sur la tâche que l'étudiant est en train d'effectuer pour pouvoir l'aider et le guider convenablement. Ce système est bien de seconde génération dans le sens où il aborde véritablement la question de l'apprentissage lexical et les types d'interactions que les apprenants peuvent avoir avec de tels programmes.

Pour autant, il ne faut pas réduire ces programmes à forte composante d'outils lexicaux à ces ressources lexicales elles-mêmes. Sans principes pédagogiques sous-jacents, un système ne peut prétendre favoriser l'apprentissage. C'est le cas, par exemple, de GLOSSER dans sa version actuelle (Nerbonne et Smit, 1996 ; Dokter *et al*., 1997 ; Nerbonne *et al*., 1998). GLOSSER est un environnement informatique composé uniquement de ressources lexicales. Il est destiné à faciliter la lecture de textes français pour des apprenants néerlandais, bien qu'en principe, il puisse s'adresser à d'autres étudiants apprenant d'autres langues pourvu que les ressources en question soient disponibles. Il est composé d'un corpus de textes français, d'un dictionnaire bilingue français-néeerlandais (*Van Dale*), d'un analyseur (analyse morphologique et désambiguïsation), Locolex (Bauer et Zaenen, 1995), développé par Rank Xerox et d'un corpus d'exemples accessible par un concordanceur. L'apprenant choisit donc un texte, le lit et par simple clic, peut consulter différentes informations dans les différentes ressources. Il peut donc avoir aussi bien la traduction du vocable dans sa langue, que sa forme canonique ou qu'une liste de phrases contenant le même vocable afin de comparer. D'après les auteurs (Nerbonne *et al*., 1998), le système facilite l'approche d'un texte en langue

étrangère. En effet, ils notent le plus grand nombre de vocables consultés dans le dictionnaire électronique par rapport à sa version papier (il faut 22,6 secondes en moyenne pour consulter un vocable dans le dictionnaire électronique contre 85,7 pour sa version papier) et la diminution conséquente de temps nécessaire pour lire le texte. Ils concluent sur le fait que ces deux phénomènes peuvent améliorer l'acquisition lexicale. Cette conclusion semble un peu rapide. En effet, ils remarquent, avec sincérité, que l'expérimentation n'a pas révélé de différence notable dans la compréhension du texte entre les deux méthodes, mais, toutefois, qu'ils s'attendent à ce que celle-ci apparaissent avec un nombre plus important de sujets. En attendant, cela reste à confirmer. Il nous semble en effet que cette constatation (pas de différence significative de compréhension entre les deux méthodes) puisse s'expliquer par d'autres raisons, notamment par la faiblesse des principes pédagogiques sous-jacents au système.

Tout d'abord, le choix des textes. A aucun endroit, les auteurs ne motivent leur choix sur les textes qu'ils présentent. Ceux-ci sont extraits d'Internet (projet Gutenberg par exemple), ou bien de projets de corpus spécialisés (MULTEXT, ECI). Il s'agit donc de textes « tout-venant », de sujets très variés, plus ou moins spécialisés et dont les niveaux de difficulté, par rapport au public visé de niveau intermédiaire (plutôt en lycée), peuvent être non pertinents pédagogiquement (ce que les auteurs reconnaissent d'ailleurs).

Le point fort du système réside dans la présence de l'analyseur Locolex. En effet, grâce à celui-ci, le lien peut être effectué entre les chaînes de caractères du texte et les entrées du dictionnaire électronique bilingue. L'analyse morphologique et la désambiguïsation permet d'orienter le système vers l'entrée la plus appropriée, avec un taux d'erreur satisfaisant et conforme à l'état de l'art. Cependant, même si l'utilisation du dictionnaire s'avère incontestablement bien plus pratique que la version papier[2], les problèmes d'accès lexical et de compréhension ne sont pas réglés pour autant. Comment se comporte le système en cas d'homonymie, en cas de collocation ? Et surtout, quelle aide apporte-t-on à l'utilisateur pour choisir entre les différentes lexies de l'article, qui peuvent être nombreuses dans certains cas ? Car ce que cherche l'apprenant avant tout n'est pas l'article complet du vocable sélectionné, mais son sens dans le texte. La désambiguïsation sémantique, il est vrai, n'est pas d'un point de vue technologique encore possible et il est probable qu'elle ne pourra jamais être entièrement automatisée. Du reste, le problème est le même avec un dictionnaire papier, comme nous le verrons au chapitre suivant. Les auteurs tentent bien une correspondance (*matching*) entre le contexte du vocable sélectionné et celui des exemples du dictionnaire. Même si cette tentative peut dans certains cas porter ses fruits, il est probable qu'elle ne sera pas applicable la plupart du temps. Une solution consiste à préparer à l'avance les textes, en désambiguïsant (attribution d'un numéro de sens en correspondance avec le dictionnaire) par exemple les vocables les plus polysémiques. Il ne serait plus alors question de travailler sur des textes tout-venant mais les limites des traitements automatiques sur ce point imposent une

[2] Il faut noter toutefois que l'accès au dictionnaire interrompt le processus de lecture (surtout s'il dure en moyenne 22 secondes, ce qui n'est tout de même pas négligeable). Il ne faut donc pas en abuser et inciter l'apprenant à ne chercher dans le dictionnaire que les vocables importants pour la compréhension et difficiles à comprendre d'après le contexte.

compensation par un apport humain manuel. En fait, une décision pédagogique doit être prise en fonction du niveau des apprenants et de la difficulté des textes. Il nous semble qu'une telle préparation est inévitable lorsque le niveau des apprenants n'est pas assez élevé par rapport à celui du texte.

Le même problème surgit avec les exemples : est-il intéressant, pour la compréhension du vocable en question, d'avoir une série de phrases dans lequel il peut être employé avec un sens différent ? On peut aussi se poser la question de la pertinence de ces concordances lors d'une lecture. L'étude de l'utilisation du système montre qu'elles sont bien moins utilisées que le dictionnaire. Ceci peut s'expliquer par le fait que le dictionnaire est un outil bien plus familier que le concordanceur. Mais la raison principale réside sans doute dans le fait que l'examen des exemples est une opération complexe qui demande beaucoup plus de temps et de réflexion de la part de l'apprenant qu'une consultation dans un dictionnaire. De plus, le résultat (trouver la signification du vocable ou des éléments de signification) est loin d'être garanti. Le concordanceur nous semble surtout pertinent pour l'exploration du lexique (nous y reviendrons un peu plus loin), mais beaucoup moins lorsqu'il s'agit de compréhension écrite. Car il ne faut pas oublier que pendant ce temps la lecture est interrompue (et sans doute beaucoup plus longtemps que dans le cas d'un dictionnaire), ce qui est toujours préjudiciable à la compréhension. Quant à l'analyseur morphologique, en dehors de son utilisation pour déterminer l'entrée du dictionnaire à consulter, il est certes intéressant pour comprendre les mécanismes de flexion et pour trouver la forme canonique, mais il n'apporte rien sur le plan de la compréhension du texte.

Enfin, sans la moindre conservation explicite, et organisation, des vocables consultés (sous forme d'enregistrements dans une base de données personnalisée par exemple), il y a fort à parier que ceux-ci ne seront pas retenus (il faut en effet une exposition répétée et une révision régulière, chapitre 1) et donc appris. L'étude de l'utilisation du système montre d'ailleurs qu'un certain nombre de vocables ont dû être consultés plusieurs fois car leur signification avait été oubliée entre-temps. Le système permet de sauvegarder les traductions des vocables consultés. Ceci nous semble pourtant insuffisant comme annotation, car une simple traduction se résume à un accès de dictionnaire optimisé (l'apprenant obtient le sens correct immédiatement). Il n'implique rien sur la mémorisation. Aucun travail sur le sens n'est effectué et le lien dans le lexique mental sera d'autant plus faible. Par ailleurs, un dictionnaire bilingue est loin de donner toutes les informations nécessaires à la connaissance d'un vocable.

En conclusion, GLOSSER, quoiqu'en disent ses auteurs, ne nous semble pas un bon exemple d'environnement favorisant l'apprentissage lexical. Il présente des idées intéressantes pour la lecture de textes mais, en l'état actuel du système, la compréhension et la mémorisation des vocables nous semble plus problématique. Il se résume à ses ressources, et n'a pas vraiment de programme pédagogique d'acquisition de vocabulaire. Il est donc nécessaire que ce système soit amélioré, notamment par l'ajout de plus d'aide à la compréhension et par celui d'un module permettant de revoir et d'organiser les annotations les unes par rapport aux autres.

Les exemples évoqués ci-dessus nous amènent à parler des concordanceurs.

2.2 Les concordanceurs

Ces programmes ne sont pas véritablement des environnements d'apprentissage mais ils sont très utiles lorsque l'apprenant désire approfondir ses connaissances sur un vocable. A partir d'un vocable donné, un concordanceur est capable, en parcourant une vaste base textuelle, de fournir toutes les phrases le contenant. Le résultat est souvent affiché en format KWIC (Key-Word In Context) ou, en français, MCC (Mot-Clé en Contexte), c'est-à-dire que le vocable-clé est affiché au milieu de l'écran ligne par ligne et est entouré par son contexte gauche et droit (figure 2.3). La régularité de cette disposition permet alors de mettre en évidence les caractéristiques du vocable. En principe, un concordanceur doit pouvoir trier les contextes par ordre alphabétique ce qui permet d'étudier, par exemple, les schémas syntaxiques ou encore les collocations. Dans la figure 2.3, le vocable-clé est en fait travail, mais par tri du contexte droit, on peut étudier travail au noir. Les autres fonctionnalités concerne les options de recherche du vocable-clé : il doit être en effet possible de chercher plusieurs vocables ou des parties de vocable grâce à l'emploi de booléens ou de jokers (par exemple, on pourra chercher toutes les occurrences du verbe travailler par la requête travaill*). Du point de vue de l'utilisation, un concordanceur doit être rapide pour ne pas perdre de temps, notamment vis-à-vis des textes à charger et des recherches, et vu que les erreurs sont fréquentes au début, surtout pour les élèves, les recherches doivent pouvoir être interrompues.

L'un des concordanceurs les plus connus est Micro-Concord (Scott et Johns) (figure 2.3), rapide et facile d'utilisation, développé par Oxford University Press qui peut travailler sur corpus dont la taille limite est donnée par la capacité du disque dur (contrairement à son prédécesseur Mini Concordancer de Longman qui gardait tout le texte en mémoire vive, ce qui limitait beaucoup plus la taille des textes). Le nombre de concordances est cependant limité (1 806 sur notre ordinateur) ce qui implique qu'il s'agit plus d'un programme destiné à être utilisé en classe de langue plutôt que pour la recherche linguistique. Ball (1997) recense un certain nombre de concordanceurs en fonction de leurs possibilités, de leur plate-forme et de leur prix.

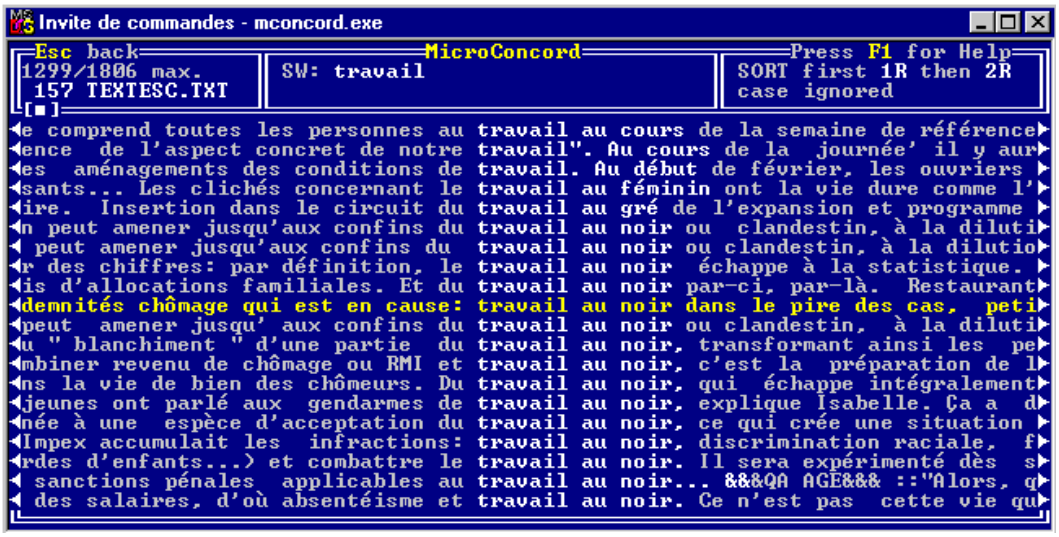


Figure 2.3 : concordances de MicroConcord pour le vocable travail au noir

L'intérêt de ces programmes réside dans le fait que les phrases (ou parties de phrases) obtenues sont extraites de textes authentiques et reflètent véritablement l'usage de la langue. Ce ne sont pas des exemples construits par des spécialistes. D'autre part l'apprenant assume le contrôle du processus et devient le « linguiste », essayant d'identifier, de classifier et de dégager des régularités dans le comportement du vocable par rapport au contexte environnant. Mis à part les différents sens du vocable, il est possible d'observer son comportement syntaxique, collocationnel et son registre. Étant donné que l'information est implicite et doit être extraite par l'observation et l'inférence, on parle de *data-driven approach*, c'est-à-dire une exploration conduite par les données. D'après Johns (1991), il est possible d'observer de nombreux phénomènes linguistiques comme par exemple les structures grammaticales (constructions différentes pour convince et persuade), les hyperonymes (révélés par l'emploi de such as : industries such as steelmaking) ou les adverbes (différence d'utilisation entre however et nevertheless). L'intérêt des concordanceurs est ici d'autant plus évident que ces informations ne sont pas toujours décrites par les dictionnaires. D'après Stevens (1993) « un concordanceur et un corpus d'anglais naturel permettent aux apprenants de raccourcir le processus d'acquisition de compétences en langue cible par le fait que l'ordinateur est capable de les aider à organiser de vastes quantités de données langagières et à discerner ainsi des régularités de schémas plus facilement ». Ce qui est résumé par Tribble (1990) : « Ce que le concordanceur fait est de rendre visible l'invisible ».

Pour autant, et bien que depuis plus d'une dizaine d'années, les corpus et les concordanceurs aient été régulièrement décrits comme l'une des idées les plus prometteuses en ALAO (Johns, 1986 ; Leech et Candlin, 1986 ; Johns et King, 1991 ; Hanson-Smith, 1993), peu d'expérimentations ont été menées pour vérifier en quoi, ou sous quelles conditions, l'utilisation de concordanceurs facilite l'apprentissage de certaines parties du vocabulaire, en comparaison avec des techniques d'enseignement plus classiques. C'est le constat que dresse Cobb (1997) qui relève, contrastant ainsi avec cette vague d'enthousiasme, une seule petite expérimentation (Stevens, 1991) menée à l'université Sultan Qaboos dans le

sultanat d'Oman. Stevens voulait montrer qu'il était plus facile de se remémorer un vocable, qui était occulté et qu'il fallait retrouver, grâce à des concordances (et donc avec différents contextes) plutôt que grâce à une seule phrase complète. Les résultats lui ont d'ailleurs donné raison, ce qui montrait que les concordanceurs avait au moins un rôle à jouer dans l'apprentissage.

En dehors de cette étude, rien qui fasse l'objet d'une évaluation d'après les formats standard d'expérimentation. Cobb fait remarquer que dans le recueil consacré à des études sur l'utilisation de concordanceurs par des apprenants (Johns et King, 1991), les articles se bornaient à décrire les activités dans lesquelles étaient impliqués les étudiants, mais qu'« aucun fondement théorique n'était exploré, aucune hypothèse falsifiable n'était formulée, aucune production d'apprenant n'était mesurée, aucune comparaison n'était tentée ». Pour expliquer cela, il avance les hypothèses que les concordanceurs commerciaux permettent difficilement l'établissement de protocole d'expérimentation, ne permettant pas de génération de traces et n'autorisant que des observations informelles, et d'autre part, que les études de nouveaux médias d'apprentissage sont difficiles à mener du fait que les étudiants mettent du temps à s'habituer à ces nouveaux outils et changent d'attitude vis-à-vis d'eux au cours du temps.

L'expérimentation menée par Cobb poursuit les travaux de Stevens et cherche à montrer la supériorité des concordances sur une phrase seule entière lorsqu'il s'agit non pas de se remémorer un vocable, comme dans Stevens, mais de l'apprendre.

Pour cela un programme spécifique PET•200 a été conçu et testé pendant une année académique sur une centaine d'apprenants arabes motivés de niveau universitaire en première année d'anglais. L'expérimentation consistait à utiliser des concordances dans diverses activités lexicales de difficulté variable. Il s'agissait d'étudier 20 vocables par semaine sur 12 semaines (soit 240 vocables au total) et de les apprendre. Les mêmes tâches étaient effectuées une semaine à l'aide de concordances et la semaine suivante, sans concordance, c'est-à-dire à l'aide d'une seule phrase complète (définition ou exemple du vocable-clé), puis on alternait à nouveau. Un programme de traces enregistrait toutes les actions des utilisateurs. Leur niveau de vocabulaire était mesuré avant, pendant et après. En effet, en plus d'un pré-test et d'un post-test, chaque semaine, un test mesurait l'acquisition des 20 items. Il était constitué de deux tâches : un exercice d'orthographe, vu comme une mesure de contrôle, et un exercice dans lequel les étudiants devaient remplir un texte à trou avec les vocables nouvellement étudiés. Quant aux activités lexicales d'étude du vocabulaire, elles consistaient :

- à trouver la définition du vocable à apprendre (QCM) parmi trois autres générées aléatoirement,
- à reconnaître le vocable, occulté, dans une série de lettres mélangées au hasard avec d'autres,
- à reconnaître un vocable d'après sa prononciation,
- à retrouver des vocables (proposés par un menu déroulant) dans un texte, l'aide étant une série de concordances avec le vocable masqué,
- à retrouver des vocables comme ci-dessus, mais en les tapant au clavier.

Les analyses ont été effectuées à partir des résultats de 11 sujets choisis au hasard parmi une centaine. On peut regretter toutefois que l'échantillon n'ait pas été plus large. Les deux tests antérieur et postérieur à l'expérimentation ont vérifié que les sujets avaient acquis en moyenne 430 vocables en trois mois ce qui est bien au-dessus de la moyenne européenne (275 vocables en six mois d'après Milton et Meara, 1995). Ce qui montre l'efficacité du programme. Celui-ci a d'ailleurs très bien été perçu par les apprenants qui l'ont noté bien mieux que les autres matériaux d'apprentissage plus traditionnels, livre de grammaire compris.

Les scores obtenus aux tests hebdomadaires reflètent l'alternance des semaines avec et sans concordance. Les résultats ont été meilleurs de 12 % en moyenne pour les semaines avec concordances (75,9 % contre 63,9 %). Dans le détail, 8 sujets sur 11 ont été meilleurs avec l'utilisation de concordances.

Cobb pour autant se demande si ce gain est dû aux concordances et avance deux raisons qui permettent de le croire : premièrement, la version sans concordance a été utilisée normalement comme le montre le test d'orthographe hebdomadaire pour lesquels les résultats restent constants d'une semaine à l'autre. Deuxièmement, on constate que si le temps passé dans les activités lexicales est globalement le même dans chacune des versions (10 heures en moyenne par sujet), il n'en est pas de même de la quantité de traces générées, qui est significativement plus faible (40 %) pour la version avec concordances. En d'autres termes, les sujets ont fait quelque chose pendant la version avec concordances qu'ils n'ont pas fait avec la version sans, quelque chose qui leur a permis d'améliorer leurs résultats de 12 % en moyenne. Il est donc difficile de conclure pour l'auteur que ce soit autre chose que la lecture des concordances.

Pour résumer, les sujets passent autant de temps et ont les mêmes résultats à des tests d'orthographe que les activités lexicales comprennent des concordances ou pas. Par contre, lorsqu'elles en comprennent, ils répondent à 40 % moins de questions du programme mais ont des résultats 12 % meilleurs dans les exercices des textes à trous. Il faut donc en déduire l'utilité des concordances pour l'apprentissage de vocabulaire.

2.3 Lexica

Goodfellow (1994) a développé un environnement informatique d'aide à l'apprentissage de vocabulaire (Lexica) s'adressant à des adultes motivés et non débutants et reposant sur un modèle comprenant trois étapes : sélection de vocables, regroupement et test. Son principe repose sur le fait que les vocables ne sont pas des étiquettes arbitraires mais sont reliés entre eux de manière systématique et forment des familles ou champs sémantiques (élément de sens en commun, dérivés ayant la même racine, orthographe ou prononciation similaire, etc.). Il s'agit donc d'aider à la mémorisation, en vue d'une utilisation productive, en essayant de prendre conscience de ces relations lexicales concernant des vocables nouvellement rencontrés. L'environnement a été conçu de manière à faciliter :

- l'enregistrement de vocables nouveaux et intéressants lors de la lecture de textes,
- la recherche et la prise de notes sur leur sens et leur emploi,

- le regroupement en fonction de relations lexicales,
- la possibilité de se tester sur sa capacité à se remémorer et produire les vocables travaillés.

Le système est composé :

- d'un corpus de textes (50 000 mots répartis dans des textes de 300 à 2 000 mots).
- d'un module de sélection qui permet d'extraire des portions de texte (mot, expression, ligne, etc.) vers une liste.
- de ressources lexicales comprenant un dictionnaire bilingue en ligne dans le sens version seulement (L2-L1) et un concordanceur, ressources servant à obtenir des informations sur un vocable donné.
- d'un module de regroupement dans lequel l'apprenant peut regrouper les vocables sélectionnés suivant trois critères : un critère de sens (synonymes, hyperonymes, etc.), de forme (suffixes, constructions grammaticales, phonétique, etc.) ou de contexte (collocations ou vocables liés). L'apprenant peut mettre des annotations sur les vocables et doit donner un nom à chacun des groupes qu'il compose.
- d'un module de test. Le concordanceur affiche la liste des phrases contenant le vocable-cible en le remplaçant par un vide. Pour trouver le vocable, des indices sont proposés comme le titre du groupe dont il fait éventuellement partie et les autres vocables qui le compose, les annotations qui y sont attachées et les voyelles, consonnes, préfixes ou suffixes qui le composent.
- d'un modèle de l'apprenant. Chaque action est « tracée » et conservée. On peut donc étudier les actions effectuées et leur durée. Ce modèle permet donc de savoir la manière dont le système est utilisé.

Tout comme Mayday, Lexica est un environnement ouvert dans le sens où l'apprenant choisit le contenu lexical à étudier et la manière dont il l'étudie, c'est-à-dire les stratégies qu'il utilise. Le procédé de sélection de vocables dans les textes et de collecte d'information à partir des ressources disponibles construit des liaisons faibles dans le lexique mental de l'apprenant. Le but du module de regroupement est de renforcer ces liaisons. L'apprenant doit en effet catégoriser les vocables qu'il sélectionne, c'est-à-dire discerner des ressemblances suivant différents critères avec les autres vocables (figure 2.4).

Fig 4.3: Lexicon-Building Module

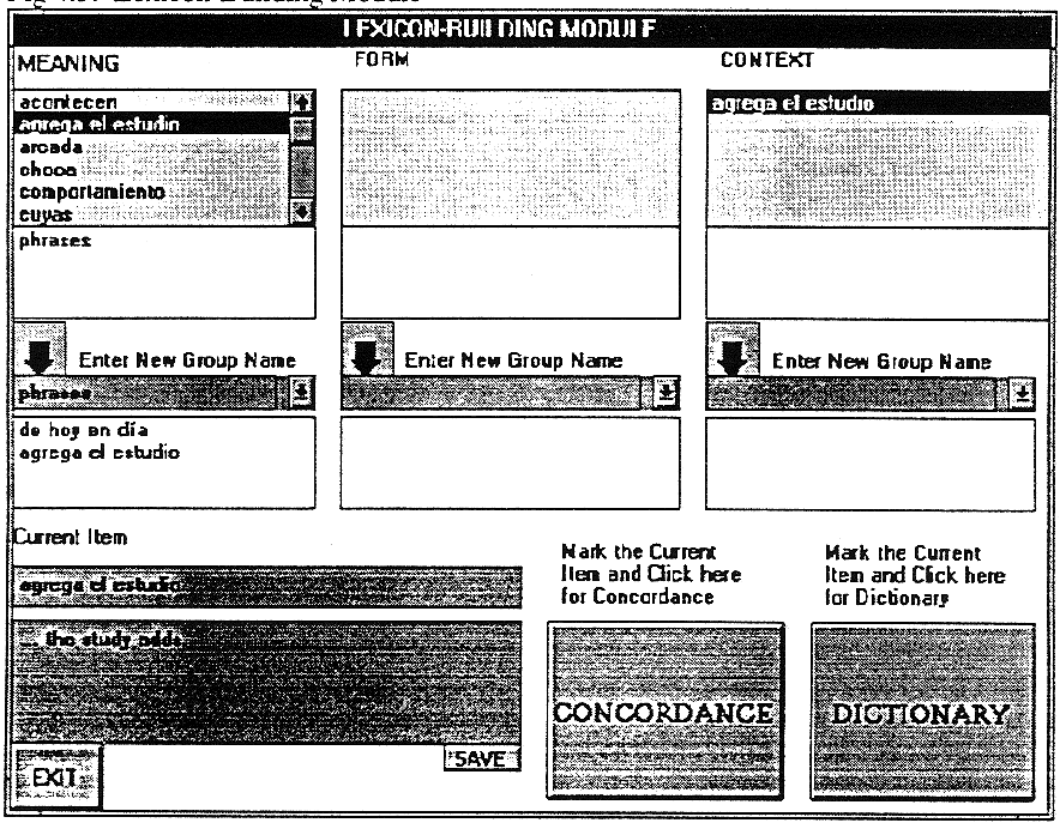


Figure 2.4 : regroupement des vocables dans Lexica

Cette activité l'oblige à effectuer un travail « en profondeur » sur les vocables, propre à mettre en relief leurs caractéristiques. L'effort mental ainsi fourni doit permettre de renforcer les traces mémorielles et de multiplier les liens avec les autres items mentaux. Le module de test, enfin, favorise l'activation des vocables stockés dans la mémoire à partir des associations (les indices proposés) créées par le module de regroupement. L'apprenant utilise ainsi le traitement effectué sur chaque vocable afin de pouvoir se le remémorer. Un système de révisions permet d'améliorer la rétention par une exposition répétée jusqu'à ce que le système décide, par une notation, que le vocable est acquis.

L'expérimentation menée a conduit à des résultats satisfaisants. Certains apprenants sont arrivés à un taux d'acquisition de huit vocables par heure d'utilisation, ce qui était le taux initialement fixé.

Les points faibles relevés par Goodfellow concernent le manque de préparation et de formation des apprenants au système. Le processus de regroupement a ainsi été jugé difficile par certains apprenants. D'autre part, le système n'a pas été utilisé exactement comme on pouvait l'espérer. Certains apprenants ont ainsi concentré leur énergie sur la recherche de traductions en langue maternelle.

Goodfellow (1997), Lamy et Goodfellow (1998), décrivent le portage de Lexica (Lexica OnLine) sur Internet, ce qui apporte des éléments nouveaux. Les apprenants, tous des adultes motivés et rémunérés pour effectuer un minimum de tâches fixé, ne sont plus directement face au système, mais l'utilisent à distance. Ce point n'a d'ailleurs pas véritablement posé

problème. Ils disposent d'un tuteur humain qui les aide en cas de problèmes. Le point le plus significatif est l'incorporation d'un forum de discussion en ligne asynchrone dans lequel les messages sont organisés hiérarchiquement et montrent la liste des réponses pour tel ou tel message. Il est dès lors possible de savoir qui répond à qui. En dehors des questions technologiques, le but de ce forum est, par le biais de la discussion, de faire prendre conscience aux étudiants des processus impliqués dans l'acquisition de vocabulaire. Il est modéré par un tuteur dont le rôle est de provoquer la discussion en encourageant les questions et les réponses. Ce forum est aussi l'occasion de savoir ce que pensent les utilisateurs du système.

L'analyse des données montrent que le forum a eu du succès, tous les utilisateurs y ayant participé. Les commentaires portent sur les fonctionnalités du système comme le concordanceur qui, bien qu'étant un outil inhabituel, a captivé les étudiants. Par contre, ceux qui ont apprécié le regroupement, le trouvant digne d'intérêt et posant des questions sur son utilité, ne sont pas nombreux. Ceci est symptomatique d'une certaine réticence à manipuler des relations langagières plus abstraites telles que la classification lexicale, la morphologie, la suffixation ou les problèmes de fréquence. Pour certains, il pourrait s'agir d'une non familiarité avec le métalangage mais Goodfellow pense qu'il y a une objection plus fondamentale, celle que ces questions touchent davantage les linguistes ou les experts de la langue mais pas ceux qui veulent « juste utiliser la langue ». Pourtant, lorsqu'on insiste, les utilisateurs sont tout à fait capables d'opérer ces regroupements.

Le gros des messages échangés sur le forum a toutefois concerné le sens des vocables et expressions auxquels les étudiants étaient confrontés et leur emploi en contexte. Ceci s'explique par le fait que les apprenants ont l'habitude et aiment discuter des aspects, notamment lexicaux, de la langue qu'ils travaillent. C'est un facteur d'intégration dans un groupe, qui permet de construire une identité et qui révèle les aptitudes, l'éducation, l'expérience et les autres aspects personnels de chacun.

Une question posée par l'analyse des messages du forum est de déterminer si des vocables nouveaux sont appris (soit de manière explicite ou plus implicite en réutilisant les tournures employées par le tuteur, ou un pair, ou une phrase d'un texte) et réutilisés. Les apprenants déclarent tous avoir appris du vocabulaire, mais comment le mesurer ? C'est un problème qui se pose lors de l'analyse des discussions en ligne et qui doit peser sur les méthodes d'interprétations de ces données.

2.4 Les traces

Afin d'analyser l'utilisation de l'ordinateur par les apprenants, les environnements se dotent de programmes de traces qui enregistrent chaque action effectuée par l'utilisateur, et cela de manière précise et transparente (sur ce dernier point, voir par exemple Hulstijn, 1993, qui préfère des enregistrements transparents, que l'apprenant ne peut soupçonner car invisibles et inaudibles, plutôt que l'utilisation de la vidéo, ceci afin d'interférer le moins possible avec l'objet de l'expérimentation). Dans l'exemple de Lexica, le programme enregistre par exemple chaque bouton cliqué, chaque texte sélectionné, chaque vocable

regroupé, etc. avec le temps nécessaire pour effectuer chaque action. Au fur et à mesure de l'utilisation du système est ainsi généré ainsi un fichier de données tracées (ou traces). La présentation de ces données est en général linéaire, chaque ligne contenant une action et éventuellement des paramètres l'accompagnant. On obtient ainsi un bon aperçu, précis et détaillé, de la manière dont est utilisé le système.

Cependant, ces données s'avèrent souvent délicates à manipuler. En effet, les études en apprentissage de la langue qui ont déjà eu recours à des programmes de traces dans le but de suivre la progression et les choix des apprenants (Noblitt & Bland, 1991; Hasebrook & Fezzardi, 1996; Lomicka, 1998) se sont toutes trouvées confrontées à la difficulté d'exploiter de manière efficace l'ensemble des données.

D'après Laurier et Renié (1999), les difficultés rencontrées dans le traitement des données tracées sont :

- la grande quantité des données;
- le fait que des données sur les actions de l'apprenant et celles du système soient juxtaposées et ainsi facilement confondues;
- le fait que les actions tracées n'aient pas toutes le même degré d'importance et apparaissent pourtant sur un même plan.

D'autre part, le contenu des fichiers de traces n'est pas toujours « parlant » et n'est souvent compréhensible que par le concepteur du logiciel.

Renié (1999) expose comment, dans le cadre de l'évaluation des stratégies d'apprentissage lors de l'utilisation du logiciel CAMILLE – *L'acte de vente* (Lotin *et al*., 1996 ; Chanier 1996a), les informations des fichiers de traces ont pu être transférées dans une base de données, ce qui a facilité leur exploitation. Il a été ainsi possible de relever des séquences répétées d'informations et de ce fait d'isoler les actions des apprenants. Des types d'utilisations ont été pu être présentés sous forme graphique à partir de traces simplifiées et de l'architecture du logiciel (figure 2.5).

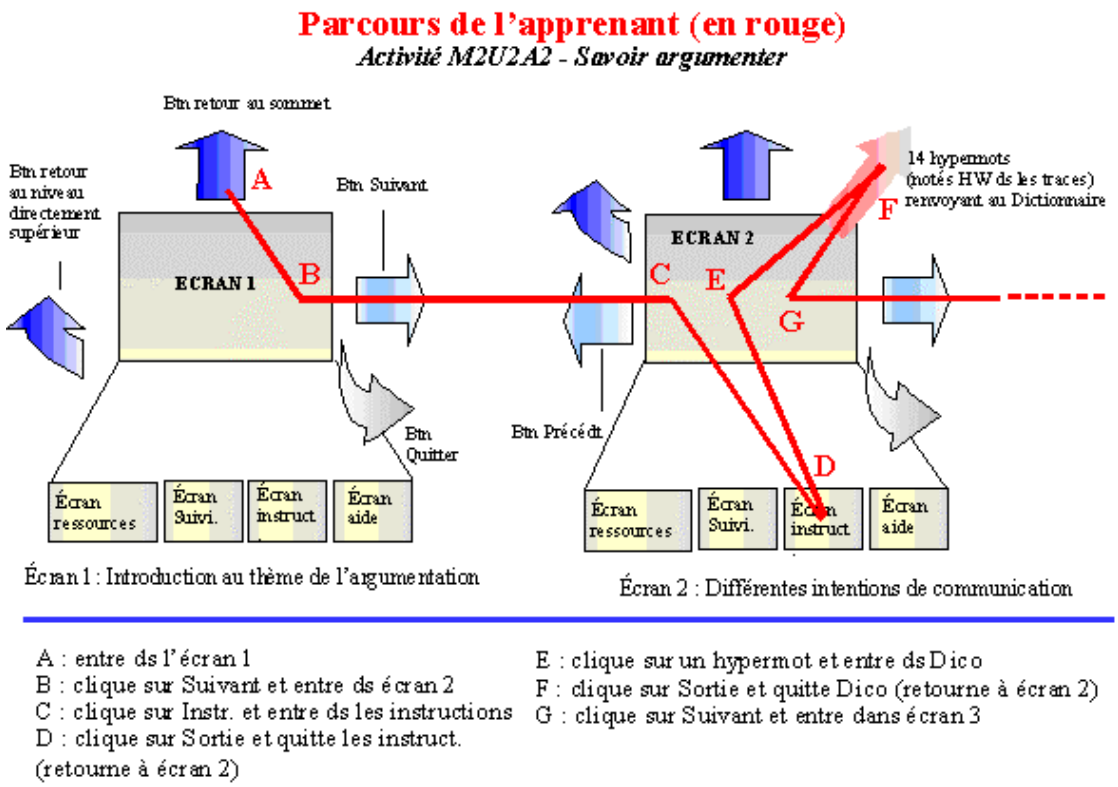


Figure 2.5 : représentation schématique du parcours d'un apprenant dans CAMILLE

Cette présentation rend l'étude des traces accessible à tout le monde, enseignant ou apprenant, ce qui permet véritablement d'examiner, en plus de ce qui est appris, la manière d'apprendre de chaque utilisateur. On peut par la suite tenter des rapprochements entre les types d'utilisations et les variables personnelles des apprenants (niveau de langue, âge, sexe, etc.). Il a ainsi été établi que le niveau de langue n'influençait pas l'utilisation du système, ce résultat devant toutefois être confirmé par un échantillon statistiquement plus significatif (testé sur 23 sujets).

Un autre problème est qu'aussi perfectionnés que soient les fichiers de traces, ils ne permettent pas de savoir ce que pense l'étudiant de ce qu'il fait, pourquoi il utilise tel ou tel dispositif à la place d'un autre, dans quel but et ce qu'il en attend. Dans certains cas, il n'est pas possible de savoir si l'étudiant utilise telle ou telle fonctionnalité à dessein ou si, au contraire, il n'en a pas une mauvaise compréhension. Pour pouvoir répondre à ces questions, il reste alors nécessaire d'utiliser des données verbales (transcriptions de commentaires simultanés enregistrés par vidéo par exemple) et de les faire correspondre avec les traces. Gay & Mazur (1993) soulignent la complémentarité de données telles que les traces, les enregistrements audio et les enregistrements vidéo. Ces nouvelles données sont moins fiables puisque collectées par des observateurs humains mais peuvent apporter des éclaircissements sur telle ou telle utilisation *a priori* incohérente.

3 Conclusion

Nous avons vu dans ce chapitre l'apport des environnements informatiques pour l'acquisition générale. Nous avons constaté que ces environnements pouvaient se partager en deux catégories correspondant à deux générations de produits.

La première génération n'a pas pu concrétiser, à cause de moyens techniques limités, les concepts pédagogiques parfois pertinents qu'elle avait dégagés pour ces environnements. Ainsi, la plupart des programmes visaient surtout à tester ce qui avait été appris, plutôt qu'à enseigner de nouvelles connaissances.

L'évolution de la technique a permis de développer l'interactivité ainsi que l'incorporation de ressources lexicales favorisant l'apprentissage en autonomie. C'est le cas par exemple des environnements Mayday et Lexica qui organisent un ensemble de ressources autour de principes pédagogiques bien cernés. Ces principes sont bien évidemment essentiels dans le cadre de l'ALAO, car il a été établi par l'expérience que les environnements où ils font défaut, tel GLOSSER dans sa version actuelle, se cantonnant à leurs ressources, ne pouvaient pas favoriser l'apprentissage.

Le chapitre suivant est consacré à l'étude d'une ressource lexicale primordiale dans le cadre de l'apprentissage lexical : le dictionnaire. Quel est son rôle dans cet apprentissage ?

CHAPITRE 3

Le rôle des dictionnaires dans l'apprentissage lexical

En matière d'apprentissage lexical, le dictionnaire semble être un outil très utile, voire indispensable de par la quantité d'informations qu'il contient. Comme nous allons le voir plus loin, les apprenants sont familiers avec cet ouvrage et l'utilisent bien plus que, par exemple, une grammaire. Néanmoins, il convient d'étudier plus précisément en quoi le dictionnaire peut favoriser la maîtrise lexicale et si ceux disponibles sur le marché conviennent aux apprenants. Dans un premier temps, nous étudierons le rapport qu'entretiennent les apprenants avec le dictionnaire (s'en servent-ils souvent, qu'y cherchent-ils, ont-ils une préférence pour tel type de dictionnaire, etc.), puis nous étudierons, à travers les dictionnaires spécialisés pour apprenants (en anglais car leur apport est le plus significatif dans le domaine), la manière dont ceux-ci sont conçus pour aider les utilisateurs vis-à-vis du problème lexical qu'ils ont à résoudre. Enfin, avec le développement de l'informatique grand public qui a permis l'émergence des dictionnaires électroniques, nous étudierons ce qu'apportent ces nouveaux outils par rapport aux traditionnels dictionnaires papier.

1 Dictionnaires et apprenants

1.1 Situations d'utilisation du dictionnaire

En premier lieu, en quelles situations est utilisé le dictionnaire ? Est-il plus utilisé pour l'oral ou pour l'écrit ? En compréhension ou en production ?

D'après Béjoint (1981), il est utilisé de préférence lors d'une tâche de décodage écrit (version, lecture) ; ensuite, Béjoint classe par ordre de fréquence décroissante l'encodage écrit (thème rédaction), le décodage oral (compréhension orale) et enfin l'encodage oral (expression orale). Cet ordre est confirmé par Hartmann (1983) pour lequel une très large majorité (enseignants et étudiants d'allemand langue étrangère) utilisent le dictionnaire pour comprendre, traduite et rédiger des textes écrits tandis que 20% d'entre eux seulement s'en servent lors de compréhension de textes oraux ou pour participer à une conversation. De

même, dans l'enquête de Bogaards (1988) concernant 371 étudiants néerlandophones, le dictionnaire est plus utilisé pour une tâche écrite, même si l'encodage prime sur le décodage. Cette préférence du dictionnaire pour les tâches écrites semble normale vu qu'il est lui-même sous forme écrite et que sa consultation demande un certain temps, temps dont on dispose rarement dans une conversation ou une écoute, au contraire d'une lecture ou d'une traduction.

1.2 Possession du dictionnaire

Le dictionnaire est un ouvrage très répandu parmi les apprenants. L'enquête d'Ibrahim et Zalessky (1989) concerne, entre autres, la possession d'un dictionnaire monolingue de français. Il en ressort que tous les scolaires consultés en France (un peu plus de 200 personnes du primaire au lycée) possèdent un dictionnaire, et que, hors de France, seule une minorité (parmi un public apprenant le français dans des instituts d'enseignement supérieur) n'en possède pas (un sixième des étudiants consultés en Chine, un dixième aux États-Unis et un huitième au Japon). Il ne faut pas oublier que pour ces derniers, il s'agissait d'un dictionnaire de langue étrangère.

Cette proportion est plus forte chez Bogaards (1988) qui constate que près d'un quart des étudiants néerlandais en première et deuxième année de français en université ne possède pas de monolingue français et que près de 10% des étudiants en deuxième année et au-delà ne possèdent pas de bilingue. Ici se dessine la tendance selon laquelle la possession du monolingue augmente au détriment du bilingue au fur et à mesure que le niveau s'élève. Les étudiants déclarent posséder en moyenne 2,6 dictionnaires et 6 % n'en possèdent aucun, ce qu'on peut considérer comme assez élevé.

Une troisième enquête, menée par Chi (1998) sur 67 étudiants hong-kongais apprenant l'anglais à l'Université, montre qu'un seul d'entre eux ne possède aucun dictionnaire tandis que 40% (la plus forte proportion) en ont de trois types différents (monolingue, bilingue ou semi-bilingue et électronique).

1.3 Utilisation effective

Pour autant, la possession n'implique pas forcément l'utilisation. Bogaards (1988) fait le point des recherches menées à cette date. Les résultats montrent qu'une très forte proportion déclare se servir du dictionnaire au moins une fois par semaine : pour Hartmann, 98% des étudiants et 87 % des enseignants se servent du bilingue, pour Béjoint, 92 % se servent du monolingue et pour Baxter (1980) , 97 % se servent du bilingue pour la version uniquement. Ce dernier fait en effet la différence avec le thème, pour lequel le dictionnaire est bien moins utilisé (24,5 % pour le bilingue et 19,5 % pour le monolingue). Concernant l'utilisation quotidienne, les résultats sont beaucoup plus disparates ce qui peut s'expliquer, selon Bogaards, par le fait qu'on ne distingue pas suffisamment les types de dictionnaires dans les questions et que les choix proposés aux questions de fréquence varient d'un test à l'autre et manquent de précision.

Dans sa propre enquête, Bogaards fait la différence entre le monolingue et le bilingue. Le bilingue est le plus utilisé (au moins une fois par semaine pour 97 % contre 60 % pour le

monolingue) mais la fréquence diminue avec le niveau : plus celui-ci s'élève et plus l'emploi du monolingue augmente. Ce qui confirme les résultats vus plus haut de la possession de dictionnaires.

Il est difficile de se prononcer sur les résultats de l'enquête menée par Ibrahim et Zalessky, les questions de fréquence n'étant pas assez précises et leur appréciation variant d'un niveau à l'autre. Quant à Chi, 66 % des étudiants déclarent se servir très souvent du dictionnaire en temps d'étude tandis que ce taux chute à 24 % pendant le temps libre ou de loisir.

1.4 Nature des informations recherchées

Que cherche-t-on dans le dictionnaire ? D'après Bogaards (1988) résumant les enquêtes antérieures, c'est le sens et les définitions des vocables qui sont les premières préoccupations. Puis viennent les synonymes et l'orthographe. Ensuite les informations grammaticales et l'emploi des vocables en contexte. Suit la prononciation. Ces trois derniers éléments sont surtout demandés par les apprenants étrangers, moins par les locuteurs natifs. Enfin, l'étymologie, les niveaux de langues ou les homonymes ne sont presque jamais demandés. Peut-être l'intérêt de ces informations n'a-t-il pas suffisamment sauté aux yeux des apprenants.

Ces résultats sont confirmés par l'étude de Chi (1998) pour qui la principale information recherchée dans un dictionnaire monolingue est la définition, puis l'orthographe. Viennent ensuite les informations grammaticales et les exemples. Par contre, près d'un étudiant sur deux n'a jamais exploité les informations sur la prononciation ou sur l'usage du vocable.

Atkins et Varantola (1997) étudient la façon dont est utilisé le dictionnaire pour une aide en traduction lors de deux expérimentations, l'une menée en 1991 lors de l'atelier EURALEX à Oxford sur l'utilisation du dictionnaire et concernant 71 participants de langues maternelles très variées et de niveaux en anglais différents, et l'autre en 1993 menée par Varantola sur un groupe de 32 étudiants finlandais. Les réponses aux questions ont permis de constituer une base de données décrivant les détails d'un millier de consultations dans les dictionnaires, ce qui est statistiquement significatif. Les utilisateurs étaient tous assistés d'un binôme qui notait soigneusement la manière dont le premier utilisait le dictionnaire. Il ne s'agissait donc pas d'une recherche par introspection.

Comme on pouvait s'y attendre, la traduction d'un vocable que ce soit dans le sens L1-L2 ou L2-L1 (43 % et 59 % des consultations) est de loin la première information recherchée. La vérification du vocable que les utilisateurs pressentaient comme traduction correcte vient en seconde position avec un tiers des consultations. Loin derrière viennent les informations concernant les collocations (11 % et 4 %) et encore plus loin, avec 4 %, les informations grammaticales. Les autres types d'informations représentent 5 %. Si l'on prend en compte le niveau des utilisateurs, on constate que plus le niveau est faible et plus le dictionnaire est utilisé pour trouver une traduction, tandis que la proportion de vérification baisse. Les pourcentages concernant les collocations et la grammaire restent comparables aux chiffres précédents, à ceci près que les utilisateurs les plus faibles ne recherchent pratiquement jamais les informations sur les collocations (1 %).

Harvey et Yuill (1997) ont, quant à eux, étudié l'utilisation d'un monolingue anglais pour apprenants, le Collins Cobuild, en production uniquement. L'étude a porté sur 211 étudiants d'anglais langue étrangère, ce qui a permis d'enregistrer 582 consultations. Les apprenants devaient composer des textes sur un thème donné qu'ils choisissaient parmi quatre. Il apparaît que l'information la plus recherchée est l'orthographe (24,4 %) puis le sens (18,3 %). Plus loin vient l'existence du vocable (12,8 %). La synonymie, les informations grammaticales, le niveau de langue, les collocations font l'objet de 10 % des consultations en moyenne chacune. A la lumière de ces résultats, il semble bien que le besoin des étudiants soit surtout d'ordre sémantique bien plus que grammatical (10,5 %) ou collocationnel (8,2 %), contrairement à ce qu'affirmait Béjoint (1981) pour lequel 53 % de ses étudiants ouvraient le dictionnaire pour des informations concernant la grammaire.

A propos de l'orthographe, Ibrahim et Zalessky (1989) notent que les besoins en orthographe varient suivant les pays : ainsi, les américains semblent beaucoup plus préoccupés par ce problème que les chinois ou japonais.

1.5 Bilingue ou monolingue ?

Un point qui semble sûr, c'est que la majorité des étudiants et utilisateurs préfèrent le bilingue au monolingue. Il suffit d'ailleurs d'aller dans n'importe quelle librairie pour constater que les dictionnaires bilingues sont bien plus nombreux que les monolingues et ont un poids économique plus important.

Pour Bogaards (1988), reprenant les résultats d'enquêtes antérieures, les apprenants préfèrent le bilingue. Bogaards (1991) constate que les utilisateurs consultent plus de vocables avec le bilingue. L'utilisation du bilingue pour une tâche de traduction L1-L2 donne de meilleurs résultats. En effet, il est difficile de trouver avec un dictionnaire monolingue la traduction d'un vocable dans la langue cible si on n'en a aucune idée. D'autre part, on trouve encore moins de traduction lorsqu'on ne dispose d'aucun dictionnaire. En matière d'apprentissage, le bilingue est beaucoup moins bénéfique. En effet, le nombre de vocables retenus quinze jours plus tard est beaucoup plus faible que celui de traductions trouvées (le groupe utilisant le bilingue a bien traduit 13,2 vocables mais n'en a retenu que 8,2 ; les résultats sont de 7,6 bonnes traductions et 8,8 vocables retenus avec le DFLE ; pour PR les résultats sont 8,0 et 7,6 ; quant au groupe n'utilisant aucun dictionnaire, les résultats sont 5,6 et 7,1). Cependant en gain, le groupe bilingue se classe entre les deux monolingues (2,7 vocables appris contre 3,3 pour le DFLE, 2,1 pour PR et 1,6 pour ceux qui n'utilisaient aucun dictionnaire). La différence entre le DFLE et PR peut s'expliquer par le fait que ce dernier n'est pas un dictionnaire pour apprenant : on peut relever, entre autres, que les définitions sont véritablement rédigées pour des natifs, aucun effort n'étant entrepris pour simplifier le vocabulaire définitoire. Cependant, le faible nombre d'item (17) étant trop petit, ces résultats ne sont pas significatifs statistiquement. On ne peut que constater une tendance avec beaucoup de prudence.

Cette tendance est toutefois confirmée par Atkins et Varantola (1997) : on consulte beaucoup plus avec le bilingue (714 fois contre 281), que ce soit pour traduire en thème ou en

version et la proportion de recherches fructueuses est plus élevée avec le bilingue (64 % contre 48 %). Toutefois la proportion de consultation n'est pas la même en fonction de l'information recherchée : le monolingue est plus utilisé que le bilingue pour trouver des informations grammaticales ou collocationnelles. D'autre part, l'étude a montré que la proportion d'utilisation du monolingue était plus importante au fur et à mesure que le niveau augmentait. Il faut donc avoir un certain niveau de langue pour tirer profit des informations du monolingue. L'étude livre un dernier aspect intéressant de l'utilisation des deux types de dictionnaire. On constate une différence d'utilisation lorsque la recherche d'informations se fait en plusieurs fois (plusieurs accès). Le premier dictionnaire ouvert est très majoritairement le bilingue. Par contre, lors du deuxième accès, cette proportion diminue de manière plus ou moins sensible suivant les groupes. Elle tend à s'équilibrer pour le groupe hétérogène d'Oxford tandis que le monolingue est le plus utilisé dès le deuxième accès dans le groupe homogène finlandais. Cela laisse supposer que lorsqu'on ne trouve pas l'information recherchée dans le bilingue, on se rabat vite sur le monolingue. Les auteurs ne précisent toutefois pas les raisons d'un tel basculement, l'analyse détaillée de tous les enregistrements demandant trop de temps pour pouvoir donner des réponses satisfaisantes. Quant aux gains en acquisition, il n'est malheureusement pas possible de comparer avec la précédente étude, puisqu'Atkins et Varantola ne se sont pas préoccupés de cet aspect.

Une troisième étude, Laufer et Hadar (1997), vient aussi confirmer ces résultats. Cette fois-ci s'ajoute un troisième type de dictionnaire, le semi-bilingue, sur lequel très peu d'études ont été menées à ce jour et qui est un terrain ouvert à l'investigation car les résultats obtenus par ce type de dictionnaire sont prometteurs. Les dictionnaires semi-bilingues sont un mélange des deux précédents dans le sens où, à la suite de la définition monolingue pour apprenant en langue seconde, se trouve la traduction du sens du vocable considéré. L'utilisateur est donc assuré d'avoir compris la définition en langue étrangère, ce qui est confortable, à la condition toutefois de la lire et de faire l'effort de la comprendre, la tentation étant grande de ne regarder que la traduction.

Mais revenons à l'étude. Elle avait pour but d'évaluer les performances des trois types de dictionnaires, monolingue pour apprenants, bilingue et semi-bilingue autant en compréhension qu'en production. Elle a porté sur 123 sujets, lycéens et universitaires étudiant l'anglais langue étrangère, de niveau différents. En compréhension, l'épreuve était un QCM (il fallait trouver l'équivalent de vocables dans une liste) et en production, il s'agissait d'employer chacun des vocables-cible dans une phrase. Les erreurs grammaticales n'étaient pas considérées. Seule la pertinence sémantique comptait. Il fallait que la phrase produite soit significativement différente des exemples qui auraient pu être donnés dans le dictionnaire.

De manière générale, les meilleurs résultats, autant en compréhension qu'en production sont obtenus avec le dictionnaire semi-bilingue. Ce dictionnaire rassemblant les deux types d'informations contenus dans les autres dictionnaires, les utilisateurs sont à même, d'une part de trouver à chaque fois les renseignements nécessaires, et d'autre part, de pouvoir les exploiter. Similairement à ce qui a été vu plus haut, le groupe utilisant le dictionnaire bilingue en production obtient nettement de meilleurs résultats sur le monolingue. Ce résultat est inversé dans le cas de la compréhension où le monolingue est légèrement plus utile que le

bilingue. Il convient toutefois de nuancer ces résultats qui sont fortement dépendants du niveau de langue des utilisateurs. Comme vu plus haut, plus le niveau s'élève et meilleurs sont les scores avec le monolingue, le semi-bilingue obtenant toujours les meilleurs résultats. Ainsi, que ce soit en production ou en compréhension, le groupe comprenant les meilleurs éléments obtient de meilleurs résultats à l'aide du monolingue qu'à l'aide du bilingue. Les résultats sont exactement l'inverse pour le groupe le plus faible, bien que la différence y soit plus marquée. Cette étude a le mérite d'établir des résultats clairs sur une base statistiquement significative. Toutefois on peut se demander si ces résultats ne sont pas un peu biaisés par le fait que les vocables-cible ne sont pas présentés en contexte, mais isolés au sein d'une liste. Ainsi, autant en compréhension qu'en production, les tâches demandées semblent quelque peu artificielles et n'ont que peu de chance d'être reproduites dans la réalité.

Bogaards (1994) récapitule les forces et faiblesses des deux types de dictionnaires.

Les points faibles du monolingue sont :

- le manque de contrastivité. Comme les monolingues sont destinés à des locuteurs de langues différentes et ne décrivent que la langue étrangère, il est impossible de traiter de façon contrastive les « faux-amis » ou les vocables qui prêtent à confusion en comparaison avec une langue maternelle donnée.
- la compréhension. Comme toutes les informations sont données dans la langue étrangère, il y a parfois des problèmes de compréhension. C'est le cas avec les définitions compliquées ou qui bouclent. Même lorsque le vocabulaire est contrôlé et simplifié, comme c'est le cas dans les dictionnaires pour apprenants, il n'est pas toujours facile de saisir le sens précis des vocables. Le problème ne vient toutefois pas toujours des dictionnaires : certains vocables sont en eux-mêmes difficiles à comprendre ou à conceptualiser. Ickler (1982) remarque avec raison que déjà en classe de langue, l'enseignant a parfois des difficultés à faire comprendre tel ou tel vocable, bien qu'il puisse donner des informations de vive voix ou poser des questions de contrôle.
- l'accès aux vocables inconnus. Comment, dans les tâches productives, un apprenant peut-il trouver le vocable qu'il lui faut mais qu'il ne connaît pas ? Nous renvoyons aux études précédentes (Bogaards, 1991 ; Atkins et Varantola, 1997). Ce défaut est encore plus sensible lorsque l'apprenant veut s'exprimer dans la langue étrangère sur des sujets qui sont propres à la culture de sa langue maternelle (Tomaszczyk, 1981).

Les points forts du monolingue sont le grand nombre d'informations différentes disponibles « authentiques » car exprimées dans la langue cible. Il est ainsi possible de voir le comportement réel des vocables dans les définitions et les exemples ainsi que les vocables « satellites » (les actants), c'est-à-dire ceux qui sont reliés aussi bien syntaxiquement que sémantiquement et qui sont, de ce fait, nécessaires pour la maîtrise du vocable initial.

Tandis que le point fort du bilingue est la compréhension, il s'avère, d'après Béjoint (1987) plus ou moins inadéquat dans les cas suivants :

- quand des items comparables dans les deux langues appartiennent à des réseaux sémantiques différents (anglais solicitor, barrister et français avocat, avoué, notaire). L'usage du bilingue est trompeur et l'influence de la langue maternelle est néfaste.
- dans le cas de vocables communs très polysémiques
- quand l'apprenant ne connaît pas le référent, même pas dans sa langue maternelle.

Concernant l'utilité respective des deux types de dictionnaires, Bogaards (1994) résume la situation :

« Tout compte fait, les deux types de dictionnaires, les bilingues comme les monolingues, peuvent être utiles aux apprenants. Sous l'influence d'une méthodologie exclusivement monolingue, qui s'opposait à la place trop importante de la traduction dans les approches traditionnelles et qui cherchait, par peur des inférences, mais en vain, de minimiser les contacts avec la langue maternelle, on a opté pour les monolingues (cf. Gairns et Redman 1986). On ne peut nier, cependant, que les bilingues sont souvent indispensables et parfois très pratiques, mais qu'il est nécessaire (et possible) de les adapter plus aux besoins des apprenants, comme on l'a déjà fait avec beaucoup de succès pour les monolingues. »

Il ressort donc des expériences précédentes que le dictionnaire monolingue demande une meilleure connaissance de la langue. Ce n'est qu'à cette condition que la richesse et la diversité des informations permettent de meilleurs résultats. D'autre part, le problème des monolingues en production est de trouver le vocable inconnu. Cette difficulté résolue, le monolingue est plus profitable.

2 Dictionnaire et compréhension écrite

Plusieurs études (Hartmann, 1983, Bogaards, 1988) ont montré que le dictionnaire est très souvent utilisé au cours d'une traduction ou de la lecture d'un texte dans une langue étrangère (voir plus haut, possession et utilisation). Si l'apprenant ne peut déduire le sens d'un vocable d'après le contexte, le dictionnaire reste son seul recours. Pourtant, son utilité n'est pas toujours évidente : selon Bogaards (1995), plusieurs expériences ont montré que le dictionnaire ne semblait pas améliorer la compréhension des textes d'une manière significative. Pour cela, il avance plusieurs raisons :

- les apprenants n'aiment pas utiliser un dictionnaire. Ils le considèrent comme une étape obligée et contraignante qui les détourne de leur lecture. Ils reculent devant leur complexité et préfèrent bien souvent ne pas s'en servir.
- ils ne savent pas utiliser un dictionnaire. Ils ont des difficultés à repérer l'information pertinente et acceptent la moindre indication qui va dans le sens de leur hypothèse initiale de manière à abréger l'« épreuve ». De plus, pour les monolingues, ils sont souvent dans l'obligation d'aller consulter d'autres entrées pour comprendre la première, soit par référence explicite, soit parce que la première définition contient des vocables peu ou mal connus. Ils ont alors toutes les chances de perdre le fil du texte. D'autre part, Müllich (1990) a constaté que, pour des raisons très différentes, la moyenne des réussites n'était que de quelque 50 %, ce qui veut dire que les réponses erronées,

inadéquates ou nulles sont à peu près aussi fréquentes que les réponses acceptables (Bogaards, 1994).

- le dictionnaire nuit au processus de lecture : des expériences (Bensoussan *et al.*, 1984, Nesi et Meara, 1991) montrent que des étudiants utilisant un dictionnaire mettaient souvent plus de temps à terminer leur tâche, sans pour autant obtenir de meilleurs résultats. Selon Müllich (1990), plus un apprenant met de temps à chercher une information, moins il a de chance de la trouver.

Face à ce constat, Bogaards en déduit qu'il faut, d'une part, avoir un niveau de connaissance avancé sur la langue pour pouvoir profiter des informations contenues dans les dictionnaires, et d'autre part, avoir une bonne dose de ténacité et de courage.

D'autre part, comme le suggère de nombreux lexicographes et professeurs de langue, il est nécessaire de former les apprenants à l'emploi efficace des dictionnaires (Béjoint, 1981 ; Kipfer, 1987). Les préfaces et introductions de dictionnaires contiennent souvent beaucoup d'informations, parfois très détaillées, mais il apparaît (Bogaards, 1988) que près de la moitié des utilisateurs (34 % pour Béjoint et 42 % pour Bogaards ; 58 % occasionnellement) ne les ont jamais lues. Par ailleurs, Bogaards (1994, pp. 219) relève que « selon une enquête menée auprès de plus de mille usagers du dictionnaire à travers l'Europe, le pourcentage de ceux qui n'ont reçu aucune explication précise ou systématique sur la manière d'utiliser les dictionnaires est de plus de 60 (près de 80 % en France ; Atkins et Knowles, 1988). »

Nous allons maintenant étudier, parmi tous les types existants de dictionnaire, ceux qui sont conçus pour l'apprentissage du lexique.

3 Dictionnaires pédagogiques de l'anglais

1995 a été un bon millésime pour la lexicographie pédagogique en anglais, puisque quatre nouveaux dictionnaires ont vu le jour : trois sont des rééditions de dictionnaires déjà existants (2^{e} édition pour le Collins Cobuild English Dictionary (COBUILD), 3^{e} édition pour le Longman Dictionary of Contemporary English (LDOCE) et 5^{e} édition pour le Oxford Advanced Learner's Dictionary (OALD)) et un 4^{e} a vu le jour : le Cambridge International Dictionary of English (CIDE). Cette profusion de dictionnaires dénote d'une grande vitalité de la lexicographie anglaise, en partie due à la compétition économique que se livrent ces quatre éditeurs, ce qui fait dire à Herbst (1996) que l'anglais est probablement, d'un point de vue lexicographique, la langue la mieux décrite au monde.

Plutôt que de faire un comparatif de ces quatre dictionnaires, nous évoquerons les spécificités des dictionnaires pour apprenants, les problèmes d'utilisation que rencontrent les apprenants et les réponses que ces ouvrages apportent.

Le dictionnaire pour apprenant est un monolingue destiné aux personnes apprenant une langue étrangère. Il se différencie sur un certain nombre de points du monolingue pour natif. Davantage de précisions sont données sur les constructions grammaticales, sur la fréquence d'emploi ou sur les notions pragmatiques comme le registre. En principe le natif est déjà au fait de ces informations qui n'ont pas besoin d'être mentionnées dans les monolingues

normaux. Cependant, pour des étrangers, elles sont des indices précieux sur le maniement de la langue.

Mais le point le plus significatif et le plus visible est l'utilisation d'un vocabulaire définitoire contrôlé pour décrire les entrées dans les définitions. Il est nécessaire en effet de définir et d'expliquer des vocables inconnus avec des vocables simples, que l'apprenant est susceptible de connaître déjà. Car s'il doit parcourir d'autres articles, l'efficacité du dictionnaire s'estompe et son utilité disparaît comme nous l'avons vu plus haut avec le problème des définitions circulaires. Nous allons étudier cet aspect un peu plus loin.

Les quatre dictionnaires ont une caractéristique commune qui devient maintenant quasi-obligatoire lors de leur construction : ils sont tous élaborés à partir de corpus. COBUILD exploite the Bank Of English, un corpus de 200 millions de mots à l'époque où le dictionnaire est sorti et qui comprend maintenant (la dernière publication remonte à juillet 1998) 330 millions de mots. OALD et LDOCE ont collaboré et exploitent, quant à eux, le British National Corpus (100 millions de mots) qui est incomparable pour sa couverture de la langue orale et qui comprend notamment le Longman Lancaster Corpus (30 millions de mots sur l'anglais oral britannique et américain). Il est toutefois difficile d'évaluer précisément les utilisations qui en ont été faites. L'utilisation de corpus, par le biais de concordances, permet de mieux cerner les sens des entrées et en particulier l'ordonnancement des lexies, les plus attestées en premier.

Pour Bogaards (1996), les problèmes d'utilisation du dictionnaire se regroupent sous trois thèmes : la compréhensibilité, l'accès lexical et l'aide à la production.

La compréhensibilité concerne non seulement la clarté des définitions, mais aussi celle des exemples ainsi que les illustrations et autres moyens de rendre les significations plus faciles à comprendre.

L'accès lexical concerne le nombre d'articles, puisqu'il faut que l'apprenant trouve effectivement le vocable qu'il cherche, et l'organisation des homonymes et des dérivés entre eux ainsi que les collocations et expressions semi-figées. Il concerne aussi les dispositifs mis en place pour aider l'apprenant à trouver les vocables et expressions qu'il ne connaît pas, ce qui est particulièrement important en production. Enfin, il concerne les vocables polysémiques et l'accès aux lexies : quels sont les indices qui orientent l'apprenant sur tel ou tel sens d'un vocable polysémique, spécialement lorsque l'article est long ?

L'aide à la production, en dehors des problèmes d'accès, concerne surtout les informations grammaticales, les synonymes ou antonymes, les fréquences, les registres et les autres aspects pragmatiques. Enfin, il faut aussi examiner l'utilité des exemples.

3.1 La compréhensibilité

3.1.1 Les définitions

Le concept de vocabulaire contrôlé à grande échelle a été introduit avec la première version de LDOCE en 1978 et a depuis remporté un grand succès puisque tous les autres

dictionnaires anglais pour apprenant l'ont adopté. Ils varient néanmoins en taille et en nature. Les listes ne comprennent pas les mêmes vocables. Concernant la taille, le vocabulaire contrôlé du LDOCE est de 2 000 vocables, celui du OALD de 3 500 et CIDE affirme moins de 2 000. Quant à COBUILD, sa politique est moins claire puisqu'il affirme que « la plupart des vocables dans les définitions font partie des 2 500 plus communs de la langue anglaise » (p. xviii). Néanmoins, il convient de parler plutôt de famille de mots, car tous les dictionnaires ne font pas figurer les dérivés dans leur liste de vocabulaire. LDOCE utilise même les affixes : ainsi independence ne se trouve pas dans la liste car on peut le construire avec in-+depend+-ence. Étant donné qu'il y a en moyenne 1,6 vocable par famille (Nation 1983a), les vocabulaires définitoires sont donc de l'ordre de 3 200 à 5 600 éléments. Il y a pourtant des cas où ce vocabulaire n'est pas suffisant. Les éditeurs se réservent en effet le droit d'employer (modérément) des vocables hors des listes, mais ceux-ci sont alors démarqués typographiquement des autres constituants de la définition. LDOCE et OALD utilisent des petites capitales et CIDE donne en plus une indication comme dans (Herbst, 1996) :

> A cream tea is a light meal of SCONES (= a type of bread) with JAM (= a sweet soft substance made by cooling fruit with sugar) and cream.

Ce procédé est utile car il évite d'aller consulter d'autres articles, mais il a l'inconvénient, dans certains cas, d'alourdir de façon notable les définitions.

D'après Bogaards (1996), COBUILD est le dictionnaire s'éloignant le plus de la politique du vocabulaire contrôlé. D'après ses classements par fréquence (la fréquence est indiquée par une échelle de 0 (le moins fréquent) à 5 diamants (le plus fréquent)), les 2 500 vocables les plus fréquents ont au moins trois diamants. Or Bogaards, parcourant les définitions de 73 lexies allant de pocket à point, a relevé 19 vocables définitoires n'ayant pas ces trois diamants.

En 1987, COBUILD a jeté un pavé dans la mare en adoptant de manière généralisée le format sous forme de phrase pour les définitions. Ce fut une véritable révolution en lexicographie car jusqu'alors, la définition devait toujours être substituable au vocable dans le texte. L'avantage est de pouvoir indiquer de manière transparente et naturelle, sans avoir recours à un métalangage ou à des codes, un certain nombre d'informations comme les structures grammaticales des verbes, les actants (il est facile par exemple de déterminer les contraintes sémantiques sur les actants en employant somebody ou something), les contextes d'emploi ou les collocations. Ce format a par contre l'inconvénient de ne pas être toujours directement adaptable au contexte du vocable en question que l'on cherche à comprendre dans le texte. Par ailleurs, il peut être délicat dans certains cas de délimiter l'ensemble des vocables possibles pour tel ou tel actant et de les regrouper sous un même hyperonyme. Par exemple, pour to employ, on a

> 1 If a person or a company employs you, they pay you to work for them.

Ici, l'employeur est une personne ou une entreprise. Mais il peut très bien s'agir de l'État, d'une administration ou d'une association, etc. La définition n'est pas fausse en soi mais il faut comprendre que la personne en question est en fait une personne morale, ce qui n'est pas forcément trivial.

3.1.2 Les exemples

Comme le souligne Herbst (1996), la vieille controverse sur les mérites respectifs des exemples authentiques ou inventés voit le jour sous un nouvel angle avec l'utilisation de corpus dans l'élaboration des dictionnaires. En effet, d'une part, l'extension du corpus de COBUILD de 20 (pour la première édition) à 200 millions de mots (pour la deuxième) fait qu'on trouve forcément un exemple approprié et que « la majorité des exemples dans le dictionnaire sont tirés mot à mot de l'un des textes de la Bank of English, des modifications mineures ayant été faites seulement occasionnellement » (COBUILD, p. xxii). D'autre part, même les exemples fait à la main sont de plus en plus influencés par les corpus. Les exemples sont utilisés différemment par les éditeurs. Pour LDOCE et OALD, les exemples sont inventés; pour COBUILD, ils sont extraits des textes et reflètent donc la langue authentique; pour CIDE, c'est un mélange des deux. Ces politiques ont plusieurs conséquences. COBUILD est le dictionnaire qui utilise le plus d'exemples et LDOCE le moins (près de 35 % moins que COBUILD). Pour COBUILD, ils servent principalement d'illustration de l'entrée définie. LDOCE et OALD leur vouent un rôle plus pédagogique, les utilisant pour suppléer d'une part la définition, et d'autre part, pour montrer les propriétés grammaticales et collocationnelles de l'entrée. De ce fait, les exemples tendent à être plus stéréotypés, mais ils sont également plus courts. Autre conséquence importante, surtout en compréhension, les exemples authentiques à la COBUILD contiennent un nombre bien plus important de vocables hors du vocabulaire contrôlé des définitions, voire des vocables de basse fréquence (paganism ou filing cabinets, Bogaards, 1996). LDOCE, OALD et CIDE en utilisent aussi, mais en moins grande quantité. OALD indique (p. xvi) que certains exemples du corpus ont été modifiés afin d'enlever les vocables difficiles.

Laufer (1992) a étudié les rôles respectifs des différents types d'exemples et de l'influence de la définition et des exemples dans la compréhension. Ces résultats montrent que l'exemple seul ne suffit pas pour comprendre, mais que, dans ce cas de figure, les exemples inventés conduisent à de meilleurs résultats. Lorsqu'on ajoute la définition comme élément d'information, les résultats sont bien meilleurs. Mais là encore, on constate que les exemples inventés sont significativement plus profitables. Laufer en conclue, d'une part, que l'exemple inventé apporte plus d'information, même si le vocable est expliqué ; d'autre part, que la définition seule est supérieure à l'exemple seul. Enfin que « le bénéfice possible que l'on peut tirer des exemples inventés, par opposition aux exemples authentiques, est moins dépendant du niveau de vocabulaire de l'utilisateur ». En d'autres termes, il est nécessaire d'avoir une meilleure connaissance de la langue, voire de la culture associée, pour pouvoir tirer parti des exemples authentiques. Les résultats de cette étude sont à rapprocher de ceux de Harvey et Yuill (1997) selon lesquels les exemples servent véritablement à l’élucidation du sens. Les apprenants ont même tendance à chercher le sens des entrées dans les exemples, y compris lorsque le vocable est défini par une phrase comme dans le COBUILD. Il est donc logique de penser que les exemples inventés qui ont des vertus pédagogiques soient plus profitables que les exemples réels destinés surtout à illustrer la définition.

En production, Laufer aboutit à des conclusions différentes. Dans ce cas, les exemples inventés sont légèrement, mais non pas de manière significative comme en compréhension, plus profitables. La performance des sujets est donc moins sensible au type d'exemple.

3.1.3 Les illustrations

Avec les tables et les diagrammes, l'illustration est très utile dans les dictionnaires pour apprenants, puisqu'elle supplée la définition lorsque les moyens purement linguistiques sont insuffisants pour expliquer un vocable ou produisent une définition trop lourde et qu'elle présente d'autres vocables reliés sémantiquement au premier comme les co-hyponymes ou les composants (relation partie-tout). Cette dernière fonctionnalité est surtout utile en production. On peut ainsi connaître différents types de lits ou les différentes parties comestibles d'un bœuf. Là encore, les politiques éditoriales varient. COBUILD n'utilise aucune illustration tandis que les trois autres en font bon usage, allant jusqu'à de grandes pages en couleur illustrant un thème comme (dans LDOCE) les fruits, le bureau ou la conduite. Il n'apparaît aucune justification dans aucun des trois dictionnaires sur le choix des entrées illustrées. Celles-ci varient d'un ouvrage à un autre et d'autre part, on peut être surpris (Bogaards, 1996) que certaines illustrations soient utilisées pour des vocables faisant partie du vocabulaire contrôlé, qui en principe ne doit pas poser problème. Cela laisse penser que ces illustrations sont utiles surtout en production.

D'autre part, s'il existe des renvois de certaines entrées aux illustrations lorsque celles-ci sont distantes dans le dictionnaire (c'est le cas par exemple, des 24 illustrations pleine page de LDOCE), ceux-ci ne sont pas systématiques. Ainsi si les entrées de sink et photocopier renvoient à leur illustration, ce n'est pas le cas de kitchen pour LDOCE ou toggle (vers Dressing and undressing) pour CIDE.

Les informations d'ordre graphique sont quasiment toujours des illustrations concernant des objets concrets. Rien n'est spécifié pour des entrées plus abstraites. Aucun réseau sémantique (synonymes, antonymes, hyper/hyponymes) n'est utilisé. Il serait intéressant, par exemple, d'avoir des graphes représentant des thèmes sémantiques, tels que, dans le champ notionnel du travail, les synonymes de chef, les différents types de salaires, etc. Dans ce domaine, il nous semble que beaucoup reste à faire. Nous explorerons cet aspect plus en détail dans le chapitre 6.

3.1.4 Autres dispositifs d'aide à la compréhension

Les quatre dictionnaires utilisent des renvois analogiques vers d'autres vocables lorsqu'ils estiment qu'une comparaison est nécessaire pour bien saisir les nuances. Les vocables auxquels on renvoie sont des synonymes, antonymes, vocables composés ou présentant des similarités morphologiques. Les renvois sont le plus souvent introduits par des see also ou compare qui ont l'inconvénient de n'être parfois pas assez explicites sur le type de la relation décrite. On aimerait bien savoir par exemple la relation entre emigrant et immigrant (sont-ils synonymes ou pas ?) Si l'utilité de ces renvois est incontestable, les relations sémantiques devraient donc être mieux repérables, plus directes et moins ambiguës. Ces critiques

s'appliquent surtout à LDOCE, OALD et CIDE, moins à COBUILD. En effet, ce dernier utilise sa colonne supplémentaire et les renvois sont d'une part, beaucoup plus repérables, car ils sont isolés du texte, et d'autre part la nature de la relation est beaucoup plus claire (synonymie et antonymie identifiées par les symboles = et). En outre, ces renvois sont bien plus nombreux, d'après Bogaards, que ceux des trois autres dictionnaires. On pourra cependant regretter que les mots sur lesquels on dirige l'utilisateur soient toujours des vocables et non des lexies. Une correspondance de sens à sens aurait évité le parcours du nouvel article.

Tout comme les illustrations, aucune politique de renvoi n'est définie et l'utilisateur ne peut compter sur aucune systématicité, même si COBUILD, encore une fois, est le plus satisfaisant sur ce point.

A côté de ces renvois, OALD, LDOCE et CIDE utilisent des notes d'usage qui discutent de manière contrastive des points précis ou des difficultés auxquelles peuvent être confrontés les apprenants. Ainsi LDOCE compare les différences entre famous, well-known, distinguished, etc., OALD le fait pour to giggle, to snigger et to titter, et CIDE pour expensive, costly, dear, etc. A l'évidence, ces notes sont d'une grande utilité pour les apprenants, notamment en production, mais une fois de plus, celles-ci apparaissent de manière aléatoire et imprédictible tout au long du dictionnaire. COBUILD ne fait pas usage de ces notes, comptant sur ses renvois dans la colonne supplémentaire et sur ses descriptions détaillées. Cependant, les définitions ne précisent pas toujours suffisamment les nuances comme grubby, grimy et filthy décrits respectivement par « rather dirty », « very dirty » et « very dirty indeed ».

3.2 L'accès lexical

L'accès lexical concerne deux aspects de la rédaction de tout dictionnaire : la macrostructure, c'est-à-dire l'organisation des entrées dans le dictionnaire les unes par rapport aux autres, et la microstructure, c'est-à-dire la description à l'intérieur de l'entrée elle-même, l'organisation des lexies, des informations grammaticales, etc. Concernant la macrostructure, il s'agit d'avoir accès au bon vocable le plus efficacement possible. Il ne s'agit pas uniquement de confort mais de nécessité lors d'une tâche de compréhension, car le temps passé hors de la lecture doit être le plus court possible pour ne pas perdre le fil du texte.

3.2.1 La macrostructure

Bogaards (1996) s'est livré à une estimation du nombre total des entrées dans chaque dictionnaire. Il est clair que plus le dictionnaire en comprendra, plus l'apprenant sera à même de trouver le vocable qu'il cherche.

LDOCE apparaît être le dictionnaire avec le plus d'entrées : de 90 000 à 100 000 tandis que les trois autres comprennent de 70 000 à 75 000 entrées, ce qui est un nombre fort respectable. Ces quatre dictionnaires s'adressent donc visiblement à des étudiants avancés. L'avantage de LDOCE n'est en fait pas véritablement significatif car il s'agit surtout de vocables vieux, démodés ou littéraires. D'autre part , LDOCE mentionne de nombreux dérivés, souvent sans explication indépendante.

L'organisation des entrées à l'intérieur d'un dictionnaire est le plus souvent alphabétique. C'est la politique de LDOCE qui l'applique strictement. Elle a l'avantage d'être claire et non ambiguë. Par contre, elle a le désavantage de scinder les familles de mots, les dérivés syntaxiques, qui, d'un point de vue pédagogique, ont leur importance car il y a une relation plus ou moins régulière entre la forme et le sens. Il est par exemple facile de deviner le sens d'un adverbe à partir de l'adjectif de la même famille. Beaucoup d'adverbes ne sont d'ailleurs pas expliqués, étant simplement placés à côté de l'adjectif correspondant ou renvoyant à cet adjectif. D'autre part, les trois autres dictionnaires regroupent aussi les vocables composés, phénomène fréquent en anglais. Toutefois les politiques ne sont pas toujours cohérentes (Bogaards, 1996). Ainsi, OALD classe le dérivé pocketful aussi bien que les vocables composés pocket knife et pocket money dans l'article de l'entrée pocket, mais classe pocketbook dans une entrée différente. Dans COBUILD, poetically est décrit dans l'entrée poetic et apparaît ainsi avant poetical, qui est une autre entrée. CIDE donne le plus de vocables dans une même entrée mais là aussi, le regroupement n'est pas systématique. D'autre part, ce dernier qui utilise des indices sémantiques (nous allons y revenir un peu plus loin) a parfois du mal à faire correspondre le sens du vocable composé avec la détermination sémantique de l'entrée sous laquelle il est classé (quel rapport sémantique entre dead end et dead [COMPLETE] ?). Le regroupement de vocables sous une même entrée offre donc des avantages en illustrant les relations morpho-sémantiques entre eux, même si une certaine incohérence peut fourvoyer l'utilisateur. Cependant, le phénomène est minimisé par le fait que les vocables, proches alphabétiquement, le sont aussi visuellement sur un dictionnaire papier. Le problème n'est pas le même pour les dictionnaires électroniques.

Le deuxième problème concernant la macrostructure est l'organisation et l'accès aux unités polylexicales, soit les collocations et expressions semi-figées. Les expressions semi-figées n'étant pas des entrées de dictionnaires, bien qu'elles aient un sens et des informations propres (synonymes, variations lexicales, transformations grammaticales, etc.) et qu'elles pourraient être illustrées par des exemples, elles sont décrites dans les articles d'un des mots qui les constituent. Comme deux ou plusieurs candidats sont possibles, il reste à savoir lequel. Pour les collocations, le problème est identique, même si leur statut d'entrée à part entière est moins envisageable. Le placement et la description des unités polylexicales n'ont rien de rigoureux et de systématique et varient fortement d'un dictionnaire à l'autre, tout comme au sein d'un même dictionnaire. Parfois, elles sont expliquées, parfois citées comme exemple, parfois ignorées. Les mêmes expressions sont décrites sous des entrées différentes suivant les ouvrages. Une même expression peut être définie, parfois différemment, à l'article de chacun de ses constituants. Il y a souvent des renvois, mais ceux-ci ne sont pas systématiques. Il faut aussi considérer le problème des variations lexicales. Idéalement, une expression devrait être décrite et expliquée à un seul endroit avec de multiples renvois à partir de chacun de ses constituants pertinents, c'est-à-dire les mots pleins, par opposition aux mots grammaticaux, sans oublier les renvois à partir des variations lexicales. Mais ceci semble être beaucoup plus difficile à mettre en application.

3.2.2 La microstructure

Il s'agit ici d'examiner les moyens mis à la disposition de l'utilisateur pour pouvoir choisir entre les différents sens d'une entrée celui qui correspond le mieux avec le mot du texte (ou tout du moins celui qu'il cherche à comprendre).

En premier lieu, il convient de définir ce qu'est une entrée de dictionnaire. En principe, une entrée correspond à un vocable, c'est-à-dire un mot polysémique dont les sens présentent une certaine unité et fonctionnent de la même manière dans une phrase, c'est-à-dire qu'ils possèdent la même catégorie grammaticale. On met ici le doigt sur la traditionnelle opposition entre l'homonymie et la polysémie. Lorsque les sens d'un vocable deviennent trop différents et que l'on perd la relation entre eux (Mel'cuk *et al*., 1995, parle de « pont sémantique »), il convient de considérer ces sens comme faisant partie de deux vocables différents. C'est le cas, par exemple, de voler, dont les sens principaux être dans les airs et dérober, ne peuvent plus être réunis sous une même entrée, car n'ayant pas assez d'intersection sémantique. Toutefois, le manque de rigueur et de systématicité des langues naturelles fait qu'il est pratiquement impossible de s'accorder sur un nombre précis de lexies par entrée. Ce phénomène se retrouve dans les monolingues anglais.

COBUILD se démarque des trois autres en définissant une entrée selon un critère principalement orthographique. Il n'est ainsi pas fait distinction des catégories grammaticales et une entrée comprendra par exemple les formes verbales et nominales ou nominales et adjectivales du même vocable lorsqu'elles sont identiques, ce qui n'est pas rare en anglais. Les lexies (qui appartiennent de fait à des vocables différents) sont classées par ordre de fréquence, le sens le plus fréquent en premier, critère qui prime sur les catégories grammaticales. Celles-ci sont indiquées dans une colonne supplémentaire à côté des articles qui guide les utilisateurs. Ce choix est toutefois discutable car la catégorie grammaticale, qui est un des éléments les plus saillants d'un vocable et qui est la plupart du temps facilement reconnaissable, est un critère majeur dans la discrimination des sens d'une forme graphique. Nous avons vu au chapitre 1 que la catégorie grammaticale était une caractéristique essentielle et indissociable d'un vocable et que ceux–ci semblaient être stockés en fonction de leur catégorie. En pratique, il faut donc « sauter » les lexies qui n'ont pas la même catégorie que celle recherchée et même si cela se fait plus rapidement par le biais de la colonne supplémentaire, cette étape n'aurait pas lieu d'être si les lexies étaient regroupés d'abord par catégorie (formant un vocable à part entière avec une catégorie définie), puis par fréquence. Cette position est d'ailleurs adoptée par les trois autres dictionnaires. Du point de vue de l'apprentissage, cela clarifie nettement le statut et la fonction d'un vocable : l'apprenant sait qu'il existe le nom play et le verbe to play. Et il perçoit mieux l'unité de sens de ces vocables.

Outre la catégorie grammaticale, LDOCE et CIDE ont innové en intégrant des *signposts* ou des *guide words* dans les lexies. Lorsqu'une entrée est longue, chaque lexie est précédée d'une précision sémantique permettant de discerner rapidement et facilement ses traits sémantiques dominants. Il s'agit le plus souvent d'un synonyme proche, d'un hyperonyme ou d'une expression définissant le domaine d'application. Ce procédé est très utile car il permet d'éliminer les lexies qui manifestement ne correspondent pas avec celle qui est attendue et de

se concentrer sur les autres. Toutefois des difficultés peuvent apparaître. Tout d'abord, ces précisions ne font pas forcément partie du vocabulaire contrôlé et il peut y avoir des problèmes de compréhension. D'autre part, il n'est pas toujours possible de « coller » à la lexie définie et les *signposts* sont parfois vagues ou leur sens est trop large. Néanmoins, cette innovation facilite certainement le repérage des lexies et apporte un avantage sur ce point à ces deux dictionnaires bien qu'il reste toutefois à évaluer par des expérimentations l'impact réel sur les apprenants de ce dispositif.

3.3 L'aide à la production

3.3.1 La recherche d'un vocable à partir d'une idée

L'un des problèmes cruciaux, lorsqu'un apprenant cherche à composer un texte, est de trouver la correspondance entre le sens, l'idée qu'il veut exprimer, et un vocable qu'il ne connaît pas. Le classement alphabétique empêche tout regroupement de vocables par thème, ce qui rend le passage du sens au texte très délicat. Pour évaluer l'utilité des monolingues anglais pour apprenant, Bogaards (1996) se propose de trouver la traduction en anglais de 27 vocables n'appartenant pas aux 2 000 vocables les plus fréquents et faisant partie de phrases exprimant des préoccupations de tous les jours. Pour cela, il essaye de trouver les chemins qui mènent aux vocables en question à partir de synonymes, d'hyperonymes ou de méronymes. A la pratique, il s'avère que le nombre de vocables dont la traduction trouvée est considérée comme satisfaisante est compris dans une fourchette allant de 9 (COBUILD) à 13 (LDOCE). Par exemple, il n'a pas été possible de trouver la traduction de délicieux, à partir de good par exemple, ou mécanicien à partir de work, job, profession ou to repair. Il est parfois possible de trouver la traduction à partir des illustrations : c'est le cas pour évier, sink, que l'on trouve dans la page consacrée à la cuisine dans LDOCE mais pas dans la figure de CIDE. Et encore faut-il savoir qu'une telle illustration existe, n'étant pas référencée dans l'entrée kitchen de LDOCE. Par contre on trouve les parties de voiture comme brakes ou exhaust pipes dans les illustrations de voitures. On remarque ici l'utilité de ces illustrations pour la production, ce qui explique en partie le faible score obtenu par COBUILD qui n'en possède aucune.

Le taux de réussite, moins de la moitié des tentatives ont été un succès, peut être considéré comme suffisamment bas pour ne pas inciter les apprenants à utiliser les monolingues. Ils se tourneront alors vers le bilingue. C'est ce qu'indique Atkins et Varantola (1997), où 81 % des utilisateurs consultent le bilingue pour une traduction en L2 contre 19 % qui utilisent le monolingue (et encore, certains ont commencé par le bilingue pour trouver les candidats, puis sont passés au monolingue pour avoir plus de précision). Tels qu'ils sont conçus actuellement, les dictionnaires monolingues pour apprenant sont donc insuffisants pour passer du sens au texte. Notons que les dictionnaires semi-bilingues se tirent plutôt bien de l'épreuve grâce à un index qui permet de passer d'un vocable en L1 en sa traduction en L2. Avec le Password de Kernerman, il a été possible de trouver 24 des 27 items du test de Bogaards, les échecs étant dus aux collocations et à un vocable, deliberation, qui ne s'y trouvait pas.

Le Longman Language Activator (LLA) est présenté par ses auteurs comme le « premier dictionnaire au monde pour la production » et diffère dans sa présentation des entrées de tous les autres dictionnaires. Il répertorie les thèmes de la langue (answer, attention, important, out, etc.) qu'il classe par ordre alphabétique puis, pour chacun d'eux, fait l'inventaire de toutes les idées qui viennent à l'esprit à partir de ces thèmes et donne les vocables, avec définitions et exemples, qui les expriment. Cette présentation semble *a priori* séduisante et utile. Pourtant, à l'usage, les résultats obtenus sont du même ordre qu'avec les autres monolingues (11 traductions satisfaisantes trouvées). Pourtant, un examen de ces traductions montre que celles qui ont été trouvées correspondent surtout à des termes abstraits, comme deliberate, compromise ou négotiate qui n'auraient pas pu être découvertes avec les autres dictionnaires, tandis que la recherche des traductions pour des vocables plus concrets mène à l'échec. Il semblerait donc que LLA soit à améliorer pour les vocables concrets, notamment par l'ajout d'illustrations qui ont prouvé leur utilité dans les autres dictionnaires.

3.3.2 Les synonymes

Même si l'apprenant a trouvé le vocable censé exprimer son idée, il peut s'interroger sur sa justesse : « Exprime-t-il vraiment ce que je veux dire ? », « N'y aurait-il pas des vocables plus précis parmi les synonymes ? », « Ne faudrait-il pas un vocable plus neutre, moins négatif, moins véhément ? » En d'autres termes, il lui faut, dans un premier temps, examiner les différentes possibilités qui se présentent pour démasquer les candidats, puis comprendre les différences, parfois subtiles, entre eux pour pouvoir retenir le plus adéquat.

Nous avons vu plus haut qu'il n'était pas toujours aisé de trouver des synonymes ou des vocables reliés par d'autres relations sémantiques. Les renvois ne sont pas systématiques, peu nombreux, la relation n'est pas toujours clairement exprimée. Les réseaux sémantiques apparaissent différents d'un dictionnaire à l'autre et ne présentent pas tous la même densité et la même qualité. Dans certains cas, voir Bogaards 1996, les définitions auxquelles on accède par les renvois ne sont pas suffisamment précises pour distinguer les nuances entre deux entrées. Grâce à ses nombreux renvois, COBUILD présente *a priori* plus de facilité pour les apprenants. Assez étonnamment, Bogaards relève des renvois qui n'aboutissent à aucun vocable. C'est le cas de natter qui indique chinwag comme synonyme, chinwag n'étant pas une entrée du dictionnaire. Il faut tout de même rappeler que la masse d'informations contenue dans un dictionnaire est telle que l'erreur n'est pas toujours évitable. Mais revenons à la recherche de synonymes. L'étude de Harvey et Yuill (1997) nous éclaire sur ce point. Il faut signaler toutefois qu'elle porte sur la première édition de 1987 du COBUILD, ce qui empêche d'établir des conclusions définitives sur l'édition de 1995, celle-ci ayant quelque peu évolué notamment dans le domaine des synonymes. D'après cette étude, le taux de réussite n'est que de 64 %, 53 % des synonymes ayant été trouvés facilement, en utilisant la colonne supplémentaire, et 11 % avec difficulté, principalement en examinant les définitions. Cependant, le fait de trouver les synonymes n'est pas le but final d'une consultation lors d'une tâche de production. Il reste à savoir si ceux-ci sont adéquats. Et là, il apparaît clairement que les apprenants ont une confiance toute relative dans les synonymes indiqués dans COBUILD 1987. En effet, seulement 34 % des synonymes trouvés dans le dictionnaire

ont été utilisés dans les compositions. Les apprenants ne considèrent pas forcément une recherche comme réussie dès lors qu'elle aboutit à un synonyme. 44 % des recherches infructueuses avaient permis de recenser des synonymes dans la colonne supplémentaire et 33 % un hyperonyme. La principale raison de cette désaffection semble être le manque d'informations suffisantes, en particulier le contexte d'emploi du synonyme. De ce fait, les apprenants sont obligés d'aller consulter l'article qui lui est consacré et rien n'indique qu'ils soient disposés à le faire. Cette méfiance est pourtant justifiée puisque les apprenants ont tendance à mal utiliser ces synonymes, c'est-à-dire qu'ils les emploient dans des contextes non appropriés d'un point de vue syntaxique, sémantique ou stylistique.

3.3.3 Les informations grammaticales

De par l'importance de la grammaire dans l'enseignement des langues ces dernières décennies, les informations grammaticales occupent un espace relativement important dans les dictionnaires pour apprenants. Les utilisateurs ont en effet besoin, plus que les natifs, d'être dûment renseignés sur ce point afin de s'exprimer correctement. Chaque dictionnaire a son système de notation, plus ou moins clair, faisant plus ou moins appel à des notions qui posent parfois des difficultés telles que la transitivité ou l'intransitivité. L'apport le plus significatif est celui de COBUILD dont les définitions sous forme de phrase font clairement apparaître les structures syntaxiques sans pour autant contraindre le lecteur à aller chercher la signification de tel ou tel code. Notons que ce procédé permet en outre, surtout en compréhension, de rester dans le même niveau de lecture : du texte, l'utilisateur passe aux phrases du dictionnaire et évite ainsi le métalangage des codes grammaticaux. COBUILD n'évite pas pour autant les codes car il n'est pas toujours possible de déduire toutes les informations des définitions ou des exemples. Par exemple, COBUILD, qui possède le système de description grammaticale de loin le plus riche, est le seul à faire la distinction entre adjectifs gradués (graded) et non gradués, c'est-à-dire pouvant être notamment accompagnés par des adverbes (assez chaud ou assez froid, mais pas assez absent). Mais pour à revenir à l'influence des codes, l'étude de Harvey et Yuill (1997) est éloquente sur ce point. Une grande majorité des apprenants localisent les informations grammaticales non pas dans les codes qui, chez COBUILD, sont disposés dans la colonne supplémentaire, mais dans les exemples et dans un moindre degré, dans les définitions. La colonne supplémentaire n'est pas en cause puisqu'elle est tout à fait utilisée pour la recherche des synonymes. Ainsi, seulement 9,8 % des recherches fructueuses ont tiré l'information grammaticale des codes de cette colonne. Il n'est pas possible de connaître l'influence des trois autres dictionnaires sur l'aide grammaticale qu'ils apportent, mais pour le COBUILD, le taux de réussite est élevé : dans 76 % des cas, l'information a été trouvée facilement et dans 14 % des cas, même si elle s'est heurtée à plus de difficultés, la recherche a été fructueuse. Les échecs sont dus à une insuffisance d'information (5,6 %), au vocable qui n'était pas dans le dictionnaire (2,8 %) et à l'entrée qui n'a pas été comprise (1,4 %). Si l'on se fie à cette étude, COBUILD présente un bilan plutôt satisfaisant.

3.3.4 Les fréquences et les registres

Signalons enfin deux derniers éléments d'informations qui peuvent guider l'apprenant à mieux choisir son vocable : la fréquence et le registre, ce dernier abordant les problèmes de pragmatique.

L'apport des corpus a contribué de manière notable à la généralisation des indications de fréquence. L'apprenant est ainsi informé de l'utilisation réelle d'une entrée et peut en mesurer, par exemple, son côté démodé ou au contraire dans le vent. Assez étonnamment, OALD et CIDE ne donnent pas cette indication tandis que COBUILD applique systématiquement une échelle de 0 à 5 diamants (pour les plus fréquents) et LDOCE, moins régulièrement, fait la distinction entre l'écrit et l'oral.

Tous les dictionnaires font la distinction entre l'emploi formel ou informel d'un vocable. COBUILD va plus loin dans ces distinctions en appliquant des registres plus spécialisés tels que journalism, legal, literary, offensive, old-fashioned, spoken, written, etc.

4 Dictionnaires pédagogiques du français

La langue française ne bénéficie pas de la situation privilégiée de la langue anglaise. En effet, tandis que pour l'anglais, l'enjeu économique favorise la compétition entre les différents éditeurs et apporte un lot d'innovations dans les *learner's dictionaries*, la situation est beaucoup plus calme pour le français. Dans son étude, Bogaards (1998) recense trois dictionnaires présentés comme étant utiles pour l'étude du français langue étrangère. Il s'agit du Micro Robert Poche (MRP), du Robert Junior Illustré (RJI) et du Dictionnaire du français langue étrangère Niveau 2 (DFLE). On notera que ces trois dictionnaires n'ont pas la même importance que leurs homologues anglais : RJI, comme son nom l'indique, s'adresse à des enfants natifs, et non à des adultes désireux d'apprendre la langue, DFLE contient un nombre d'entrées assez restreint, voire insuffisant (5 000 et 12 500 lexies d'après les estimations de Bogaards) et l'adjectif du MRP, Micro, laisse supposer qu'il s'agit d'un élément annexe de la collection Robert, impression renforcée par la petite taille et le caractère « touffu » et peu lisible de la mise en page. Signalons enfin que DFLE est très difficile à trouver en librairie et que sa dernière édition date de 1986.

Le constat de Bogaards, reprenant les mêmes éléments d'analyse que pour les dictionnaires anglais (voir ci-dessus), est sévère à l'instar de ces trois dictionnaires :

- Aucun d'entre eux n'est satisfaisant pour la compréhension (MRP obtient malgré tout les meilleurs résultats grâce à un nombre approprié d'entrées et des définitions assez précises, DFLE est faible).
- DFLE est satisfaisant pour la rédaction de texte (grâce notamment à des descriptions riches et détaillées de certaines de ses entrées), mais ce n'est pas le cas du MRP malgré un système assez riche de renvois analogiques.
- RJI est insuffisant pour la compréhension et la production.

Si l'on compare les notes obtenues par ces dictionnaires avec celles de leurs homologues anglais, on constate que ces derniers sont loin devant. Les dictionnaires français ne

bénéficient pas notamment de certaines des caractéristiques qui font la force des *learner's dictionaries*. Ils ne sont pas rédigés à partir de corpus ; il n'y a donc aucune indication de fréquence et l'intuition du linguiste, parfois trompeuse, est le seul élément d'appréciation pour la description des entrées lexicales. Il n'y a aucun contrôle sur le vocabulaire des définitions ce qui amène MRP à définir certaines entrées avec des vocables compliqués (bouche : « cavité située à la partie inférieure du visage de l'homme, bordée par les lèvres, communiquant avec l'appareil digestif et avec les voies respiratoires ») ou bien à présenter des définitions circulaires (corpulent : « qui est d'une forte corpulence »). Enfin, il n'y a pas de *signposts* ou d'indices permettant d'aiguiller l'apprenant sur les entrées longues.

Bogaards termine son étude en citant Zöfgen (1994) : « il y a certains indices qui révèlent que la lexicographie du français langue étrangère, face à celle de l'anglais, risque de se retrouver dans l'arrière-garde et que la France, qui peut se prévaloir d'une tradition glorieuse – et ininterrompue depuis le XVIIe siècle – dans le domaine de la lexicographie monolingue, ne peut plus prétendre à une position de premier plan qu'en ce qui concerne les dictionnaires monolingues en plusieurs volumes ». Il s'étonne « devant la qualité très pauvre des dictionnaires qui sont proposés aux très nombreux apprenants du français langue étrangère. », surtout par rapport aux « multiples discours français concernant la mission universelle de la langue française ». Cette situation était déjà relevée par Rey (1989) qui expliquait cette insuffisance par les différences entre les marchés des langues et préconisait une aide à l'édition. Signalons que le ministère des affaires étrangères français recense 57 millions d'élèves et d'étudiants du français dans le monde, le français se situant globalement au deuxième rang des langues vivantes enseignées. Même si l'anglais est loin devant, ce nombre n'est tout de même pas négligeable.

5 Les dictionnaires électroniques

Les progrès technologiques de l'informatique ont permis l'accroissement de la rapidité et des volumes de données traités. C'est ainsi que depuis le début des années 90, le grand public a pu avoir accès à de nombreux dictionnaires électroniques. Il convient donc d'étudier, dans le cadre de ce présent travail, ce que ces dictionnaires peuvent apporter dans l'étude du lexique d'une langue étrangère, c'est-à-dire en quoi ils peuvent faciliter l'accès et l'usage des informations qu'ils contiennent. Nous avons principalement étudié les dictionnaires électroniques de français dont nous disposions, c'est-à-dire, pour les monolingues, la première version du Robert Électronique (RE, 1994, l'équivalent du Grand Robert), le Petit Robert sur CD-Rom (PRCD, 1996) et le Petit Larousse sur CD-Rom (PLCD, 1998), et pour les bilingues, les versions électroniques du Grand Larousse bilingue français-anglais (GLCD, 1996) et du Hachette-Oxford (HOCD, 1996), lui aussi bilingue français-anglais. Par ailleurs, nous avons pu obtenir des informations sur les dictionnaires électroniques anglais pour apprenant d'après la revue de Nesi (1996), traitant du Longman Interactive English Dictionary (LIED, 1993), du Electronic Oxford Wordpower Dictionary (OWPD, 1994) et du Collins Cobuild on CD-Rom (COBUILDCD, 1995), et les sites Internet des éditeurs Longman (1999) et Collins (1999). Signalons que parmi tous ces dictionnaires électroniques, seul COBUILDCD est en fait un véritable dictionnaire pour apprenant, même si cet éditeur a produit par ailleurs le Collins

COBUILD Learner's Dictionary, qui s'adresse à un public de niveau intermédiaire. Depuis la revue de Nesi, Cobuild a produit d'autres dictionnaires électroniques, parmi lesquels COBUILD E-Dict Dictionay on CD-Rom (E-Dict, 1998), dont nous allons examiner quelques fonctionnalités intéressantes, et le COBUILD English Collocations on CD-ROM (CECC) qui, comme son nom l'indique, traite des collocations. Cobuild semble donc, pour l'instant, l'éditeur de dictionnaire électronique le plus prolifique et le plus innovateur.

Pour Nesi, les dictionnaires électroniques peuvent exceller dans quatre domaines : les renvois dans et entre différentes sources qui constituent le CD-Rom, les liens directs avec les autres applications, notamment les traitements de texte, les possibilités de recherche complexe à l'intérieur du texte des entrées du dictionnaire et les interactions possible avec les utilisateurs pour aider à développer le vocabulaire et les facilités de consultation. Sur ces quatre points, un seul, la recherche sur les entrées, concerne le dictionnaire lui-même. Les renvois entre différentes sources (surtout pour les dictionnaires anglais ; LIED, par exemple, propose une compilation de quatre sources de référence : le *Longman Dictionary of Common Errors*, le *Longman English Grammar*, le *Longman Pronunciation Dictionary* et le *Longman Dictionary of Language and Culture*, plus des vidéos et du son) concernent les facilités de navigation entre les différents volumes électroniques, les liens avec les autres applications permet d'exporter les résultats dans des traitements de textes et les interactions concernent surtout les sources audio et vidéo extérieures ainsi que divers activités ludiques sur les vocables. Même si ces fonctionnalités apportent une richesse incontestable au dictionnaire de base, notre propos portera plutôt sur le dernier point.

L'étude part d'une remarque générale qui concerne tous les dictionnaires cités plus haut : tous sont des adaptations électroniques de version papier. Ceci à des conséquences à la fois sur la macrostructure et sur la microstructure du dictionnaire. D'autre part, on trouve les mêmes informations lexicales que celles contenues dans les versions papier. Dès lors, nous n'examinerons pas les informations elles-mêmes, mais plutôt la manière dont on y accède et dont elles sont présentées.

5.1 L'accès lexical

Les possibilités de traitement automatique et de recherche d'un vocable dans un index sont un des grands atouts des dictionnaires électroniques par rapport à leur équivalent papier. En effet, une entrée peut être sélectionnée soit en cliquant dessus, soit en tapant ses premières lettres. Ce procédé efficace permet d'accéder au vocable très rapidement (Dokter *et al*., 1997), ce qui est très apprécié par les apprenants (Guillot & Kenning, 1994a & b). Ce gain de temps s'avère très important dans un processus qui doit être le plus court possible. De plus, l'apprenant sera davantage tenté d'utiliser le dictionnaire pour rechercher une information.

Passer de la forme fléchie dans le texte à la forme canonique dans un dictionnaire n'est pas toujours évident dans une langue morphologique comme le français qui contient beaucoup de formes irrégulières (irai pour aller, yeux pour œil, etc.). Même si certaines formes irrégulières sont citées à leur place dans l'ordre alphabétique (par exemple les pluriels très irréguliers), ce n'est

pas toujours le cas (la plupart du temps les verbes conjugués ne sont pas dans un dictionnaire). Dès lors, l'apprenant ne peut compter que sur sa connaissance de la langue.

La plupart des dictionnaires électroniques récents résolvent simplement ce problème, soit en listant l'ensemble des formes fléchies, soit en faisant fonctionner un analyseur morphologique.

Dans un premier temps, et pour la plupart des recherches, l'accès aux vocables simples peut donc être considéré maintenant comme étant résolu de manière satisfaisante par les dictionnaires électroniques. Cependant, deux problèmes subsistent. D'une part, il y a le problème de l'homonymie : que faire lorsque deux vocables s'écrivent de la même manière ? Lorsque les deux vocables (car il s'agit bien de deux mots différents et non pas d'une polysémie plus large) n'ont pas la même catégorie grammaticale (boucher verbe et boucher nom), le problème est résolu par l'ajout, comme dans le cas du PRCD, de la catégorie grammaticale à côté du vocable dans la liste des entrées. Par contre, si ce n'est pas le cas (voler, dérober et voler, être dans les airs), aucun moyen n'est possible de les distinguer (PRCD liste trois poste). Il faut donc lire l'article de chaque entrée pour rester ensuite sur celui qui nous intéresse. A l'évidence, il manque une information.

Un deuxième problème, plus important, concerne les collocations, idiomes, expressions semi-figées, c'est-à-dire tous les vocables composés de plusieurs mots. Là, les solutions proposées sont très différentes. On trouve d'une part les dictionnaires qui ne tiennent pas compte de ce phénomène (PRCD). HOCD propose un index contenant principalement des expressions semi-figées mais qui sont listées telles quelles (annoncer qch de but en blanc à qqn) et même avec leur variation (avoir la tête or gueule de l'emploi[3]). Il faut taper exactement les premières lettres de l'expression pour avoir accès à sa traduction (en fait, à l'article de l'entrée qui contient sa traduction). Inutile de dire que cette fonctionnalité ne sera que de peu de secours pour celui qui précisément ne connaît pas l'expression ou qui ne la devine pas dans un texte. Plus souple est le moyen employé par GLCD. On obtient en tapant un mot, une liste d'unités polylexicales qui contiennent ce mot. Malheureusement, d'une part, les expressions listées sont le plus souvent des locutions adverbiales (en sorte que) même si, il faut quand même le signaler, on trouve parfois de véritables expressions semi-figées telles que au coup par coup, en tête à tête ou tête de mort. D'autre part, et surtout, cet accès n'est pas systématique (on ne trouve pas coup de barre, ni à coup, ni à barre, on ne trouve pas travail au noir ni à travail, ni à noir) et on ne peut pas accéder à ces expressions à partir de n'importe lequel de ses constituants (on accède à tête de mort à partir de tête mais pas à partir de mort). L'initiative la plus avancée en ce domaine est sans doute celle de E-Dict, si l'on s'en tient aux informations délivrées sur le site Internet, où l'utilisateur a véritablement accès aux idiomes et expressions à partir des constituants. Ainsi on a accès à foul play à partir de foul (il faudrait vérifier si on peut y accéder à partir de play). Signalons que dans ce dictionnaire, certaines expressions ont leur entrée à part entière. Cette fonctionnalité semble particulièrement intéressante pour avoir accès rapidement à ces expressions dont la localisation pose souvent problème dans les versions papier, mais aussi aide au repérage des expressions dans certaines tâches de compréhension comme la lecture

[3] Notez la présence de or.

d'un texte : certaines expressions n'étant pas toujours faciles à repérer, l'apprenant peut facilement et rapidement vérifier si le mot qu'il ne comprend pas est en fait un constituant d'une expression et avoir accès dès lors à sa définition.

CECC traite des collocations. A partir d'un mot nœud, il est possible d'avoir la liste, par ordre de fréquence, des mots qui apparaissent le plus souvent proches du premier, les mots grammaticaux exclus. Les statistiques sont calculées à partir d'un corpus de 200 millions de mots (The Bank of English). Il est possible, par ailleurs, d'avoir les concordances elles-mêmes. Ce programme sert principalement à attester l'existence d'une collocation et de ses variations. Il s'agit davantage d'un concordanceur que d'un dictionnaire électronique.

Certains dictionnaires, notamment PRCD, proposent des possibilités de recherche dans le texte des entrées du dictionnaire. Il est ainsi possible de recenser tous les articles contenant un ou plusieurs vocables donnés. La recherche peut être précisée par l'emploi de connecteurs logiques tels que ET, OU ou PRES/n (mot1 et mot2 à n mot d'écart) : la requête « travail ET noir » donne tous les articles contenant obligatoirement ces deux vocables, mais sans qu'ils aient forcément un rapport l'un avec l'autre. L'emploi de PRES (qui a pour paramètre le nombre de mots d'écart), permet de repérer les collocations et expressions semi-figées. Ainsi la requête « travail PRES/2 noir » donne tous les articles qui contiennent l'expression « travail au noir ». Cette fonctionnalité, bien qu'incontestablement utile pour repérer les unités polylexicales, a pourtant quelques limites, car il faut parfois consulter plusieurs articles pour trouver l'information recherchée (notamment une explication). De plus, l'accès n'étant pas immédiat (il faut passer par un menu spécial, taper deux mots en entier et un connecteur puis parcourir les articles), il est peu probable qu'un apprenant utilise cette fonctionnalité lors de la lecture d'un texte, surtout s'il n'est pas certain de l'expression elle-même. Utile pour l'expert-linguiste, cette fonctionnalité semble plutôt une astuce informatique pour l'utilisateur normal, destinée à pallier l'insuffisance de traitement et de démarcation des unités polylexicales, à la fois dans la macrostructure (pas d'article propre) et dans la microstructure (noyées dans l'entrée et sans explication ou exemple systématique[4]) du dictionnaire.

Concernant la production, les problèmes d'accès restent les mêmes que ceux rencontrés avec les versions papier et il n'est pas plus facile de passer de l'idée au vocable, quelque soit le support. Pour l'instant, l'apport en production des dictionnaires existants se situe à un niveau plus limité, mais traité toutefois de manière efficace et satisfaisante : l'orthographe. En effet, dans la même fonction recherche (vue plus haut) du PRCD, il est possible de trouver la bonne orthographe d'un vocable à partir d'une écriture purement phonétique. Ainsi, plus de problème pour trouver les orthographes de labyrinthe ou ornithorynque à partir de, respectivement, labirinte ou ornitorinke (utile pour la dictée de Pivot !). Mais rien n'est prévu dans les domaines sémantiques : pas de réseaux de synonymes, pas d'application de fonctions lexicales (comment dit-on travailler beaucoup), etc.

[4] On trouvera une explication (indirecte) de formation continue dans le renvoi à recyclage ; climat social est « expliqué » (cité) par rapport à social, sans toutefois plus de précisions : si le natif comprendra sans doute, ce ne sera peut-être pas le cas pour l'apprenant.

5.2 Compréhensibilité

L'un des avantages des dictionnaires électroniques par rapport à leur version papier est l'interactivité. L'utilisateur peut influer sur la nature et la quantité des informations qui lui sont présentées et s'affranchit des limites du papier. Malheureusement, il faut constater que tous les dictionnaires reproduisent à l'écran l'agencement de la version papier, du plus clair comme le COBUILD au plus enchevêtré comme PR où définitions, explications, exemples, citations, renvois sont entremêlés dans le corps du texte de l'entrée. C'est d'autant plus dommage que le support informatique n'est plus tributaire du manque de place des versions papier (ce qui explique la densité de certains dictionnaires).

Pour autant, certains programmes permettent d'agir sur la présentation en laissant la possibilité à l'utilisateur d'afficher tel ou tel élément d'information. Ainsi, dans PR, on peut avoir seulement les définitions, ou les exemples, ou les renvois analogiques, etc. ou bien des combinaisons de plusieurs d'entre eux, comme les définitions et les renvois (mais pas sur la version Macintosh). Cette fonctionnalité facilite le parcours de l'entrée et limite les inconvénients de l'agencement de la version papier. Il est dommage qu'il n'y ait pas d'indépendance dans les lexies : si l'on étudie les lexies en n'affichant que les définitions mais que l'on veut avoir plus de détails sur une lexie particulière, il n'est pas possible d'obtenir plus d'informations sur elle seulement sans faire afficher les mêmes informations pour toutes les autres lexies. La programmation de cette fonctionnalité apparaît donc comme assez sommaire. Le même problème était constaté sur la première version Macintosh du RE : il était possible de visualiser soit des définitions abrégées, soit complètes, ce qui signifiait qu'après avoir repéré la lexie recherchée, il fallait retenir son numéro de sens, afficher les définitions complètes et faire défiler le contenu de la fenêtre (assez petite d'ailleurs et non redimensionnable !) jusqu'au bon endroit. Assez bizarrement, la version PC (en MS-DOS) faisait apparaître une fenêtre supplémentaire contenant les détails de la lexie cliquée, indépendamment des autres.

Mises à part les définitions abrégées du RE, aucun dictionnaire n'a mis en place un dispositif visant à faciliter le repérage des différentes lexies. LDOCE et CIDE sont les seuls dictionnaires à utiliser des *signposts* pour catégoriser sémantiquement les différentes lexies. Mais leur version électronique n'est pas encore distribuée (ou conçue).

6 Conclusion

Nous avons étudié dans ce chapitre le rôle que jouaient les dictionnaires dans l'apprentissage lexical. Après avoir abordé les rapports qu'entretiennent les apprenants avec le dictionnaire (possession, utilisation, préférence pour le bilingue ou le monolingue suivant le niveau de langue, etc.), nous avons vu les problèmes d'accès lexical et le fait que les dictionnaires ne contribuent pas forcément de manière satisfaisante à la compréhension de textes écrits.

Nous avons ensuite passé en revue les différentes propriétés des dictionnaires monolingues pour apprenants, dont la principale différence par rapport aux dictionnaires pour natifs est

l’utilisation d’un vocabulaire définitoire contrôlé ou simplifié. Nous avons pour cela examiné les dictionnaires d’apprentissage anglais, ceux-ci étant les plus innovateurs (utilisation de corpus écrits et oraux, définitions sous forme de phrase, *signposts*, informations de fréquence, etc.) et les plus performants en la matière. En effet, nous n’avons pu que constater le grand retard de la lexicographie pédagogique française.

Pour finir, nous avons étudié les dictionnaires électroniques qui sont principalement pour l’instant des retranscriptions de leur équivalent papier, en dépit de la nature différente du support électronique et des possibilités de traitements automatiques.

Avant d’aborder aux chapitres 5 et 6 les solutions envisagées dans ALEXIA pour tenter de remédier à cet état de fait, nous allons examiner les activités lexicales que nous comptons incorporer dans le système et dont va dépendre une partie de la conception de la base de données lexicales.

CHAPITRE 4

Les activités lexicales

Nous avons vu dans le chapitre précédent les problèmes inhérents aux dictionnaires d'apprentissage qui concernent principalement l'accès aux informations lexicales et l'aide à leur compréhension. Néanmoins, avant de passer à la modélisation et à la conception de notre propre base lexicale, il convient d'examiner un des modules qui utilisera ces connaissances : les activités lexicales. Il importe en effet de savoir quel type de connaissances elles nécessitent et la manière dont celles-ci doivent être organisées pour pouvoir être générées automatiquement par le système.

Dans ce chapitre, nous allons étudier les exercices existants sur le lexique en vue de l'intégration d'un module d'activités visant à entraîner l'apprenant sur certains vocables et à favoriser ainsi leur acquisition lexicale.

1 Intérêt des activités lexicales

Pour Paribakht et Wesche (1997), la présence d'activités lexicales dans l'apprentissage du vocabulaire se justifie par de meilleurs résultats des apprenants à des tests d'évaluation de compétences lexicales par rapport à d'autres processus d'apprentissage telle l'exposition à de nouveaux vocables par la lecture seule de textes. Les activités lexicales accéléreraient donc l'acquisition lexicale. Leurs travaux reprennent l'idée, déjà suggérée par Hulstijn (1989) et reprise par Goodfellow (1994), selon laquelle la rétention est favorisée par la quantité de travail effectué sur un vocable. La lecture de texte et l'exposition seule ne suffisent pas pour retenir les vocables. Paribakht et Wesche ont donc mesuré les incidences pour l'apprentissage, d'une part, du processus de lecture seule (LS), et d'autre part, du processus de lecture suivi d'activités lexicales (LA). L'expérimentation a porté sur 38 jeunes adultes apprenants d'anglais langue étrangère de niveau intermédiaire et ayant des langues maternelles différentes (français, arabe, chinois, etc.). Les mêmes sujets ayant subi les deux traitements, les textes sélectionnés dans les deux cas appartenaient à des thèmes différents. Les activités lexicales définies dans LA étaient nombreuses et variées et faisaient intervenir plusieurs compétences lexicales. Elles étaient regroupées en cinq catégories :

- attention sélective : activités visant à s'assurer que les étudiants repéraient certains vocables-cible (extraits par exemple d'une liste)
- reconnaissance : activités visant à s'assurer que les étudiants reconnaissaient les vocables-cible et leur sens (connaissance partielle des vocables-cible)
- manipulation : activités impliquant des connaissances morphologiques sur les vocables (dérivés syntaxiques, construction de vocables à partir d'affixes et de racines)
- interprétation : activités impliquant l'analyse du sens des vocables vis-à-vis du contexte (collocations, synonymes, etc.)
- production : activités demandant aux étudiants de produire des phrases contenant les vocables-cible dans des contextes appropriés.

Le temps passé dans les activités lexicales était compensé dans l'autre groupe par un supplément de lecture.

Les tests ont montré que les résultats étaient meilleurs lorsque le groupe pratiquait les activités lexicales. Ils ont permis de vérifier les hypothèses suivantes :

- Les étudiants possèdent une meilleure connaissance des vocables-cible après la séance de lecture suivie d'exercices mais aussi après la séance de lecture seule (ceci pour vérifier que la lecture était utile à l'apprentissage)
- Pour un temps donné et égal dans les deux cas, les gains en apprentissage étaient plus grands pour la lecture suivie d'exercices que pour la lecture seule.
- Les gains en vocabulaire étaient à la fois quantitatifs (plus de vocables connus à la fin) et qualitatifs (meilleure connaissance des vocables, mesurée par l'application d'une échelle de connaissance spécifique)
- Les gains dans le cas de la lecture avec exercices concernaient davantage les mots pleins (verbes, noms) que les mots grammaticaux.

L'expérimentation a aussi montré que les étudiants avaient une opinion favorable des activités, pensant que celles-ci amélioraient leurs compétences lexicales.

2 Les activités lexicales

Dans le chapitre 2, nous avons vu que le système Lexica développé par Goodfellow (1994) proposait des idées intéressantes pour l'apprentissage lexical. Cet environnement proposait à l'apprenant une exploration autonome de vocabulaire et une exposition aux vocables par la lecture de textes déjà dans le système, tout en lui mettant à disposition des ressources lexicales pour l'aide à la compréhension de vocables nouveaux (dictionnaires, concordanceur). L'utilisateur était invité à constituer un dictionnaire électronique personnalisé en regroupant et en annotant certains vocables choisis dans les textes. Ce travail de catégorisation permet de fortifier certains liens faibles du lexique mental venant d'être créés et aide ainsi à la rétention des nouveaux vocables.

L'environnement était complété par un test de « rappel », au cours de laquelle, le contexte du vocable étant donné, l'apprenant devait retrouver l'item manquant. On mesurait ainsi le taux d'apprentissage.

Comme nous allons le voir dans la suite de ce travail, l'environnement ALEXIA reprend certaines des idées développées par Goodfellow tout en apportant une contribution sur certains des constituants. Nous cherchons donc à organiser des activités lexicales plus riches s'appuyant sur les ressources textuelles et lexicales dont nous disposons : le corpus de textes, le dictionnaire général et le dictionnaire personnel.

Il s'agit donc dans ce chapitre de spécifier les activités lexicales qui feront partie d'ALEXIA, orientant de ce fait la constitution et l'organisation des ressources textuelles et lexicales.

Ces activités seront informatisées et leur intérêt computationnel doit prévaloir : en effet, ces activités ne doivent pas être reproductibles sur papier, auquel cas leur informatisation demeurerait simplement anecdotique et on se priverait des moyens que l'informatique propose.

Ces activités doivent donc présenter quatre caractéristiques :

- Elles doivent être pertinentes en termes d'apprentissage, c'est-à-dire que l'apprenant doit fortifier ou valider son acquisition grâce à elles.
- Elles doivent être reproductibles, c'est-à-dire que l'apprenant pourra avoir de nouvelles épreuves à volonté.
- Elles doivent se prêter à la mise en place d'un système d'aide qui confortera le rôle pédagogique de l'environnement, c'est-à-dire que le système ne se contentera pas de répondre vrai ou faux, mais aiguillera l'apprenant pour trouver la solution.
- Elles doivent être réalisables en termes informatiques, c'est-à-dire que leur mise au point ne doit pas révéler une explosion combinatoire au-delà d'un certain nombre d'items considérés, et le temps de réponse du système, tant au niveau de la préparation des épreuves que dans l'aide qu'il apporte, doit être convenable et acceptable pour que l'activité puisse se dérouler normalement.

Les système à base de ressources lexicales comprennent généralement un corpus de textes et un dictionnaire. Leur rôle principal est de pouvoir fournir, d'une part, une exposition authentique aux vocables, et d'autre part, une aide à l'utilisateur pour comprendre les vocables qu'il ne connaissait pas ou mal. Cependant, il serait dommage de considérer ces ressources uniquement sur un plan consultatif par l'utilisateur humain et leurs données doivent de ce fait pouvoir être utilisables par le système. Il importe donc dans ALEXIA que les textes soient disponibles pour le système qui doit pouvoir en extraire un mot, une phrase, un paragraphe ou un texte donné suivant des critères paramétrables déterminés par l'utilisation. De même, le dictionnaire électronique constitue une base de données lexicales dans laquelle les vocables sont reliés par des relations linguistiques standard. L'utilisateur manipule simplement une interface qui extrait les informations utiles dans la base lexicale. De ce fait, il faut donc tenir compte, pour l'élaboration des activités, des textes et informations lexicales disponibles. Une troisième source de données est également disponible : le dictionnaire personnel. En effet, la base que l'étudiant construit est une mine d'informations, non seulement sur les vocables eux-mêmes, mais sur la façon dont ils sont perçus par

l'apprenant. Il est donc tout à fait pertinent de tenir compte de ces informations pour la spécification des activités lexicales.

Enfin, en plus de ces données, le système dispose d'un outil, un analyseur morphologique, développé par le LIPN de l'Université Paris Nord.

Nous allons donc examiner les activités lexicales qui ont déjà été élaborées dans la littérature pour retenir, en fonction des caractéristiques énoncées ci-dessus, des données et des outils dont dispose le système, celles qui nous semblent intéressantes et que nous retiendrons.

Il existe de nombreux exercices qui mettent en jeu différentes approches ou différents aspects des vocables et de leur utilisation. Il y a par exemple les activités en compréhension ou en production, les activités communicatives ou non communicatives, les exercices hors contexte ou en contexte. Le niveau doit aussi être pris en compte. Il faut savoir cependant sur quels aspects du vocable on compte travailler : un vocable comporte plusieurs facettes (son sens, sa forme, son utilisation, etc.). Veut-on travailler tous les aspects en même temps ou au contraire seulement quelques-uns ?

L'informatisation de l'activité est une forte contrainte. En effet, le système doit fournir un retour, une correction, une aide, sur les réponses de l'apprenant. Au vu de la technologie actuelle et des moyens dont nous disposons, il n'est pas possible que l'ordinateur intervienne sur une production libre de l'apprenant en vue d'une analyse et d'une correction automatique. La correction sémantique et syntaxique est un vaste sujet, qui aborde le problème de la représentation des connaissances, implique une meilleure compréhension des langues en général et nécessite des travaux à long terme.

Cependant, le problème n'est pas vraiment de l'ordre de la production libre ou non. La langue est un phénomène riche et complexe et tous ses aspects ne peuvent être abordés en même temps. Il convient donc d'en isoler certains pour pouvoir les travailler plus efficacement.

Nous nous tournerons donc plutôt vers des exercices à réponse fermée qui s'orienteront davantage vers la compréhension (contexte -> vocable) que vers la production (vocable -> contexte), ce qui n'empêche pas tout de même un grand éventail de possibilités comme nous allons le voir. De plus certains exercices sont suffisamment souples dans leurs variantes pour pouvoir aborder les deux côtés.

3 Les activités hors contexte

On définit par activité hors-contexte (ou décontextualisée) les activités faisant intervenir les mots seuls, isolés de la phrase et du texte. Dès lors, le sens du mot n'est pas déterminé s'il est polysémique et l'on travaille au niveau du vocable, voire de la graphie (si le mot présente des homonymes).

Dans ce cas, les seuls indices pouvant déterminer le sens résident dans le vocable lui-même (racines, affixes, ressemblances avec un vocable de la langue maternelle). Il est possible néanmoins d'explorer de nombreuses facettes du vocable dans différents types d'exercices.

Bogaards (1994), Tréville et Duquette (1996) ou Verlinde (1993) en propose un certain nombre dont nous reprenons les plus caractéristiques ci-dessous.

Il y a d'une part les tests de reconnaissance de forme (par ex. Meara et Jones, 1988) : telle ou telle suite de lettres forme-t-elle un vocable de la langue ou pas ? Ces tests s'appliquent au niveau élémentaire. On donne une liste de vocables et on invite l'apprenant à indiquer ceux dont il connaît le sens (sans vérification pour autant ; il s'agit donc d'auto-évaluation). Dans la liste sont inclus des pseudo-vocables visant à limiter la surestimation de l'élève ou les réponses aléatoires. L'auto-évaluation par une liste de contrôle est un des moyens le plus direct de découvrir ce que l'on sait et ce que l'on ne sait pas et permet donc à l'apprenant d'évaluer ses propres connaissances lexicales en compréhension.

Les épreuves qui reviennent le plus souvent sont les tests à choix multiples (QCM) et une variante, les tests d'appariement. Le QCM permet d'explorer les côtés sémantiques (on demande une traduction en langue maternelle, un synonyme, un antonyme, un hyponyme, une définition), morpho-syntaxique (quel adjectif ou quel nom correspond à tel ou tel vocable), collocationnel, pragmatique, etc. A chaque question correspond un certain nombre de réponses (au moins quatre, de préférence), une étant la réponse correcte, les autres étant des distracteurs. L'informatisation de cette épreuve ne pose pas de problème, de même que la correction. Par contre, la conception est bien plus délicate : Tréville et Duquette (1996) mettent en évidence de nombreux facteurs et contraintes pesant sur les distracteurs par rapport au stimulus et sur l'établissement des tests. Les distracteurs doivent être par exemple de même catégorie grammaticale, de même niveau de difficulté que le stimulus, liés à un même domaine de référence ; en outre, il ressort d'après les expérimentations qu'ils doivent être de même longueur (les candidats ont tendance à choisir le distracteur le plus long), que la bonne réponse ne doit pas être disposée suivant un système observable (comme toujours en dernière position), que deux mauvaises réponses ne doivent pas être synonymes ou antonymes (automatiquement éliminées) et qu'elles doivent également présenter un certain attrait. Il devient dès lors difficile de pouvoir générer automatiquement ces tests et notamment les distracteurs.

Les tests d'appariement sont une variante des QCM car ils résultent, pour les mêmes questions, de la mise en commun des distracteurs. A une série de questions correspondent un même nombre (ou un nombre supérieur) de réponses et il faut donc faire le lien entre une question et une réponse. D'un point de vue informatique, les tests d'appariement présentent les mêmes caractéristiques que les QCM et il est difficile d'envisager de pouvoir les générer automatiquement.

Ces exercices s'avèrent rentables dans des approches pré-communicatives en compréhension (en préparation à une situation communicative telle que la lecture ou l'écoute d'un texte, pour travailler les vocables-clés) ou post-communicative (après la situation communicative, pour approfondir certains vocables auxquels l'apprenant aura été sensibilisé).

Il faut cependant garder à l'esprit que les exercices hors-contexte ne déterminant pas le sens dans le cas de vocables polysémique, il est délicat de présenter des activités dans lesquelles les réponses impliquent des relations linguistiques reliant non pas des vocables mais des lexies (synonymie, antonymie, dérivation syntaxique, etc.) qui sont déterminées justement par le contexte de la phrase. Ces activités doivent donc être appliquées avec précaution et plutôt sur des vocables monosémiques, pour éviter toute confusion de sens.

Les activités hors-contexte sont avant tout des activités papier. Il en existe toutefois des versions informatisées dont on peut avoir l'illustration sur le site Internet de La Passerelle (1999) (figures 4.1 et 4.2) :

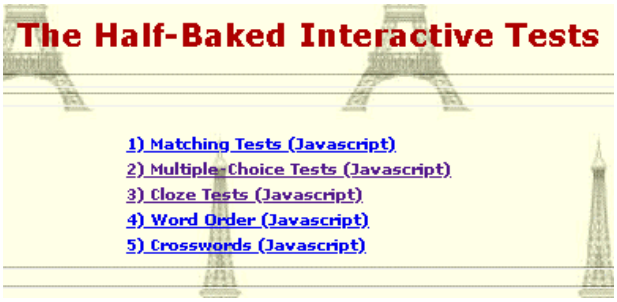


Figure 4.1 : un choix d'exercices en ligne sur le site de la Passerelle

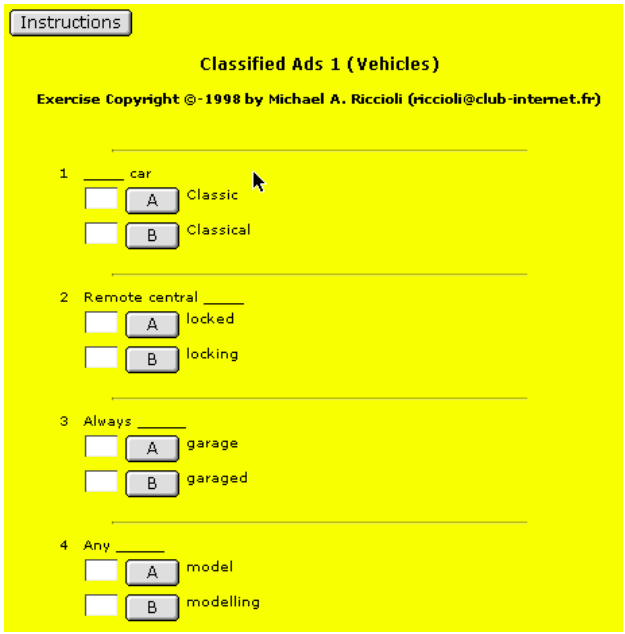


Figure 4.2 : QCM sur la vente de voiture

Cependant, il n'y a nulle génération automatique et ces exercices sont conçus et écrits à la main. Le corpus est donc « consommé » dès lors que le dernier exercice est effectué.

En raison de ces difficultés et de l'ambiguïté sémantique qui pèse sur les vocables hors-contexte, nous nous tournerons vers les activités en contexte.

4 Activités en contexte

Les activités lexicales en contexte s'appuient sur un corpus de textes duquel il est possible d'extraire des sous-parties (texte, paragraphe, phrase) mettant en évidence tel ou tel vocable.

Replacé dans son contexte, le sens du vocable est donc déterminé par ceux qui l'environnent, par la phrase, ou, plus occasionnellement, par le paragraphe. Ces activités ne portent pas seulement sur la connaissance du sens du vocable mais montrent aussi comment celui-ci s'emploie et fonctionne dans la chaîne discursive. Les différentes propriétés du vocable sont ainsi illustrées de manière plus naturelle et plus complète que dans le cas des activités hors-contexte. En passant du vocable à ses lexies, il est possible de concevoir des activités mettant en œuvre les différentes relations linguistiques vues plus haut, sans oublier les propriétés morpho-syntaxiques, collocationnelles ou pragmatiques.

Les activités en contexte permettent de plus à l'apprenant de développer ses capacités d'inférence. En effet, l'information manquante peut, dans certains cas, être déduite des éléments du contexte. Nous avons vu dans le premier chapitre combien l'inférence était un procédé utile à l'apprentissage en favorisant la connaissance de nouveaux vocables par une exploitation des indices contextuels.

Explorons donc ce deuxième type d'activités.

L'exercice le plus classique, très répandu, est le test de closure. Il consiste à « supprimer des vocables d'un texte et à inviter les sujets à restituer les vocables manquants » (Mothe, 1975). Il existe plusieurs critères pour supprimer les vocables que l'expérimentateur veut faire deviner. Ceux-ci peuvent être tout simplement retirés à intervalle fixe (tous les 5 à 7 vocables, cas du test dit « classique » ou « aléatoire », Hughes, 1989 ; Tréville et Duquette, 1996). Comme le retrait peut concerner n'importe quelle catégorie grammaticale (des verbes, des noms, mais aussi des articles, des pronoms, etc.), une variante consiste, par un étiquetage syntaxique préalable qui détermine la catégorie grammaticale de chaque vocable, à éviter les catégories les moins intéressantes comme les articles par exemple (il est facile de reconnaître un article et le nombre de valeurs possibles est faible). Un autre critère concerne la fréquence des vocables à retirer, partant du principe qu'il y a une corrélation entre les compétences langagières d'un apprenant et la taille de son vocabulaire. Les vocables à retrouver sont donc ceux qui sont les moins fréquents. Coniam (1996 et 1997), dans son évaluation de ces trois variantes, conclue que le critère de fréquence conduit à des meilleurs tests.

Quant aux vocables à trouver, on peut soit les donner dans le désordre, soit, plus difficile, ne pas les donner, soit préparer une version à choix multiple (pour chaque vocable manquant, le sujet doit choisir parmi quatre réponses).

Les tests de closure présentent l'avantage d'être faciles à générer automatiquement (sauf le cas de la version QCM pour les mêmes raisons, problème des distracteurs, évoquées plus haut) à partir de n'importe quel texte. En termes de performances, l'évaluation de Coniam indique que les meilleurs tests (ceux produits à partir du critère de fréquence) mènent à la génération d'un tiers d'items non acceptables (c'est-à-dire trop faciles ou trop difficiles à trouver ou bien pas assez discriminateurs). Ce taux chute à moins de la moitié d'items acceptables dans le cas des tests de closure à intervalle fixe. Il semble donc malgré tout qu'une correction humaine soit nécessaire, même si la machine a fait le plus gros du travail.

Ces tests ont néanmoins montré leur efficacité et sont assez répandus. Toutefois nous les écarterons pour deux raisons. Premièrement, il n'y a pas forcément de relations entre les vocables à deviner. Même si ce n'est pas nécessaire, il semble plus intéressant de faire

travailler l'apprenant sur un champ sémantique déterminé en l'invitant à retrouver par exemple un ensemble de synonymes ou d'antonymes donné, ou bien des vocables appartenant à la même famille lexicale (dérivés syntaxiques). Tout en exerçant les capacités d'inférence propres à tout test de closure, l'apprenant travaille et renforce les liens entre les réseaux de son lexique mental.

Deuxièmement, il n'y a pas de lien entre les vocables à deviner et ceux faisant partie de son dictionnaire personnel. Autrement dit, il ne connaîtra pas forcément les vocables qu'il doit retrouver, et aura d'autant plus de mal à réussir son épreuve. Il ne semble en effet pas pertinent de déclencher des activités lexicales sur des vocables inconnus. Le phénomène peut être limité en générant les tests à partir des textes lus, mais d'une part cela restreint le choix des textes et d'autre part la réussite peut s'expliquer par un simple effet de remémorisation, sans entraîner de réflexion sur le vocable lui-même.

L'une des activités lexicales que nous préconisons consiste en un exercice de recontextualisation de vocables (figure 4.3) auxquels l'apprenant a déjà été exposé (Tréville et Duquette, 1996). Il s'agit de redonner un contexte, le plus souvent une phrase, à des vocables déterminés, de préférence dans le dictionnaire personnel. Comme le disent ces auteurs (pp. 116-117), « L'association de ces vocables avec les contextes appropriés met en œuvre des stratégies d'inférence qui s'exercent à partir des vocables eux-mêmes. Ceux-ci, étant partiellement connus des apprenants, permettent d'établir des liens (sémantiques ou morpho-syntaxiques) avec certains éléments de leur entourage et de tenter une interprétation du sens des phrases ».

A partir des données du dictionnaire personnel, le système définit une quantité de vocables parmi ceux suffisamment travaillés mais non encore maîtrisés par l'apprenant (par exemple entre 5 et 10), recherche dans le corpus de textes les phrases qui contiennent ces vocables et les présente à l'apprenant tout en retirant les vocables en question (figure 4.3). Tout comme dans le test de closure, l'apprenant doit deviner les vocables manquants.

L'exercice peut être orienté soit en compréhension, si l'on donne dans l'énoncé les vocables à placer dans les trous, soit en production, en laissant l'apprenant les deviner. Cependant, si aucune condition ne pèse sur les vocables, les phrases étant disjointes, cet exercice peut s'avérer dans la plupart des cas, très difficile à résoudre, et donc rebutant. Pourtant, nous allons voir qu'en liant les vocables à trouver par certaines relations linguistiques, l'exercice devient abordable et pertinent d'un point de vue productif.

Utilisez les mots suivants pour compléter les phrases (les mots sont présentés sous la forme qui convient au contexte approprié) :

salaires – patron - usines - licencierait – dirigerai – condition – emploi - impôts

1) Une entreprise qui _____ massivement aurait beaucoup de mal à continuer à attirer les jeunes diplômés.

2) Il éprouve de plus en plus de difficultés à se "vendre" à un employeur qui lui préférera souvent un non-chômeur cherchant à changer d'_____.

3) Le rachat du constructeur tchèque par Volkswagen, a entraîné un accroissement de la productivité, et aussi des _____ …

4) Pourtant les _____ tournent au ralenti avec un chômage partiel important.

5) Les revenus nécessaires pour les financer sont levés principalement sous forme d'_____ et de charges sociales.

6) Il en est ainsi parce que se former n'est qu'une _____ nécessaire à l'emploi.

7) Le _____ d'une P.M.E., c'est un travailleur comme un autre.

8) À l'avenir, je _____ aussi mes recherches du côté des collectivités locales.

Figure 4.3 : activité de recontextualisation

Cet exercice peut être complexifié en donnant les vocables à placer uniquement sous leur forme canonique. La tâche est alors un peu plus difficile pour l'apprenant qui ne peut plus compter sur la flexion des vocables pour les placer dans les différentes phrases, mais doit aussi les fléchir pour les intégrer correctement.

Pour Tréville et Duquette, cet exercice sous cette forme est plutôt destiné aux apprenants de niveau élémentaire. Les activités s'adressant à des niveaux plus élevés (intermédiaire à avancé) doivent mettre en jeu des relations linguistiques qui permettent d'explorer le lexique de manière plus systématique. Il s'agit des relations sémantiques (synonymie, antonymie, hyperonymie, actance), des dérivations syntaxiques (pour passer d'une catégorie grammaticale à une autre), des rapports des co-occurrences ou collocationnels et enfin des valeurs stylistiques et pragmatiques.

Il semble donc judicieux de reprendre l'exercice vu plus haut et de l'améliorer de manière à faire intervenir ces relations. Le principe de recontextualisation étant toujours le même, les

relations interviennent dans le choix des vocables à retrouver qui ne sont plus pris indépendamment les uns des autres.

La figure 4.4 montre un exercice de recontextualisation sur une relation d'actance. A partir du vocable de départ étudier (choisi dans le dictionnaire personnel pour son niveau de traitement), le système calcule les actants déclarés à cette entrée dans la base de données lexicales, recherche les concordances contenant ces vocables et génère l'activité (version avec formes canoniques) :

Utilisez les mots suivants, en les adaptant au contexte, pour compléter les phrases :

Discipline – étude – étudiant – étudier – matière - université

1) Soit qu'ils _____, soit qu'ils soient au chômage.

2) Cette opération impose des _____ qui sont en cours et une validation scientifique incontestable.

3) Philippe, un _____ de 22 ans, ne gardera pas un très bon souvenir de son séjour en Angleterre.

4) Je me lève tôt, je m'habille correctement : il faut se donner une _____, structurer sa vie.

5) Diplômé de Supélec, je n'avais aucune compétence en _____ fiscale, marketing ou encore commerciale.

6) La persévérance de Vincent Decloitre, jeune diplômé de l '_____ en Sciences Eco, a payé.

Figure 4.4 : relation d'actance dans une activité de recontextualisation

Dans cet exercice, l'apprenant renforce les liens dans son lexique mental entre les vocables « satellites » (les actants) d'étudier en réfléchissant (travail en profondeur) sur les relations et les sens propres de ces vocables.

Une autre version de cette activité fait intervenir une autre relation linguistique : la dérivation syntaxique. Les vocables à retrouver sont les dérivés d'un vocable précisé dans l'énoncé. Ici, la relation est suffisamment évidente et forte pour permettre une activité orientée plus en production en omettant les dérivés et en invitant l'apprenant à les trouver et à les placer (figure 4.5).

Complétez les phrases suivantes avec les mots appartenant à la même famille lexicale qu'emploi (les mots doivent être adaptés au contexte) :

1) La qualification sera de plus en plus déterminante dans l'obtention d'un _____.

2) GL : Ben, non, c'est-à-dire que le, l'_____ cherchait quelqu'un qui était opérationnel tout de suite.

3) Au bout de deux mois, j'ai trouvé mon premier travail : _____ de magasin.

4) Mais sur la totalité des entreprises existant en France, seule une sur deux _____ plusieurs salariés.

Figure 4.5 : dérivation syntaxique dans une activité de recontextualisation

Dans cette activité, l'apprenant doit dans un premier temps découvrir (en plus d'emploi) les vocables employeur, employé et employer, puis les placer dans les phrases. La dérivation est une relation intéressante qui fait intervenir plusieurs facettes des vocables. L'apprenant doit effet maîtriser la morphologie (isolation de la racine du vocable de départ), la syntaxe (dérivation et production des vocables correspondant aux différentes catégories grammaticales) et la sémantique (les dérivés ne se distinguent pas uniquement par la catégorie grammaticale mais par leur sens comme employeur et employé qui sont tous les deux des substantifs).

Une dernière variante concerne les relations sémantiques telles que la synonymie, l'antonymie ou l'hyperonymie (figure 4.6).

Complétez les phrases suivantes avec les mots suivants, synonymes d'entreprise (les mots doivent être adaptés au contexte) :

boîte – entreprise – firme – industrie – société - usine

1) En mobylette, j'ai fait le tour de toutes les _____ autour de chez moi.

2) J'ai profité de ce stage pour parler aux délégués des _____ pharmaceutiques qui passaient au cabinet.

3) Péchiney s'apprête à construire, à Dunkerque, une _____ géante d'aluminium et à créer, dans cette cité du chômage, 2 000 emplois.

4) L'_____, le "high-tech" et même l'informatique n'ont rien offert de tel.

5) Vous avez le droit d'installer votre _____ chez vous.

6) Le perpétuel mouvement de création et de destruction des _____ entraîne une valse des effectifs

Figure 4.6 : Relations de synonymie dans une activité de recontextualisation

La synonymie représente sans doute la version la plus difficile de l'exercice, car les distinctions entre les différents candidats est souvent uniquement sémantique et les indices contextuels ne sont pas toujours faciles à détecter. Par exemple dans la phrase 1) :

En mobylette, j'ai fait le tour de toutes les _____ autour de chez moi.

l'apprenant a le choix entre entreprise, société ou boîte, ces trois vocables étant tout à fait corrects à insérer. Toutefois le plus approprié semble être boîte, du fait de la présence de mobylette qui laisse à croire que le locuteur est jeune et utilise donc un langage plus familier. Interviennent ici des notions de valeur stylistique et de registre.

D'autre part, dans certains cas, les différences entre les vocables sont tellement minimes que plusieurs solutions sont possibles. En effet, le contexte fait converger les lexies vers un sens identique. Par exemple, dans la phrase 6), firme, entreprise et société sont acceptables (firme est en fait la solution). Il convient donc de tenir compte de ce phénomène en ne pénalisant pas les apprenants lorsqu'ils fournissent une solution acceptable, même si elle n'est pas rigoureusement exacte (c'est-à-dire correspondant au vocable occulté).

Ces variantes avec relations linguistiques ne sont possibles qu'en faisant intervenir un dictionnaire fonctionnant sous la forme de thésaurus/réseau sémantique. Il s'agit donc d'en tenir compte lors de la conception du dictionnaire général du système qui dès lors sort de son rôle uniquement consultatif.

5 Le jeu du Mai

Dans l'activité précédente, les concordances extraites des textes correspondaient chacune à un vocable différent. On peut faire en sorte qu'elle s'appliquent toutes à un même vocable : c'est le jeu du Mai (Descamps *et al*., 1996). Ce jeu consiste à prendre un vocable relativement polysémique et à présenter plusieurs phrases qui contiennent ce vocable, en prenant soin de faire varier les sens. L'apprenant doit pouvoir retrouver le vocable de départ (figure 4.7) :

Trouvez le mot manquant en l'adaptant au contexte de chaque phrase :

Si vous avez déjà occupé plusieurs _____, choisissez ceux qui peuvent vraiment intéresser l'entreprise.

Envoyer des CV en grand nombre par la _____ ne suffit plus, poursuit le consultant, "Ce qui compte, c'est l'enthousiasme qu'on fait passer dans sa candidature qu'il faut toujours relancer !

La grande salle de montage des _____ de T.V. couleur a une disposition symbolique.

Déjà chacun des deux euroconseillers en _____ à Nice et à Vintimille dispose de banques de données susceptibles de renseigner les demandeurs d'emploi sur les conditions de vie et de séjour dans les deux pays.

La police est intervenue et nous avons vite fait de nous retrouver au _____ de garde.

Figure 4.7 : Le jeu du Mai autour du vocable poste

Cette activité travaille surtout les différentes acceptions d'un vocable (voire de différents homonymes) mais de manière approfondie en explorant la totalité du contenu sémantique d'un vocable (ou des homonymes). Les indices servant à repérer le vocable appartiennent à des contextes différents et il faut en connaître plusieurs pour pouvoir arriver à ses fins. Ceci est important car l'apprenant se contente le plus souvent de connaître uniquement les lexies les plus courantes ou centrales d'un vocable, ignorant celles qui sont moins répandues et plus spécifiques, de fréquence moindre, mais qui participent directement à la compétence lexicale, en déterminant la taille du vocabulaire du locuteur.

L'exercice donne donc la possibilité à l'apprenant d'être exposé à des acceptions inconnues de lui. Il garde sa pertinence pédagogique car l'apprenant devrait être capable de reconnaître quelques acceptions dans les différentes concordances. Ce genre d'exercice est proposé sur le site Internet des dictionnaires Collins Cobuild (Collins, 1999 : comp_entry.html) (figure 4.8).

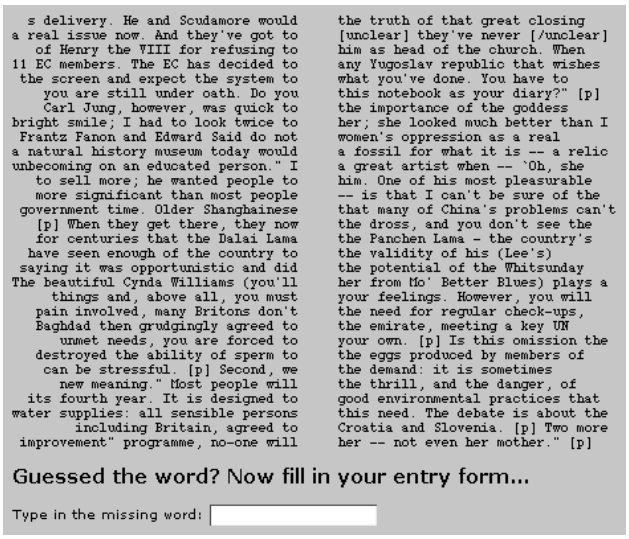


Figure 4.8 : sorte de jeu du Mai, version Collins Cobuild

Il s'agit ici d'un jeu destiné au grand public anglophone (ou apprenants d'un bon niveau) qui met en œuvre des concordances générées automatiquement à partir du corpus The Bank of English. Même s'il s'agit de concordances avec un contexte droite et gauche limité (il ne s'agit pas de phrases), même si on n'a pas été confronté consciemment et précédemment au vocable à trouver, chacun conviendra que cette activité n'est pas vraiment facile. Il importe donc, dans le cadre d'activités lexicales dédiées spécialement aux apprenants, de pouvoir proposer une aide. C'est ce que nous allons examiner maintenant.

6 L'aide

L'autre intérêt d'un dictionnaire sous forme de réseau sémantique est de pouvoir fournir une aide à l'utilisateur. En effet, le système ne doit pas se contenter de répondre vrai ou faux aux solutions proposées par l'apprenant, auquel cas son rôle pédagogique ne serait pas vraiment évident. Il doit être capable, en cas de mauvaise réponse, de pouvoir fournir un indice permettant de mettre sur la voie de la bonne solution. Une aide possible est justement un synonyme, à condition toutefois que le vocable en question en ait un et qu'il soit dans la base. Dans la négative, le système pourrait quand même venir en aide en indiquant un vocable relié sémantiquement (par exemple un actant, un dérivé) et en explicitant la relation.

L'aide peut aussi provenir du dictionnaire personnel : le système peut donner les autres vocables du groupe qui contient le vocable à trouver ou le nom de ce groupe. L'apprenant est ainsi invité à se remémorer la manière dont il a structuré son propre vocabulaire et ce travail de remémorisation renforce les liens du lexique mental.

Enfin, pour aider l'apprenant, il reste toujours les indices classiques qui font partie du vocable lui-même : initiale, premières lettres, nombre de lettres, etc…

7 Conclusion

Nous nous proposons donc d'intégrer ces deux types d'activités (recontextualisation et jeu du Mai) dans ALEXIA. Comme nous avons pu le constater, ce chapitre a encore souligné l'importance des réseaux lexicaux pour l'acquisition. Dès lors, la modélisation de notre base lexicale nécessite de prendre en compte ce phénomène en permettant notamment l'extraction d'informations spécifiques à cette organisation (ensemble de synonymes, d'actants, de dérivés). Nous verrons dans les chapitres 7 et 8 consacrés à nouveau aux activités lexicales la préparation nécessaire et les problèmes qui peuvent se poser lors de la conception de ces activités.

CHAPITRE 5

Le dictionnaire électronique d'ALEXIA

Après quatre premiers chapitres consacrés à l'examen de l'existant, nous abordons la conception des modules formant le système ALEXIA. Nous commencerons par le dictionnaire général. Avant cela, pour mieux nous rendre compte de l'environnement lui-même, nous présentons rapidement par une figure son architecture, évoquant par cela les cinq modules qui le constituent : le corpus de textes et le module de lecture, le dictionnaire général, le dictionnaire personnel, le module d'activités lexicales et le modèle de l'apprenant (figure 5.1).

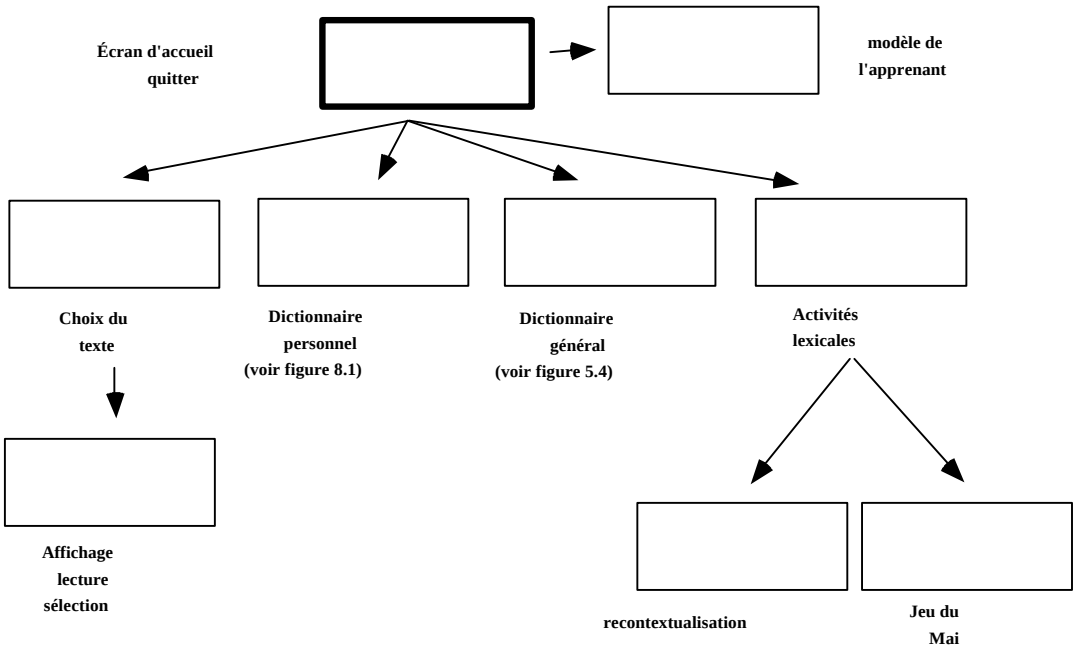


Figure 5.1 : Architecture générale et navigation du système ALEXIA

Nous allons maintenant décrire dans ce chapitre les concepts qui ont mené à l'élaboration du dictionnaire général. Nous abordons en premier lieu l'organisation interne de la base de

données lexicales, pour passer ensuite à la manière dont l'apprenant accède et visualise l'information.

L'objectif de ce travail n'est pas de produire un dictionnaire complet, dans lequel chacun des problèmes soit traité et bien résolu. Comme nous l'avons vu dans le chapitre 3, les dictionnaires sont des ouvrages d'une grande densité et d'une grande variété d'informations. Ils font appel à de nombreux champs de la linguistique. De plus, dans le cas présent des dictionnaire électroniques, il y a les considérations informatiques qui viennent se rajouter. Aussi, toutes les différentes parties n'ont pas pu être traitées complètement. Nous avons mis l'accent sur l'accès lexical, les définitions, la synonymie, mais d'autres aspects tels que l'antonymie ou la présentation des collocations n'ont été abordés que partiellement.

Naturellement, pour pouvoir illustrer notre propos et présenter les problèmes, nous devons nous appuyer sur des données lexicales. Nous avons vu au chapitre 3 qu'aucun dictionnaire pour apprenants du français n'était satisfaisant. Aussi, nous avons dû développer notre propre base lexicale, à partir du corpus dont nous disposions et des autres dictionnaires existants.

Nous avons fait de notre mieux, mais n'étant pas lexicographes de formation, nous demandons une certaine indulgence au lecteur qui jugera peut-être contestable tel ou tel aspect de notre description.

1 La lexie, unité sémantique de base

La grande majorité des vocables du vocabulaire courant sont polysémiques. La plupart des vocables monosémiques appartiennent à des vocabulaires de spécialité (ou terminologies) et sont utilisés dans des domaines bien précis. La plupart des expressions semi-figées sont aussi monosémiques. Mais dans le reste du lexique, l'apprenant est confronté la plupart du temps à des vocables qui ont plusieurs acceptions. Même s'il pourrait être utile, d'un point de vue cognitif et pédagogique, de dégager et d'enseigner l'unité profonde d'un vocable et les mécanismes qui ont abouti à ses différentes acceptions, celle-ci peut s'avérer bien vague et difficile à isoler (Bogaards, 1994, pp. 16). De même, les mécanismes de génération ne sont pas toujours simples à mettre en évidence et ne sont pas toujours réguliers, du moins si l'on veut expliquer toutes les acceptions du vocable.

Les choses semblent bien plus simples et bien plus claires si l'on place l'acception d'un vocable comme unité sémantique de base. Comme vu dans l'introduction, nous reprenons la terminologie de I. Mel'cuk (Mel'cuk *et al.*, 1995) en parlant de lexie. Les différents sens d'un vocable sont sentis distinctement et l'on peut alors les décrire par une définition. Lorsqu'on demande le sens d'un vocable à une personne, celle-ci répond le plus souvent par le sens d'une de ses lexies, souvent la lexie de base, dont toutes les autres découlent, ou lexie vedette. Outre les informations propres au vocable qui sont d'ordre morphologique et phonétique, c'est la lexie qui porte toutes les autres. En effet, les combinaisons syntaxiques, les dérivés, le niveau de langue, les synonymes, antonymes, fonctions lexicales, etc. sont toutes fonctions de l'acception d'un vocable. Connaître un vocable, c'est donc connaître la signification et l'emploi de ses lexies sur les plans syntaxiques et pragmatiques. D'un point de vue quantitatif, accroître son vocabulaire, c'est en fait accroître le nombre de lexies que l'on connaît, que ce soit en

apprenant d'autres vocables (d'autres lexies) ou en apprenant d'autres sens de vocables déjà (partiellement) connus, c'est-à-dire dont une lexie au moins est connue. Plusieurs lexies pouvant avoir le même signifiant, il est commode de les regrouper sous un même vocable qui sont les entrées du dictionnaire standard, organisé par ordre alphabétique.

Dans les réseaux du dictionnaire d'ALEXIA, la lexie est un élément fondamental. En effet, toutes les relations linguistiques s'appuient sur ce concept : les vocables ne sont reliés entre eux que de sens à sens. Dire que travail est synonyme d'emploi n'a aucun sens hors de tout contexte.

2 La modélisation de la base lexicale

Nous présentons ici les points saillants de l'implémentation de la base de données lexicales. Celle-ci est abordée en détail dans les annexes.

Nous avons vu au chapitre 3 que les dictionnaires anglais d'apprentissage actuels étaient tous élaborés à partir de corpus représentatifs de la langue qui se distinguent en outre par leur taille, une centaine de million de mots au minimum. Les corpus sont en effet source d'authenticité et guident le lexicographe dans sa description : il est possible à l'aide de concordanceurs d'observer le fonctionnement réel de la langue et ces derniers permettent d'attester le sens des vocables (notamment l'évolution du sens), leur combinatoire syntaxique, les collocations, etc. Les corpus s'opposent donc aux présupposés et à l'intuition du linguiste sur la langue.

La taille de ces corpus font qu'en principe ils peuvent pratiquement prétendre à l'exhaustivité, tout du moins dans l'anglais général. Pour nous qui ne disposons pas de ces moyens réservés à des équipes entières de lexicographes et qui sont le fruit de longues années de travail, le corpus dont nous disposons (caractéristiques décrites au chapitre 9) nous a permis de déterminer quels étaient les vocables les plus représentatifs, c'est-à-dire les plus fréquemment attestés dans les textes, du domaine que nous avons isolé.

A l'aide du logiciel HyperBase (Brunet, 1994), nous avons dressé la liste par ordre de fréquence décroissante des formes fléchies (ce logiciel n'intègre pas de lemmatiseur). Cette liste se limitant, de manière assez étonnante, aux 165 premières formes, nous avons pu isoler 11 vocables pôles : emploi, travail, entreprise, chômage, formation, salarié, chômeur, activité, société, social, et cadre. Cette liste contient uniquement, parmi les formes les plus souvent attestées, celles en relation avec notre domaine de description. Nous avons donc éliminé, outre les mots grammaticaux, les verbes et les noms généraux indépendants du sujet que l'on retrouve dans n'importe quel texte (être, faire, etc.). Nous avons ensuite constitué des réseaux lexicaux autour de ces vocables pôles (synonymes, antonymes, actants, dérivés syntaxiques, etc.) et avons obtenu une deuxième liste plus complète. Soucieux de traiter le phénomène important des collocations, nous avons associé aux vocables pôles celles qui avaient été extraites automatiquement du corpus par les travaux de Rabetifia (1994). Un certain nombre ont été ajoutées à la main, car elles n'étaient pas ou peu attestées en raison de la taille insuffisante de notre corpus pour décrire correctement ce phénomène. Bosser comme un fou, ne pas chômer tombent dans cette catégorie. Quelques vocables, présents dans ces collocations ont été aussi été

ajoutés en tant qu'entrée simple (cas de condition dans conditions de travail ou durée dans durée du travail) afin que l'apprenant puisse avoir plus d'information sur les constituants des collocations. A ce jour, le dictionnaire général (par opposition au dictionnaire personnel) d'ALEXIA est composé de 210 vocables, ce qui représente 383 lexies. La liste de ces vocables est présentée dans l'annexe A. Nous sommes conscients que cette base de données lexicales constitue en quelque sorte un dictionnaire « idéalisé » dans le sens où l'on s'affranchit des lourdes contraintes et des nombreux phénomènes qu'impose une masse de données correspondant à 60 000 entrées. Néanmoins, la délimitation d'un domaine d'étude suffisamment généralisable vise à reproduire les phénomènes lexicaux qui apparaissent à l'échelle de la langue. Ainsi, par exemple, les vocables sont reliés en réseau par la plupart des relations sémantiques standard (synonymie, actance, dérivation syntaxique,…), les définitions de leurs lexies sont décrites grâce à d'autres entrées du dictionnaire, nous sommes confrontés aux problèmes de polysémie et d'homonymie, aux collocations, etc.

2.1 Structure de la base lexicale

Le lexique est un élément fondamental d'une langue. Chacune de ses unités, le vocable, est porteuse de nombreuses informations, partagées entre les lexies, comprenant le sens et l'emploi dans le discours.

Les informations pertinentes pour les apprenants sont de trois types :

- les informations sémantiques : sens d'un vocable, relations avec d'autres appartenant au même paradigme (synonymie, antonymie, hyperonymie et hyponymie), actance
- les informations morpho-syntaxiques : forme canonique et flexions, genre et nombre, phonétique, décomposition en morphèmes, constructions syntaxiques, dérivés
- les informations pragmatiques : niveau de langue, fréquence, intention du locuteur, connotation

Pour la modélisation de la base lexicale, on peut distinguer :

- les informations propres aux vocables et à ses différentes lexies : morphologie, définitions, exemples, niveaux de langue, constructions grammaticales, etc.
- les liens de lexies entre elles : synonymies, antonymies, actance, dérivation syntaxique, collocations.

Pour les premières, le dictionnaire peut être conçu comme une base de données relationnelles. Il y aurait une première table qui contiendrait les vocables et une deuxième pour les lexies qui serait liée à la première par une relation 1:N. Chaque vocable ou lexie serait un enregistrement et les différents types d'information (morphologie, définitions, syntaxe) un champ.

Même s'il est possible de traiter les liens entre lexies de la même manière en considérant des champs synonymes, actants, collocations, etc., ceci n'est pas souhaitable car source de beaucoup de duplication et donc de possibilité d'incohérence et d'erreur.

Prenons l'exemple de emploi2 (la notation signifie lexie emploi sens 2). Cette lexie admet 13 synonymes au sens large, comme travail2a (équivalent), boulot1a (équivalent), activité2a

(hyperonyme), gagne-pain1 (hyponyme), etc. Il est tout à fait justifié de présenter ces 13 lexies à un apprenant lorsqu'il demande les synonymes de emploi2, soit sous forme textuelle (liste), soit sous forme de graphe, de manière à cerner la notion emploi-travail. Mais le système doit être aussi capable de lui présenter ces 13 lexies lorsqu'il demande les synonymes de travail2a (avec emploi2 à la place de travail2a) ou lorsqu'il demande ceux de boulot1a. En effet emploi2, travail2a et boulot1a sont équivalents, au registre près, et possèdent de ce fait les mêmes synonymes.

Si l'on garde la structure d'une base de données relationnelle, le champ synonyme de l'enregistrement emploi2 aura 13 éléments, celui de travail2a et boulot1a aussi. Qui plus est, ils seront quasiment les mêmes. Ceci est source non seulement de duplication inutile, de possible incohérence et est difficile à maintenir. La modification des synonymes d'une lexie implique celle des lexies qui lui sont liées.

On s'aperçoit que les liens entre lexies sont de type N:N et donc difficile à gérer sous forme de champs et d'enregistrement. Il faut plutôt adopter une structure de réseaux, à l'image de ce qui a été fait dans WordNet (Miller *et al*., 1993).

WordNet est un thésaurus électronique anglais issu de travaux en psycholinguistique sur l'organisation du lexique mental. Cette base lexicale a été pensée et conçue pour le support électronique. Son objectif est de décrire comment les lexies s'organisent les unes par rapport aux autres. Plutôt que de lexies, les réseaux de WordNet sont constitués de *concepts*, véritables unités sémantiques de base qui sont en fait des ensembles de synonymes. « Ces ensembles de synonymes (synsets) n'expliquent pas ce que sont les concepts ; ils en posent l'existence. Nous supposons donc que les locuteurs anglophones, ayant déjà acquis ces concepts, sont capables de les reconnaître à partir des mots listés dans le synset. » (Miller *et al*., 1993, pp. 5-6). Les concepts sont reliés par des relations d'hyponymie, d'antonymie, de méronymie, d'implication ou de dérivation morphologique. Mais la relation privilégiée est bien sûr la synonymie, interne aux concepts. WordNet privilégie quatre catégories grammaticales : les noms, les verbes, les adjectifs et les adverbes. Chaque catégorie a sa propre structure interne : « ce sont des expériences sur les associations de vocables qui ont mis en évidence à l'origine que l'organisation varie d'une catégorie syntaxique à l'autre ».

De par son importance (WordNet couvre toute la langue anglaise), WordNet est probablement la base de connaissance électronique la plus utilisée. Elle constitue une référence dans le domaine.

A l'image de WordNet, la base lexicale d'ALEXIA devient donc un réseau, un graphe, dans lequel les nœuds sont les lexies avec toutes leurs informations propres (et notamment le vocable) et les arcs sont les différentes relations sémantico-syntaxiques. L'accès aux lexies et aux informations se fait donc soit par accès direct à un nœud du graphe, soit à l'aide d'opérations de parcours de graphe.

Pour implémenter une base de données lexicales qui a les propriétés vues ci-dessus, notre choix s'est porté sur Prolog. En effet, ce langage s'avère pratique tant pour établir la base de données que les opérations parcourant le réseau. Ainsi chaque vocable est identifié par un nom d'identificateur et chaque lexie reliée à ce vocable par un foncteur reprenant l'identificateur et un numéro de sens. Les données propres au vocable, c'est-à-dire communes à toutes ses lexies, sont les arguments du prédicat `entree` (identificateur, catégorie

grammaticale, code pour le genre et le nombre, graphie, trait distinguant les éventuels homonymes et fréquence du vocable). Celles propres à une lexie sont les arguments du prédicat `entreesec` (identificateur et numéro de sens, mini-définition, définition, schémas grammaticaux (transitivité, préposition, etc.), contraintes sémantiques sur les sujets et compléments, variations lexicales des collocations, registre et exemple). Chaque vocable et chaque lexie est un fait Prolog que l'on peut formaliser par :

```
Entree(Ident,Catgram,GenreNombre,Graphie,Homonyme,Frequence).
Entreesec(Ident,Sens,MiniDef,Definition,SchemaGram,Contraintes,VarLex,Registre,Exem
ple)
```

Exemple pour le vocable emploi (figure 5.2) :

```
entree(emploi,cat:n,gn:1,chaine:"emploi",hom:"",freq:139).

entreesec(emploi('1'),mini:"manière d'utiliser une chose",def:["l'emploi d'une
chose est l'action ou la manière de l'utiliser, de s'en
servir",3,8,0,[]],formes:[['N']],nil,nil,reg:2,exemple:[148]).

entreesec(emploi('2'),mini:"ce que fait une personne pour gagner de
l'argent",def:["un emploi est le travail qu'une personne fait de manière régulière
pour gagner de
l'argent",4,9,1,[[18,24,travail,'2a']]],formes:[['N']],nil,nil,reg:2,exemple:[149])
.

entreesec(emploi('3'),mini:"ensemble de ceux qui ont un emploi",def:["l'emploi est
un fait économique et social qui représente l'ensemble des personnes qui ont un
emploi et le travail qu'elles
font",3,8,3,[[36,41,social,'1a'],[94,99,emploi,'2'],[107,113,travail,'2a']]],formes
:[['N']],nil,nil,reg:2,exemple:[150]).
```

Figure 5.2 : extrait de l'entrée emploi

Nous allons maintenant examiner l'implémentation des relations entre lexies. Les informations propres au vocable ou à ses lexies (morphologie, construction des verbes, registre, etc.) sont simplement déclarées dans les faits Prolog et ne présentent pas de difficultés. Les détails de ces déclarations sont donnés dans les annexes.

Les liens entre lexies sont décrits par d'autres faits Prolog de nature différente qui représentent un arc du graphe. En principe, cet arc devrait être un triplet qui comprend les deux lexies liées et leur relation. En pratique, d'autres informations s'y ajoutent comme le registre ou la fréquence dans le cas où l'un des deux nœuds n'est pas une entrée du dictionnaire (et sur lesquelles le dictionnaire n'a aucune information). Il peut aussi y avoir des informations plus spécifiques comme la catégorie grammaticale ou le numéro d'argument dans le cas des actants et des dérivés syntaxiques. Un prédicat est réservé pour chaque type de relation : `semantique` est utilisé pour les synonymies, antonymies et fonction lexicales ne débouchant pas sur des collocations, `cooc` pour les co-occurrences et collocations, `actant` pour les actants et `derive` pour les dérivés syntaxiques. Les prédicats `semantique`, `cooc`, `actant` et `derive` peuvent être formalisés de la manière suivante :

```
semantique(A,B,Rel).
cooc(A,B,Rel).
```

```
actant(A,B,CatGram,NumArg).
derive(A,B,CatGram,NumArg).
```

où `A` est une lexie, `B` une lexie ou une graphie suivant le cas, `CatGram`, la catégorie grammaticale de `B` et `NumArg`, son numéro d'argument (1 pour le sujet, 2 pour le premier complément, etc.)

Exemples avec la figure 5.3 :

```
semantique(travail('2a'),job('1'),sem:syneq).
semantique(travail('1'),corvee('1'),sem:magn).
semantique(travail('1'),"repos",sem:antoeq,reg:2).
cooc(travail('1'),"bourreau de ~~",sem:magn(agent),reg:2).
actant(travailler('2'),employeur('1'),cat:n,arg:3).
derive(travailler('1'),travail('1'),cat:n,arg:0).
```

Figure 5.3 : exemples de relations sémantiques (synonymie, antonymie, actance, etc.)

Les opérations de parcours de graphe sont des règles Prolog et ont été écrites pour tirer parti des propriétés linguistiques, que nous allons développer, et informatiques. Elles ont été appliquées principalement aux synonymes, actants et dérivés syntaxiques.

2.2 Les différentes relations lexicales

2.2.1 Les relations synonymiques

Nous décrivons les relations suivantes : quasi-synonymie, hyperonymie et hyponymie, synonymie intersective.

Quasi-synonymie

Nous appelons quasi-synonymie la similarité de sens de deux lexies. Deux lexies sont quasi-synonymes si la substitution de l'un par l'autre dans une phrase ne change jamais la valeur de vérité de cette dernière. Nous utilisons le terme de quasi-synonymie car il y a toujours de petites différences entre deux lexies : différences fines de sens, d'emploi, d'occurrence avec d'autres vocables, de constructions syntaxiques, de connotation, etc.

Lorsqu'on décrit des relations de quasi-synonymie entre vocables, il faut toujours tenir compte d'un contexte linguistique. En ce sens, deux vocables ne sont jamais synonymes exacts dans l'absolu, c'est-à-dire dans tous les contextes, mais ils peuvent l'être dans des contextes spécifiques. Par exemple, on pourra remplacer emploi par utilisation pour exprimer l'idée de « se servir de » mais cette substitution ne sera plus appropriée pour les autres sens d'emploi. Il faut donc toujours prendre en compte un contexte d'une manière ou d'une autre lorsqu'on considère une relation de quasi-synonymie entre vocables. En général, il est donné par la phrase. Cette contrainte n'existe plus lorsque la description se situe au niveau de la lexie car le contexte n'est plus nécessaire pour déterminer le sens, celui-ci étant spécifié explicitement.

Cependant, lorsque ce contexte disparaît, comme c'est le cas dans les dictionnaires, il devient nécessaire de le réintroduire en spécifiant explicitement quels sont les lexies quasi-

synonymes, ce qui est malheureusement rarement le cas dans les dictionnaires du français (excepté le DEC, Mel'cuk, 1992). Tout au plus avons-nous une synonymie entre une lexie (présentée dans tel sens du vocable) et un vocable. Une des spécificité d'ALEXIA est de décrire systématiquement les relations sémantiques entre lexies. Cette caractéristique nous semble très importante pour les apprenants. Établir une quasi-synonymie entre vocables (quand ce n'est pas entre mots) à la place de lexies ne fait qu'augmenter le risque de confusion.

Hyperonymie et hyponymie

L'hyperonymie et l'hyponymie sont deux relations sémantiques qui découlent de l'organisation hiérarchique du lexique. Une lexie A est un hyponyme d'une lexie B (ou une lexie B est un hyperonyme d'une lexie A) si le sens de A est plus spécifique que celui de B, si tous les traits sémantiques de B appartiennent à A. Les lexies sont liées par une relation « sorte de ». Ainsi, enseignant est un hyponyme de métier et métier est un hyponyme d'activité.

La relation d'hyper/hyponymie est tout à fait évidente pour les noms désignant des objets concrets (comme chaise et meuble). Mais lorsque l'on est confronté à des noms plus abstraits, la relation est plus difficile à concevoir. Concernant les verbes, il apparaît que la relation d'hyper/hyponymie est plus complexe que pour les noms. En effet, verbes et noms ne partagent pas les mêmes caractéristiques (Fellbaum, 1993) et, tandis que la relation principale de l'hyper/hyponymie est une relation « sorte de » pour les noms, il s'agit d'une relation d'implication pour les verbes. Dans son article, Fellbaum montre que l'implication regroupe en fait quatre implications lexicales. Dans ALEXIA, nous utilisons principalement la troponymie : un verbe A est un troponyme d'un verbe B si l'on peut dire :

A, c'est B d'une certaine façon

Par exemple, « destituer, c'est renvoyer d'une certaine façon ». Ici, « certaine façon » doit être interprété assez vaguement pour permettre de multiples interprétations sémantiques.

La troponymie implique la co-extensivité. Ceci signifie que A est un troponyme de B si A implique B et si A et B partage la même étendue de durée, c'est-à-dire que les activités de A se déroulent en même temps que celles de B (Fellbaum, 1993).

Synonymie intersective

La synonymie intersective est une relation entre deux lexies qui ont seulement certains traits sémantiques en commun. Bien qu'il y ait intersection de sens, elles ne peuvent être considérées comme quasi-synonymes. Par exemple, profession et emploi sont deux synonymes intersectifs.

Concernant les synonymes, les deux propriétés linguistiques sont la réciprocité et la transitivité.

2.2.2 Réciprocité

Soit deux synonymes équivalents A et B. Si A est synonyme de B, alors naturellement B est synonyme de A. Cette relation (arc) ne doit pourtant être exprimée qu'une fois dans la base. Or un fait Prolog étant non symétrique, il faut donc construire une règle de réciprocité.

La relation synonymique de base est exprimée par

```
syn(A,B,Rel).
```

où les lexies `A` et `B` sont reliées par la relation `Rel` (quasi-synonymie, hyperonymie, hyponymie ou synonymie intersective).

Pour obtenir une relation réciproque, on a donc la règle `synrec` telle que :

```
synrec(A,B,eq) <= syn(A,B,eq) ∨ syn(B,A,eq)
```

`eq` exprime l'équivalence (ou quasi-synonymie).

La réciprocité s'applique de la même manière à la synonymie intersective :

```
synrec(A,B,inter) <= syn(A,B,inter) ∨ syn(B,A,inter).
```

L'hyperonymie et l'hyponymie sont des relations réciproques par nature, à savoir que si `A` est hyperonyme de `B`, alors `B` est hyponyme de `A`. Soit `B`, hyponyme de `A`, on a donc

```
synrec(A,B,pe) <= syn(A,B,pe) ∨ syn(B,A,pl)
```

Et de même

```
synrec(A,B,pl) <= syn(A,B,pl) ∨ syn(B,A,pe)
```

2.2.3 Transitivité

Cette propriété permet de regrouper plusieurs synonymes équivalents. Soit A, B et C trois synonymes, si A est synonyme de B et B est synonyme de C, alors A est synonyme de C. La transitivité peut d'ailleurs être étendue à n synonymes et il faudra n-1 relations (arcs) plus une règle Prolog de transitivité pour décrire l'ensemble des liens.

On aura donc, par exemple, pour quatre synonymes équivalents `A1`, `A2`, `A3`, `A4` :

```
syn(A1,A2,eq).
syn(A2,A3,eq).
syn(A3,A4,eq).
```

et la règle de synonymie transitive `syntran` :

```
syntran(A,B,eq) <= synrec(A,C,eq) ∧ syntran(C,B,eq).
```

La relation de synonymie égale est une relation privilégiée qui implique certaines règles concernant les relations d'hyperonymie et d'hyponymie. En effet, plusieurs synonymes égaux forment un concept, et toutes les relations que possède l'un des synonymes sont valables pour les autres. On peut ainsi dégager les règles suivantes d'hyperonymie étendue :

```
syntran(A,B,pl) <=   (synrec(A,C,pl) ∧ syntran(C,B,eq)) ∨
                     (syntran(A,C,eq) ∧ synrec(C,B,pl)) ∨
```

```
                    (syntran(A,C,eq) ∧ synrec(C,D,pl) ∧ syntran(D,B,eq)).
```

et d'hyponymie étendue :

```
syntran(A,B,pe) <=   (synrec(A,C,pe) ∧ syntran(C,B,eq)) ∨
                     (syntran(A,C,eq) ∧ synrec(C,B,pe)) ∨
                     (syntran(A,C,eq) ∧ synrec(C,D,pe) ∧ syntran(D,B,eq)).
```

Nous pourrions aussi considérer les transitivités concernant les hyperonymies ou les hyponymies seules à savoir qu'un hyperonyme d'un hyperonyme d'une lexie est aussi hyperonyme de cette lexie et de même pour les hyponymes. A l'opposé de Wordnet, cette relation n'a pas été développée au sein de la base lexicale principalement pour des raisons didactiques. En effet, on arrive assez rapidement à des lexies de sens trop général ou trop spécifique et qui ne possèdent qu'un rapport étroit avec la lexie de départ. En général, il n'est pas pertinent lorsqu'un apprenant cherche un vocable proche de travail de voir acte (hyperonyme d'activité qui est hyperonyme de travail) ou action humaine. Nous nous sommes donc limités à un seul niveau de transitivité concernant l'hyper/hyponymie.

2.2.4 L’antonymie

L’antonymie a été implémentée sur le même modèle que la synonymie en ne reprenant cependant que la relation de réciprocité. Les relations de transitivité sont beaucoup moins évidentes, à supposer qu’elles existent, car l’antonyme d’un antonyme n’est pas forcément antonyme du premier.

L’antonymie implique aussi d’autres relations linguistiques complexes telles que la gradation, l’opposition ou les relations entre antonymes converses. Ces relations n’ont pas été suffisamment étudiées pour faire l’objet d’une implémentation dans ce travail de thèse.

2.2.5 Actants et dérivés syntaxiques

A côté des synonymies et antonymies, il existent d'autres relations sémantiques intéressantes : l'actance et la dérivation syntaxique.

Nous reprenons l'actance telle qu'elle est explicitée dans Mel’cuk (1995, pp. 75-76), à savoir qu'un actant sémantique correspond à un argument d'un prédicat sémantique. Les prédicats sémantiques désignent des actions, des évènements, des processus, des états, des propriétés, des relations, etc. - en un mot, des faits qui impliquent nécessairement des participants, que l'on désignera par actant. Par exemple, dire est un prédicat à trois actants (quelqu'un (1) dit quelque chose (2) à quelqu'un (3)) ; diriger (sens de commander quelqu'un) est un prédicat à deux actants (quelqu'un (1) dirige quelqu'un (2)) ; courir n'a qu'un actant (celui qui court). Les prédicats peuvent aussi être des noms (crime a deux actants, celui qui tue, le criminel, et celui qui est tué, la victime), des prépositions (devant, sur, qui ont deux actants), des adjectifs (qui qualifient quelqu'un ou quelque chose), etc.

Certains actants peuvent être des dérivés syntaxiques. Les dérivés syntaxiques sont les vocables qui appartiennent à une même famille lexicale, c'est-à-dire un ensemble de vocable ayant même radical et un certain rapprochement sémantique. Ainsi travailler, travail, travailleur,

travaillé sont des dérivés syntaxiques les uns par rapport aux autres. De même pour emploi, employé, employeur, joie, joyeux, joyeusement, etc.

L'intérêt des actants pour la lexicologie et la lexicographie est évident car ils participent directement à la définition et au sens de la lexie. Une définition doit obligatoirement expliciter les actants pour pouvoir décrire le sens de la lexie considérée. Il est impensable de définir aboyer sans employer chien. Les actants sont donc indissociables du prédicat auquel ils sont liés et participent directement à la description du sens exprimé par le prédicat (lexie à sens prédicatif). Les actants sont parfois lexicalisés et il convient donc de présenter à l'apprenant le nom, verbe, adjectif, etc. typique pour désigner tel ou tel actant. Celui-ci peut être mentionné explicitement dans la définition (Lorsqu'un chien aboie, ...), mais dans le cas contraire, il doit être présenté en dehors. C'est le cas pour toutes les lexies et définitions qui contiennent les pronoms indéfinis quelqu'un et quelque chose. Dans la définition « Lorsque quelqu'un travaille quelque chose, ... » , le quelqu'un en question est désigné par le nom générique de travailleur (au sens large) et le quelque chose par l'adjectif générique travaillé (encore que là le sens n'est pas tout à fait direct, travaillé impliquant un certain degré de perfectionnement, une certaine quantité de travail).

L'intérêt des dérivés syntaxiques est tout d'abord la possibilité de paraphraser une expression en changeant la catégorie grammaticale du vocable principal. Par exemple, on peut paraphraser employer telle ou telle chose par l'emploi de telle ou telle chose. Cependant, dans une même famille lexicale, il peut y avoir plusieurs membres qui ont la même catégorie. Il importe alors de spécifier plus précisément les rapports sémantiques dans le cadre d'une même famille. On retrouve alors une certaine similarité avec les actants sémantiques.

Prenons le cas de la famille d'allouer. On y trouve allocation et allocataire. Même si une rapide analyse morphologique peut indiquer que le fait d'allouer est plutôt allocation (le suffixe -tion désignant le plus souvent des actions), celle-ci relève d'une compétence que l'apprenant ne possède pas toujours. En l'occurrence, allocation est aussi, dans un sens différent, la somme d'argent que l'on alloue. Il importe donc de spécifier explicitement qu'on alloue une allocation à un allocataire et non pas le contraire (un allocataire à une allocation). En fait, allocation et allocataire sont les actants du prédicat allouer.

Cependant, l'intérêt des dérivés syntaxiques ne se limite pas seulement à la paraphrase avec changement de catégorie. Les dérivés sont liés par la forme mais aussi par le sens et montrent les possibilités de génération morphologique par ajout de suffixe. Il est possible dès lors de cerner le trait (sens) commun à une famille et de voir quelle nuance sémantique apporte tel ou tel suffixe. D'un autre côté, il faut éviter les phénomènes de surgénération, la composition n'étant pas régulière, par la présentation exhaustive de la famille. L'apprenant évite ainsi de construire à l'aide de suffixe des formes qui n'existent pas. Dans le cas de la famille d'allouer, l'agent allocateur n'existe pas, bien que ce terme soit prédictible en génération et compréhensible pour celui qui connaît le rôle du suffixe -teur (on devine qu'il pourrait s'agir de celui qui alloue).

La présentation la plus naturelle des actants d'un prédicat est celle qui fait intervenir le verbe comme vocable central. Les actants sont alors le sujet et les compléments de ce verbe, que les vocables appartiennent à la même famille lexicale ou pas. Il faut donc utiliser un schéma syntaxique qui définit la syntaxe du verbe à l'aide de variables et assigner un couple

(terme, catégorie) à chaque variable, pour peu que celui-ci existe. Par exemple, pour allouer on a le schéma syntaxique

A1 alloue A2 à A3

avec A2 allocation et A3 allocataire (dans le sens verser une subvention pour allouer) et A0 (variable non-visible mais qui par définition porte sur le verbe) allocation, fait d'allouer, dans le sens octroyer (en informatique, une allocation de mémoire).

On note de la même manière qu'A1 n'est pas lexicalisé.

Concernant la famille crime, criminel, on utilisera le schéma syntaxique du verbe tuer qui exprime le même prédicat. On aura donc

A1 tue A2

avec A1 criminel, A0 crime et A2 victime.

On s'aperçoit donc que les actants ne sont pas toujours de la même famille lexicale, ce qui est une information importante pour l'apprenant. D'un autre côté, les dérivés ne sont pas toujours des actants. En effet, pour ce dernier exemple, comment considérer criminalité ? Ce vocable fait incontestablement partie de la famille, à la fois par la forme et par le sens, mais il n'est pas possible de lui assigner une variable du schéma syntaxique. Il est cependant pertinent de le présenter à l'apprenant, en dehors du schéma et en le renvoyant à une définition pour comprendre le sens précis si nécessaire, pour bien cerner le radical crim-. Plus tard, il pourra se servir de ce radical pour inférer par exemple le sens de criminologue, qu'il pourrait rencontrer dans un texte.

La présentation des dérivés à l'aide d'un schéma syntaxique du verbe de la famille s'avère donc efficace, car l'assignation des actants aux variables permet de comprendre directement les relations fonctionnelles entre les différents vocables (sujet, objet, complément, etc.) et par là leur sens, sans passer par une définition. L'exposition répétée à un même radical augmente les chances de rétention de celui-ci par l'apprenant, ce qui pourra l'aider à l'inférence dans les textes.

Cependant, il est des cas où il n'y a pas de verbe associé au prédicat, que le verbe fasse partie de la famille lexicale ou non. Par exemple, comment présenter la famille joie, joyeux, joyeusement ? En l'absence de schéma syntaxique, on aura recours à une présentation plus classique, sous forme de liste, avec la possibilité d'avoir accès à des définitions, si besoin est. Par contre, l'objectif de cerner le radical reste le même.

Il est aussi des cas où, à l'intérieur d'une même famille lexicale, il peut y avoir des termes plus ou moins proches à la fois sémantiquement et syntaxiquement, c'est-à-dire qui dérivent directement ou pas les uns des autres. C'est le cas avec la famille d'industrie.

Cette famille est constituée de : industrie, industriel nom et adjectif, industriellement, industrialiser, industrialisation et industrialisé. On constate que l'on peut distinguer deux sous-groupes : d'une part, industrie, industriel (adj), industriellement, et d'autre part industrialiser, industrialisation, industrialisé. Industriel (n) est un peu à l'écart, n'étant pas dérivé directement des autres termes. Même si l'intérêt est moindre, il n'est pas inutile de mettre en évidence ces deux sous-groupes, en les séparant par exemple par une ligne, pour montrer les mécanismes de dérivation.

Les actants d'un même prédicat sémantique et les dérivés syntaxiques d'une même famille forment donc un groupe et tout comme les synonymes, il doit être possible d'accéder à chacun des constituants à partir de n'importe lequel d'entre eux. Dès lors, une structure sous forme de graphes avec des opérations de parcours semble la plus adaptée pour une implémentation informatique en Prolog.

La déclaration de relation entre dérivés se fait de la manière suivante :

```
derive(A,B,CatGram,NumArg).
```

où `A` est le plus souvent le verbe de la famille lexicale, `B` le dérivé, `CatGram`, la catégorie grammaticale de `B` et `NumArg` son numéro d'argument.

Ce qui donne, par exemple pour allouer :

```
derive(allouer('1'),allocation('1'),cat:n,arg:2).
derive(allouer('1'),allocation('2'),cat:n,arg:0).
derive(allouer('1'),allocataire('1'),cat:n,arg:3).
```

Les membres de la famille lexicale devant être accessibles à partir de n'importe lequel d'entre eux, il faut donc introduire une opération de transitivité `derives` :

```
derives(A,B,CatGram,NumArg) <=   derive(A,B,CatGram,NumArg) ∨
                                 (derive(Verbe,A,C1,A1) ∧ derive(Verbe,B,CatGram,NumArg)).
```

A côté des actants sémantiques, il y a les actants circonstanciels. Les actants circonstanciels sont des participants optionnels du prédicat qui décrivent les circonstances dans lesquelles celui-ci s'effectue. Nous avons retenu les actants temps, résultat, lieu et prix[5]. C'est ainsi qu'on peut déclarer :

```
actant(etudier('1c'),"université",cat:n,arg:lieu,reg:2).
```

Ce fait prolog stipule que l'on étudie (dans le sens « suivre des études supérieures ») à l'université. Le dernier paramètre, `reg`, spécifie le registre courant, université ne faisant pas partie des entrées du dictionnaire. Le registre doit être indiqué à ce niveau car aucune information n'est disponible par ailleurs. Il y a sans doute conservation du registre au sein de la même famille lexicale. Mais cette hypothèse n'ayant pas été vérifiée à grande échelle, nous préférons déclarer le registre explicitement.

Les dérivés, actants sémantiques et circonstanciels ont été implémentés de manière satisfaisante et sont présentés à l'apprenant grâce à une interface que nous décrivons plus loin.

Par contre, deux autres cas plus délicats, vus plus haut, n'ont pas été traités. Il s'agit d'une part des cas où il n'y a pas de verbe dans la famille lexicale. Comme nous l'avons vu, une des solutions pourrait être de lister les dérivés en indiquant leur définition. Mais que faire dans le cas où les dérivés sont fortement polysémiques, comme industrie, nom et industriel adjectif qui possèdent chacun quatre lexies ? Les relations de dérivations et d'actance étant toujours déterminées de lexie à lexie (car les dérivés et les actants ne sont pas forcément les mêmes d'une lexie à une autre dans un même vocable[6]), doit-on faire figurer plusieurs lexies (toutes

[5] Prix est un actant moins standard que les autres et peut-être trop spécifique. Nous l'avons toutefois retenu en raison de l'importance du salaire dans le travail. En effet, le salaire est le prix du travail.

[6] Par exemple, un travailleur est le sujet du verbe travailler (premier actant) dans le sens gagner sa vie mais pas dans le sens agir pour ou contre les intérêts de qqn (le temps travaille pour moi).

celles d'industriel ayant un rapport étroit avec le sens général d'industrie), ce qui donne beaucoup trop d'information, ou bien doit-on faire simplement figurer le vocable avec, par hyperlien, un renvoi à son article ce qui oblige à des allées et venues pas toujours désirables ni nécessaires. Doit-on alors faire figurer les définitions de manière interactive, dans une fenêtre supplémentaire ? Des recherches sur ce point précis s'avèrent nécessaires pour régler le problème de manière satisfaisante.

D'autre part, il y a les cas où la famille lexicale se scinde en plusieurs sous-groupes (comme pour industrialier-industrie). Traiter ces cas alourdit la programmation qui tient déjà compte des dérivés, des actants sémantiques et circonstanciels, du schéma syntaxique du verbe (qui n'est d'ailleurs pas toujours lexicalisé, voir ci-dessus) et de leur présentation à l'écran qui est gérée à la fois par Prolog (disposition dans une liste avec mise en évidence des numéros d'actants ou des circonstants) et par HyperCard (mise en gras). Là aussi, la programmation doit être approfondie pour tenir compte de tous ces paramètres.

2.2.6 Fonctions lexicales et collocations

Les dernières relations linguistiques traitées au sein de la base de données lexicales concernent les fonctions lexicales et les collocations. Illustrées par les trois tomes du DEC (1984, 1988 et 1992), ces fonctions permettent d'exprimer des relations sémantiques générales en plus des traditionnelles relations de synonymie, d'antonymie, d'actance ou de dérivation. Une fonction lexicale **f** est une dépendance, ou une correspondance, qui associe à une lexie (l'argument de **f**) un ensemble de lexies (la valeur de **f**).

Parmi les fonctions lexicales, on distingue les fonctions standard qui peuvent s'appliquer à un nombre important de lexies. Leur domaine d'application est grand et leur sens présente un caractère général et universel, c'est-à-dire indépendant de toute langue. Par contre, les valeurs sont fonction de la langue.

On peut illustrer ces fonctions lexicales par deux d'entre elles : Magn (intensification) et Operi (verbe sémantiquement vide)

Magn(dormir) = profondément, comme un loir, à poings fermés
Oper1(plainte) = porter [~]
Oper1(cri) = pousser [un ~] (le tilde reprend l'argument).

Le DEC fait l'objet actuellement de deux projets d'informatisation (Mel'cuk *et al.*, 1995, chapitre 6 ; projet NADIA-DEC, Sérasset, 1997), même si le but dans les deux cas n'est pas le même : le premier projet tente d'informatiser les données du DEC, au prix d'une certaine simplification, de manière à les rendre disponibles pour des traitements automatiques, tandis que le deuxième vise plutôt à une informatisation du processus de production du DEC, via un éditeur spécialisé.

Dans les deux cas, les travaux se heurtent à des problèmes liés à la complexité des informations contenues dans le DEC (difficultés de formalisation).

Dans ALEXIA, nous simplifions beaucoup cet aspect en ne considérant que les fonctions lexicales et, qui plus est, les fonctions lexicales standard. Notre souci n'est pas de formaliser à tout prix les relations sémantiques entre différentes lexies mais plus de rendre compréhensible ces relations à un public non initié. C'est pourquoi nous ne retenons que les fonctions

lexicales standard, généralisables et applicables à un grand nombre d'arguments. Nous préférons exprimer les relations sémantiques des fonctions lexicales non standard par la définition de la valeur de la fonction.

De ce fait, les fonctions lexicales sont donc simplement exprimées par des faits Prolog sous forme de triplets reliant l'argument, la valeur et la fonction lexicale :

```
cooc(allocation('1'),"percevoir une ~~",sem:oper(3),freq:nc,reg:2).
```

Les deux arguments supplémentaires concernent la fréquence et le registre. Ils doivent être rajoutés pour les vocables ne faisant pas partie des entrées du dictionnaire.

D'autre part, nous ne prenons pas en compte les phénomènes de fusion et de non fusion (représenté dans le DEC par le symbole //) en présentant l'expression entière qui rend mieux compte des co-occurrences. Au lieu donc d'écrire Oper3(allocation) = percevoir, nous écrivons Oper3(allocation) = percevoir une allocation qui met en évidence la co-occurrence. Le double tilde est en effet remplacé par l'argument de la fonction lorsque la valeur est affichée dans la carte des collocations et co-occurrences.

On constate que l'implémentation des fonctions lexicales pose beaucoup moins de problème que dans le cas de la synonymie ou de l'actance, car il n'y a pas de propriétés transitives à prendre en compte. Néanmoins, la difficulté consiste à les présenter et à rendre accessible leur signification aux apprenants. Ce point comporte quelques aspects délicats, c'est pourquoi il n'a été que peu abordé. Nous proposons toutefois quelques idées de présentation (3.3.4).

3 Le dictionnaire électronique de l'environnement ALEXIA

Indépendamment du contenu lexicographique et linguistique, un dictionnaire électronique pour apprenants se doit de faciliter l'accès lexical et privilégier son ergonomie. Il s'agit en effet d'exposer les données abordées ci-dessus pour les rendre compréhensibles, donc utiles, d'une part, et facilement accessibles d'autre part. La figure 5.4 montre l'architecture du dictionnaire :

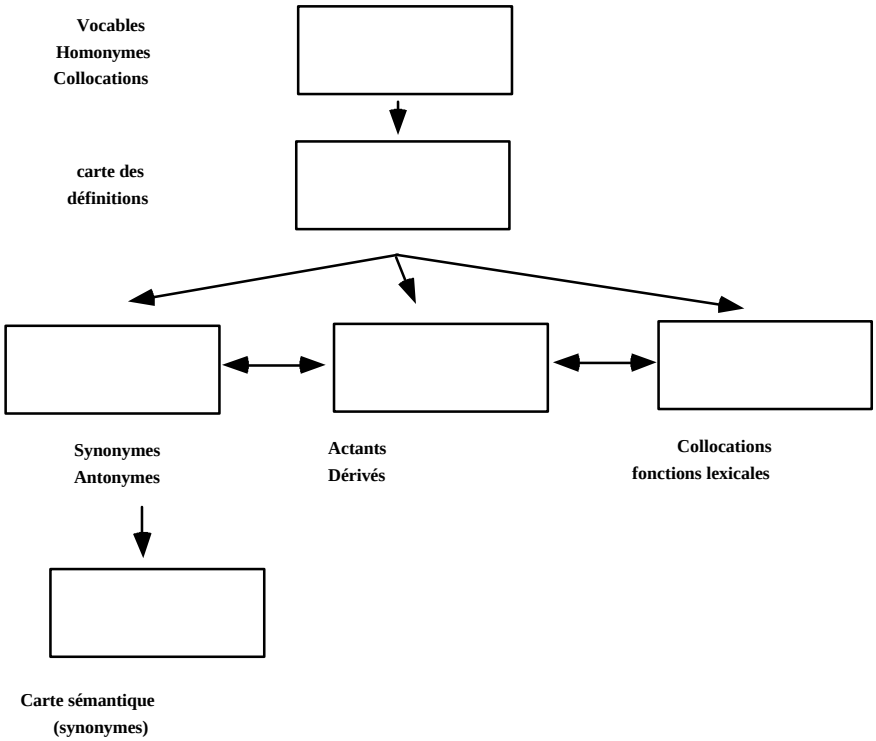


Figure 5.4 : Architecture du dictionnaire général

3.1 L'accès lexical

Nous avons vu dans le chapitre 3 que l'accès lexical était un aspect primordial de la consultation du dictionnaire. Un dictionnaire qui ne favorise pas l'accès lexical a toutes les chances de n'être pas utilisé.

L'accès lexical consiste à passer de la graphie d'un mot dans un texte à la lexie correspondant au sens qu'il exprime dans un article de dictionnaire. Cette étape essentielle doit être minimisée au possible de manière à ce que la consultation du dictionnaire soit utile et efficace. Il convient en effet que l'apprenant soit coupé de la lecture de son texte le moins longtemps possible pour qu'il n'en perde pas le fil.

L'accès lexical concerne les problèmes de flexion, d'homonymie et de collocation. Ces problèmes ont été développés dans le chapitre 3 et nous nous en tiendrons ici à ce qui a été réalisé dans ALEXIA. Pour cela, nous nous appuyons sur le modèle de la figure 5.5 qui montre que le statut du mot évolue au cours de la recherche (graphie – vocable – lexie) :

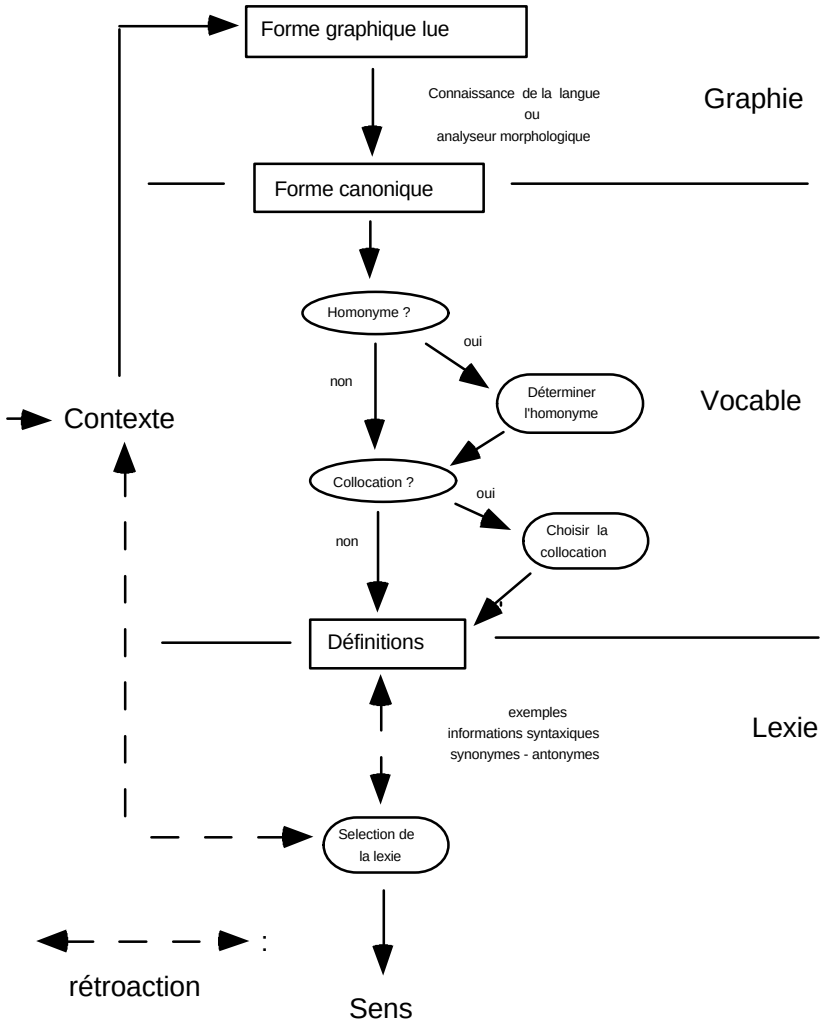


Figure 5.5 : Modèle d'accès lexical avec dictionnaire

Nous nous proposons d'étudier les étapes de ce modèle.

3.1.1 Forme canonique et flexion

Le problème de la flexion d'un vocable du texte (passage de irai à aller) est maintenant résolu de manière satisfaisante pour tout dictionnaire qui possède un analyseur morphologique capable d'indiquer la forme canonique ou qui liste l'ensemble des formes fléchies (cas du PRCD ou HOCD). Dans ALEXIA, les mots (graphies) sont listés sous leur forme canonique et sont accessibles par leurs premières lettres tapées. L'analyseur du LIPN est disponible pour déterminer la forme canonique d'un mot (figure 5.6).

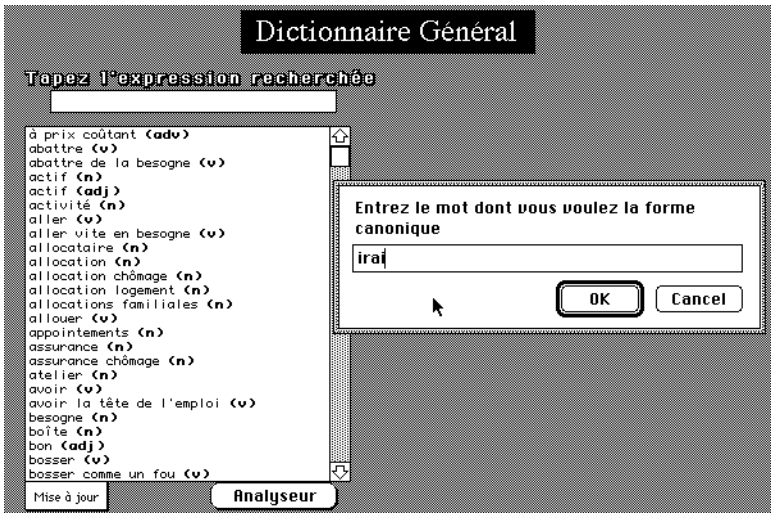


Figure 5.6 : liste des vocables et analyse morphologique

3.1.2 Homonymie

L'homonymie est l'un des problèmes importants auxquels est confronté l'apprenant dans sa recherche d'information sur une entrée. En effet, malgré une similitude de forme écrite, il est essentiel de savoir clairement à quel vocable on a affaire, de manière à avoir des informations fines et précises sur ses sens, ses constructions syntaxiques, ses synonymes, etc. L'homonymie entre vocables n'ayant pas la même catégorie grammaticale (exemple : boucher verbe et nom) ne pose pas généralement problème. Il en est autrement pour les vocables qui ont la même catégorie grammaticale. Dans ce cas, seul un critère sémantique peut aider à choisir la bonne entrée.

Dans une hypothèse synchronique, en décrivant la langue de 1999, il est important de faire comprendre à l'apprenant que les homonymes qui autrefois étaient fortement liés sémantiquement sont devenus à présent des vocables différents. Certains dictionnaires toutefois ne mettent pas l'accent sur ce phénomène et donnent la priorité à la dimension historique en montrant l'évolution des vocables à partir d'un « ancêtre commun » : PR adopte un point de vue diachronique et essaye de réduire le plus possible le nombre d'homonymes en n'opérant les dissociations que lorsque celles-ci s'avèrent inévitables (poste, emploi et poste de télévision sont dans la même entrée). Nous noterons cependant que la description diachronique intéresse plus les linguistes et les natifs que les apprenants qui sont confrontés à des problèmes plus immédiats. La même remarque s'applique aussi à COBUILD et HO, ce dernier proposant plusieurs entrées seulement pour les homonymes de différentes catégories grammaticales (poste n'a ainsi qu'une entrée). On a alors l'impression qu'il s'agit d'une large polysémie. Et dans ce cas, c'est à l'apprenant de faire le travail de dissociation. Cette étape ne fait qu'augmenter la durée de la suspension de l'activité de lecture.

Dans ALEXIA, nous attirons l'attention de l'apprenant sur ce phénomène en faisant apparaître une fenêtre intermédiaire dans le cas d'une homonymie (figure 5.7). Le vocable est alors suivi d'une précision sémantique qui permet à l'apprenant d'opérer une discrimination

(exemple : contracter : passer un accord et contracter : raidir ou encore formation : enseignement ou formation : fait d'avoir une certaine forme[7]). Cette fenêtre force l'apprenant à se décider sur un vocable particulier. Il faut toutefois ajouter que ces précisions sémantiques ne sont pas toujours aisées à déterminer : elles peuvent être trop précises, ne tenant pas compte de toutes les lexies du vocable, ou au contraire, trop vagues, pas assez discriminantes ou pas assez parlantes. Il ne faudrait pas que l'apprenant soit aiguillé (ou plutôt s'aiguille) trop souvent vers le mauvais vocable, auquel cas cette étape ne lui fait que compliquer sa tâche.

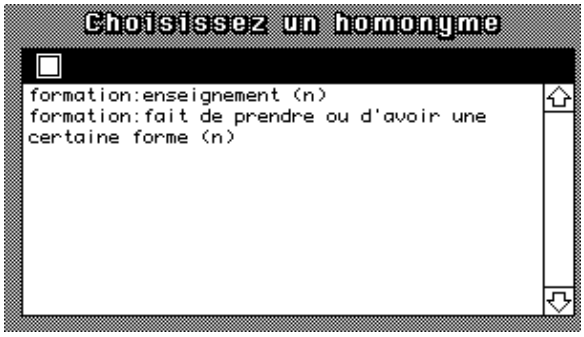


Figure 5.7 : Homonymie sur formation

3.1.3 Les collocations

Il reste un dernier problème à résoudre avant d'accéder à l'article d'une entrée : est-ce que cette entrée est isolée ou fait partie d'un groupe de mots dans lequel son sens pourrait être modifié (cas des expressions opaques) ? En d'autres termes, l'apprenant est-il en train de lire une collocation ?

Bien qu'il s'agisse d'un phénomène important (il existe une plus grande proportion de collocations que de vocables simples dans la langue), les dictionnaires classiques ne mettent pas du tout ce phénomène en évidence. Comme les collocations ne sont pas considérées comme entrées à part entière et comme elles peuvent subir des variations lexicales, elles ne sont pas citées dans l'ordre alphabétique. Il est de ce fait difficile de les localiser dans les dictionnaires (va-t-on trouver coup de barre dans l'article de coup ou dans celui de barre ?).

Dans ALEXIA, une collocation est un vocable car elle possède sa propre définition, ses exemples, sa structure syntaxique et lexicale, etc. Une partie des collocations a été extraite automatiquement du corpus et le reste complété à partir de dictionnaires ou à partir de notre intuition, le corpus n'étant pas assez important. Elles sont codées à la main, de même que leurs variations lexicales.

Le système aide l'apprenant à repérer les collocations en lui montrant (figure 5.8), dès qu'il sélectionne un mot dans un texte, toutes les expressions dont il peut faire partie (par exemple, si l'apprenant sélectionne emploi, le système lui montrera toutes les collocations qui contiennent emploi). Il peut alors décider si le problème provient ou non d'une collocation et quel est le vocable le plus judicieux à consulter.

[7] Par exemple formation nuageuse.

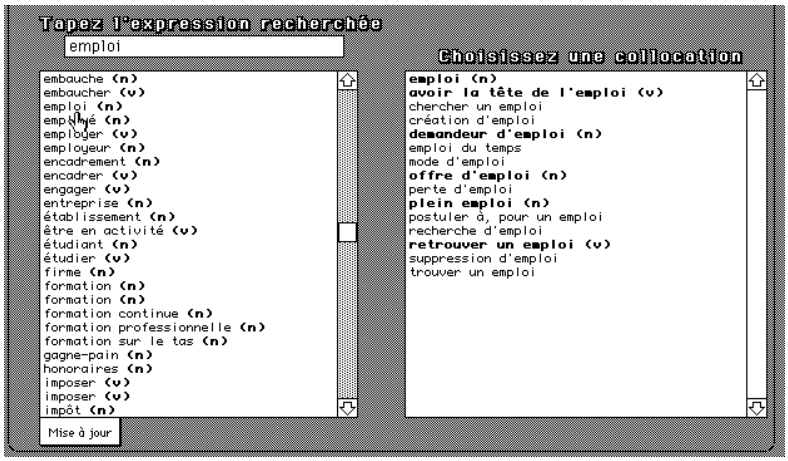


Figure 5.8 : choix d'une co-ocurrence ou d'une collocation pour la graphie emploi
(en gras les vocables décrits dans le dictionnaire)

Afin de limiter l'espace d'affichage, les collocations sont seulement listées sous leur forme principale (celle qui apparaît le plus souvent). Les variations lexicales d'une collocation font partie des informations qui doivent être décrites dans l'article qui lui est consacré. Néanmoins l'accès à une collocation peut s'effectuer à partir de chacun de ses constituants. Dans la liste, on peut par exemple accéder à l'article d'avoir la tête de l'emploi, qui a pour variante avoir la gueule de l'emploi, à partir d'avoir, de tête, de gueule ou d'emploi.

Soulignons que pour certaines entrées, le nombre de collocations peut être important. Il convient alors de les structurer, par exemple en les classifiant suivant la catégorie grammaticale des constituants (N+N, N+V, etc.). Il pourrait être aussi question d'un dispositif qui permettrait d'appeler le dictionnaire en cliquant un mot dans le texte, explorerait les contextes gauche et droit et ne présenterait que les collocations détectées. Dans la présente implémentation, ce point n'a pas été abordé et les collocations sont affichées par ordre alphabétique.

Une différence est faite à ce stade entre collocation et co-occurrence. Les premières sont en gras, ce qui indique qu'un article leur est consacré. On accède à cet article en cliquant sur la collocation. Lorsqu'on clique sur une co-occurrence, on est dirigé sur la carte des collocations, co-occurrences et fonctions lexicales (figure 5.17).

3.2 Les articles du dictionnaire

Avant d'aborder le problème des définitions et de leur compréhension par l'apprenant, il convient de se pencher sur le problème de la présentation des différentes informations dans l'article de l'entrée.

3.2.1 Présentation de l'entrée

Ces informations sont de natures très différentes : il y a bien sûr les définitions, les exemples, les informations morphologiques (genre, nombre) et grammaticales (par exemple la construction du verbe), le registre, la fréquence ainsi que l'accès aux autres vocables liés à l'entrée (synonymes, actants, collocations, etc.). Face à une telle diversité, il convient de présenter l'information de manière claire, structurée et identique pour chaque entrée, afin que l'apprenant puisse acquérir certains automatismes qui lui faciliteront l'accès visuel. Les informations doivent donc être affichées en des endroits différents suivant leur nature (figure 5.9). De plus, afin de limiter le plus possible la quantité de texte à lire et pour plus de clarté, l'information est structurée en fonction des lexies. Les exemples et les informations morphologiques et grammaticales sont affichées lorsqu'on clique sur la lexie correspondante. De même pour les synonymes, actants, collocations, etc.

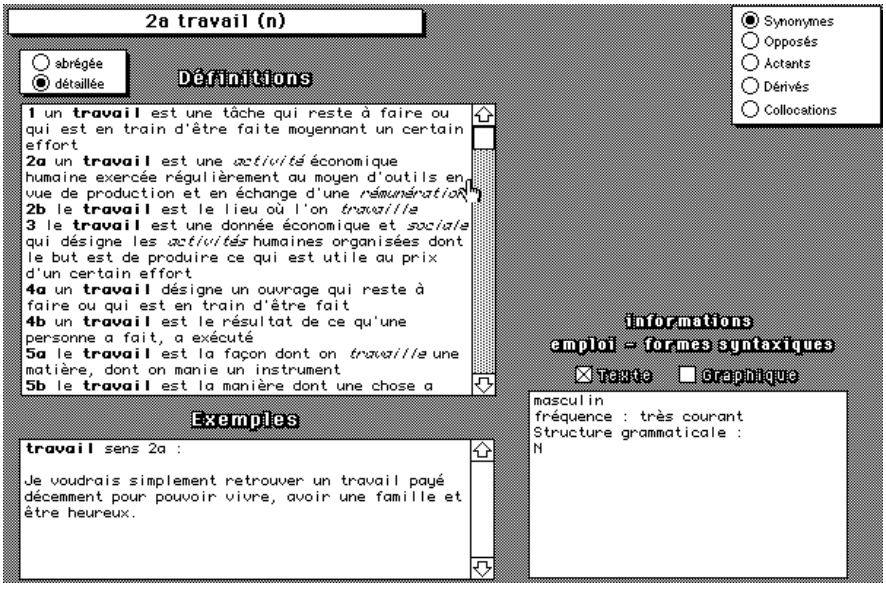
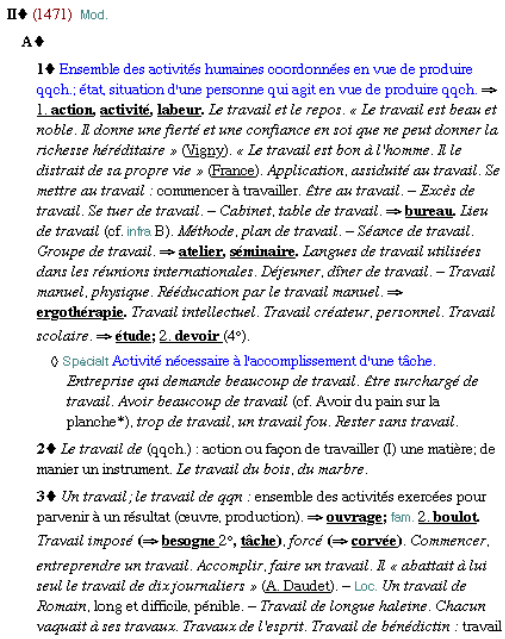


Figure 5.9 : article de travail, sens 2a

Nous évitons ainsi les traditionnels enchevêtrements dans les articles des dictionnaires papier, inévitables à cause de contrainte de place, mais beaucoup moins justifiables dans leur version électronique même si la couleur et les différents styles arrangent un peu les choses (comme le PRCD, figure 5.10).



II♦ (1471) Mod.

A♦

1♦ Ensemble des activités humaines coordonnées en vue de produire qqch.; état, situation d'une personne qui agit en vue de produire qqch. ⇒ 1. **action, activité, labeur.** *Le travail et le repos. « Le travail est beau et noble. Il donne une fierté et une confiance en soi que ne peut donner la richesse héréditaire »* (Vigny). *« Le travail est bon à l'homme. Il le distrait de sa propre vie »* (France). *Application, assiduité au travail. Se mettre au travail :* commencer à travailler. *Être au travail. – Excès de travail. Se tuer de travail. – Cabinet, table de travail.* ⇒ **bureau.** *Lieu de travail* (cf. infra B). *Méthode, plan de travail. – Séance de travail. Groupe de travail.* ⇒ **atelier, séminaire.** *Langues de travail utilisées dans les réunions internationales. Déjeuner, dîner de travail. – Travail manuel, physique. Rééducation par le travail manuel.* ⇒ **ergothérapie.** *Travail intellectuel. Travail créateur, personnel. Travail scolaire.* ⇒ **étude;** 2. **devoir** (4°).

◊ Spécialt Activité nécessaire à l'accomplissement d'une tâche. *Entreprise qui demande beaucoup de travail. Être surchargé de travail. Avoir beaucoup de travail* (cf. Avoir du pain sur la planche*), *trop de travail, un travail fou. Rester sans travail.*

2♦ *Le travail de* (qqch.) : action ou façon de travailler (I) une matière; de manier un instrument. *Le travail du bois, du marbre.*

3♦ *Un travail; le travail de qqn :* ensemble des activités exercées pour parvenir à un résultat (œuvre, production). ⇒ **ouvrage;** fam. 2. **boulot.** *Travail imposé* (⇒ **besogne** 2°, **tâche**), *forcé* (⇒ **corvée**). *Commencer, entreprendre un travail. Accomplir, faire un travail. Il « abattait à lui seul le travail de dix journaliers »* (A. Daudet). – Loc. *Un travail de Romain,* long et difficile, pénible. – *Travail de longue haleine. Chacun vaquait à ses travaux. Travaux de l'esprit. Travail de bénédictin :* travail

Figure 5.10 : extrait de l'entrée travail du PRCD

Revenons maintenant aux définitions. Toutes les étapes préliminaires précédentes d'accès lexical ont tenté d'aider l'apprenant à bien choisir le vocable qui lui pose problème et à chercher son sens directement dans l'article le plus approprié. Maintenant, se pose le problème crucial de la compréhension du vocable et de la confrontation aux définitions. En effet, une des principales difficultés est de bien déterminer l'acception qui vient d'être lue et de comprendre son sens. C'est, rappelons-le, la cause de consultation la plus fréquente. De plus, elle conditionne les informations propres aux lexies. Il convient donc de leur consacrer une attention particulière.

3.2.2 Les définitions

Comme le constate Bogaards (1995) en parcourant l'article de just dans différents dictionnaires anglais pour apprenants, on voit facilement que les vocables ayant beaucoup de lexies peuvent représenter des obstacles décourageants. Pour résoudre (une partie de) ce problème, il n'est nullement question de réduire le nombre de lexies, la version électronique doit contenir exactement les mêmes informations. Par contre, il est possible de faire jouer la souplesse d'utilisation de l'informatique pour présenter les choses différemment.

L'idée développée dans ALEXIA et dans d'autres dictionnaires électroniques (RE, PRCD et le GLCD) est de présenter en premier lieu des définitions abrégées (ou un plan de l'article). Celles-ci ont pour but uniquement de repérer rapidement la lexie qu'on pense être la bonne d'après le contexte. Pourtant, comme il n'est pas toujours évident de comprendre une acception particulière seulement d'après une mini-définition (Guillot & Kenning, 1994a), il est possible dans ALEXIA, en cliquant sur chaque lexie, d'avoir des exemples en contexte et des constructions syntaxiques qui peuvent aider au choix. Comme nous l'avons vu plus haut, ceci n'est pas toujours le cas dans les autres dictionnaires électroniques.

Les définitions abrégées remédient en partie au défaut que l'on peut reprocher aux dictionnaires électroniques : la faible quantité (par rapport aux versions papier) des informations affichées simultanément. En effet, on ne peut pas avoir accès, si l'article du vocable est long, à l'ensemble des informations afin d'en avoir une vue globale, si besoin est. L'article est fragmenté et cela implique de changer fréquemment ce qui est affiché à l'écran. Cependant, d'après Guillot & Kenning (1994a), les apprenants disent ne pas être trop dérangés par cette fragmentation.

Si la définition abrégée n'est pas suffisante pour bien comprendre le sens, on peut alors avoir accès aux définitions complètes en cliquant sur le bouton « détaillée ». On a alors sous les yeux un article classique d'un dictionnaire, à ceci près qu'il y a, dans la même fenêtre et au même endroit, seulement des définitions. Cependant, il n'est pas toujours utile d'avoir sous les yeux toutes les définitions longues. C'est pourquoi il est possible de visualiser une seule définition longue (celle qu'on juge pertinente) indépendamment des autres en cliquant sur le numéro de lexie précédant la mini-définition (figure 5.11).

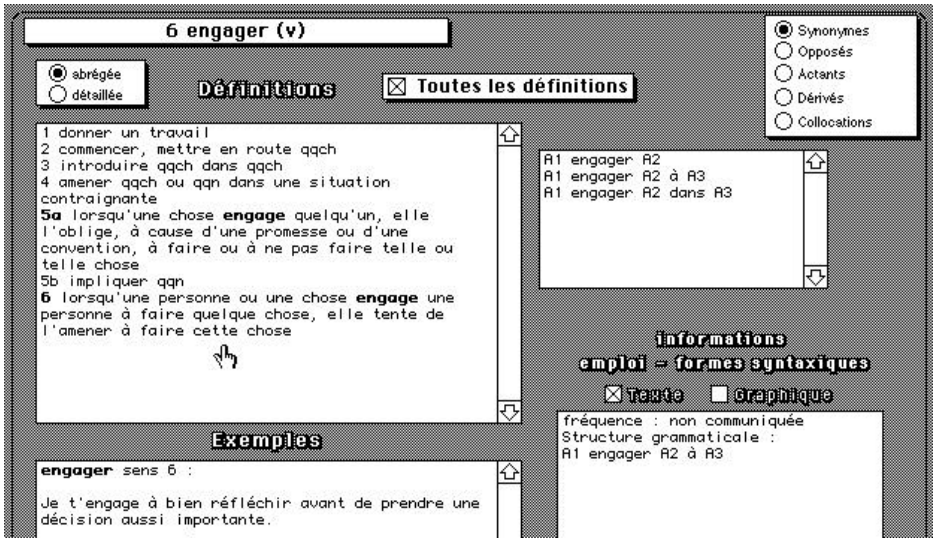


Figure 5.11 : définition longue (sens 5a et 6 d'engager)

On constate donc que la présentation de la définition longue d'une lexie indépendamment des autres autorise une vision synthétique de l'ensemble des significations du vocable et permet ainsi de choisir la lexie qui est la plus appropriée en évitant l'écueil de la lecture de l'article entier.

Il peut parfois être utile de faire fonctionner des filtres qui éliminent les définitions ne correspondant pas manifestement avec le sens que l'on recherche. Dans le cas des verbes, on peut appliquer des filtres syntaxiques : les seules définitions apparaissant à l'écran sont celles où le verbe se construit de la même manière que son occurrence dans le texte. Dans l'exemple de la figure 5.12, si l'on recherche le sens de diriger dans la phrase « cet article était vraisemblablement dirigé contre lui », seule la construction du verbe avec la préposition contre nous intéresse et on peut éliminer les définitions des autres constructions de diriger (forme transitive, construction avec sur et vers).

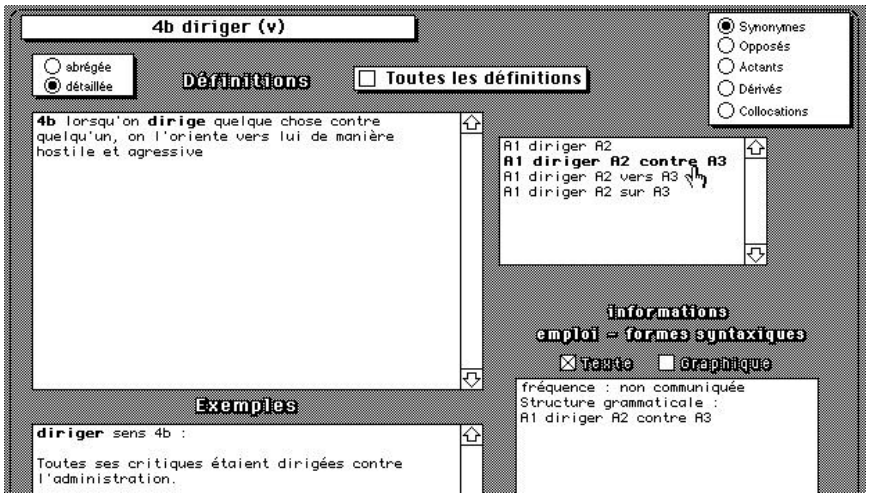


Figure 5.12 : présentation sélective des sens de diriger suivant son schéma syntaxique

Naturellement, ce filtre n'est valable que pour les verbes et implique, de la part de l'utilisateur, une certaine compétence linguistique : il doit bien faire la différence entre une préposition faisant partie effectivement de la construction verbale et une préposition faisant partie d'un complément circonstanciel (la phrase « Je travaille sur la plate-forme » indique-t-elle que la personne est physiquement sur la plate-forme ou que celle-ci est l'objet de son travail ?)

Le dernier obstacle dans la consultation d'un dictionnaire consiste en la compréhension de la définition elle-même, et donc du vocable sur lequel l'apprenant s'est arrêté. Ce problème n'a pas à être résolu par l'emploi des mini-définitions. Celles-ci ne servent qu'à éliminer les acceptions trop éloignées du contexte de la lecture. Dans certains cas, elles peuvent s'avérer suffisantes mais la plupart du temps, l'apprenant devra parcourir les définitions complètes.

Dans son étude, Bogaards (1995) relève que la compréhension des définitions est une difficulté majeure, en particulier parce que l'apprenant ne comprend pas toujours les vocables qui la composent.

Il est maintenant un fait acquis que dans le cadre d'un dictionnaire pour apprenant, les définitions doivent être rédigées avec un vocabulaire contrôlé. Cela varie, suivant les dictionnaires entre les 2 000 et 3 500 vocables les plus fréquents de la langue. Cependant, cette contrainte n'est pas toujours suffisante, soit parce que le niveau de l'apprenant n'est pas assez élevé, soit parce que les dictionnaires ne la respectent pas toujours (Bogaards, 1996). Par exemple, si l'on cherche la définition de to crinkle dans COBUILD (1994), on trouvera : « if something crinkles or if you crinkle it, it becomes slightly creased or folded ». Même si folded donne une petite indication de sens, il est difficile de comprendre le sens de to crinkle sans comprendre celui de creased. Tous deux étant rangés dans le même ordre de fréquence (1 diamant), il n'est pas évident que l'apprenant connaisse la signification de creased.

L'avantage, bien utilisé d'ailleurs, des dictionnaires électroniques sur leur version papier, est de pouvoir passer facilement et rapidement d'une définition à une autre. Là aussi (Guillot & Kenning, 1994b), les utilisateurs apprécient grandement cette facilité de consultation « horizontale », d'un vocable à un autre. On compte d'ailleurs beaucoup trop sur cette

caractéristique pour pallier l'insuffisance de qualité des définitions. Cet avantage, incontestable, n'est, par contre, pas totalement et pas judicieusement exploité dans la plupart des dictionnaires électroniques actuels. En effet, lorsqu'on passe d'un vocable à un autre, comme il n'y a toujours qu'une seule fenêtre activée, on perd visuellement la trace du premier vocable. Dès lors, il est plus difficile de comparer et d'essayer de comprendre. Par exemple, dans la figure 5.13, si l'apprenant a des doutes sur le vocable métier, il est dommage de perdre visuellement lorsqu'il consultera métier la définition du sens 2 de travailler. D'autre part, il serait particulièrement intéressant d'avoir seulement le sens du vocable qu'on ne comprend pas dans la première définition. Ceci est tout à fait possible vu que les définitions sont fixées. Malheureusement, ce principe n'est jamais appliqué pour d'évidentes raisons pratiques.

Dans ALEXIA, on tient compte de ces deux problèmes en présentant, lorsqu'on clique sur un vocable, la lexie particulière dans une petite fenêtre à côté de la première définition (figure 5.13).

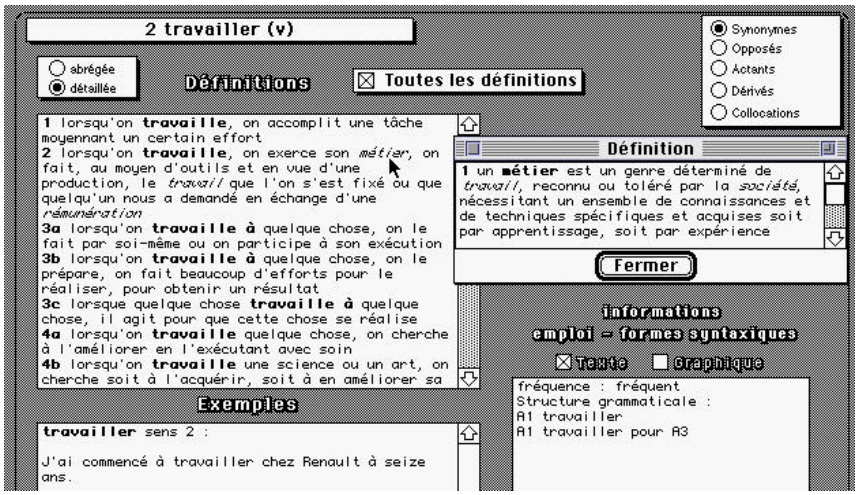


Figure 5.13 : définition d'une lexie extraite d'une précédente définition

Cette fonctionnalité est possible dans ALEXIA, vu le faible nombre d'entrées. Cela implique aussi que chaque vocable d'une définition ne peut être défini dans une petite fenêtre. C'est pourquoi seuls les vocables en italique, étant eux-mêmes une entrée, présentent cette caractéristique.

Néanmoins, même si le liage explicite de lexie à lexie demande une grosse quantité de travail, l'extension à un dictionnaire couvrant la langue entière ne nous paraît pas insurmontable. En effet, il n'est pas nécessaire de déterminer le sens de tous les vocables de la définition, mais seulement celui des vocables les plus pertinents ou les plus difficiles. C'est cette dernière solution qui a été adoptée par LDOCE, ce qui montre que l'entreprise est tout à fait réalisable pour peu que l'on y mette les moyens.

3.2.3 Les autres informations lexicales

Les exemples

Les exemples ont été tirés du corpus et adaptés dans certains cas pour rendre la lecture plus facile. Pour les lexies non représentées dans le corpus, les exemples ont été inventés. Ils sont présentés par phrase entière, une seule phrase étant disponible par lexie. Ils jouent plutôt un rôle illustratif qu'explicatif, même si les exemples peu parlants, comme les exemples trop courts ou possédant un contexte trop pauvre, ont été écartés.

La fréquence

Pour le calcul des fréquences, nous nous sommes servi de la liste de Beaudot (1992), celle-ci étant la plus récente pour le français, bien qu'elle ait été constituée à partir de textes écrits dans les années 60.

L'échelle reprend celle établie dans COBUILD. Elle possède cinq rang de fréquence. Le premier, intitulé « très courant » contient environ 700 vocables (parmi lesquels travail ou activité), le deuxième, « courant », contient environ 1 200 vocables (boîte, cadre), le troisième, « assez courant », 1 500 vocables (dirigeant, employeur), le quatrième, « relativement peu courant », 3 200 vocables (industrialiser, recrutement, rémunération) et le dernier, « peu courant », contient 8 100 vocables (honoraires, encadrement, allouer). Au-delà, les vocables sont considérés comme rares.

La liste et les résultats en termes de classement doivent cependant être relativisés. D'une part, elle porte sur des textes relativement anciens, qui ne reflètent pas toujours le vocabulaire actuel (par exemple, intérim ou intérimaire sont classés comme « peu courant ». D'autre part, il s'agit de textes généraux, ce qui explique que des vocables comme professionnel (nom), industriel (nom) ou qualification soient considérés comme « relativement peu courant ». Enfin, les différents registres sont inégalement représentés puisque les vocables familiers (comme boulot ou job) sont « peu courant » (bosseur, aussi bien nom qu'adjectif, ne fait pas partie de la liste).

Cette particularité est due au manque de listes de fréquence récentes sur le français. Les dictionnaires français ne les utilisent pas, contrairement au COBUILD ou au LDOCE, ce dernier faisant même la distinction pour certains vocables entre langue orale et langue écrite.

Nous avions pensé dans un premier temps établir une liste de fréquence des vocables du corpus, mais ce dernier n'étant pas assez important et pas assez significatif, les résultats n'auraient été valables que pour les vocables les plus fréquents.

Signalons enfin, même si elles sont peu nombreuses, que certaines collocations ont leur rang de fréquence. Celui-ci est généralement faible, et de fait peu significatif.

Le genre et le nombre

Ces informations concernent principalement les noms et adjectifs. Il faut tenir compte des noms et adjectifs qui sont uniquement au singulier ou uniquement au pluriel, ou qui n'ont pas de genre, etc. Pour cela, une série de dix codes parcourt toutes les possibilités (voir annexe B).

Il est d'usage dans les dictionnaires d'indiquer des modèles de conjugaison des verbes. Cette information n'est pas présente dans ALEXIA car, d'une part, nous n'avons pas l'ensemble de ces modèles, et d'autre part, il n'y a pas d'outil de conjugaison pour pouvoir les

illustrer. Nous n'avons pas accordé la priorité à ce module, car ce n'est pas l'objet de recherche de ce travail. L'implémentation ne poserait pas de problème *a priori*.

Les registres

Ils sont au nombre de quatre : soutenu, courant, familier et grossier. Ces distinctions sont assez classiques. Il conviendrait d'en rajouter de plus fines, à l'instar par exemple du COBUILD qui propose des registres (ou niveau de langue) tels que littéraire, technique, oral, écrit, injurieux, démodé, etc.

Les informations grammaticales

Elles concernent les verbes et montrent la façon dont ceux-ci sont construits et s'emploient (figure 5.13, en bas à droite). La description est faite de manière explicite en listant les différentes constructions possibles (transitivité, intransitivité, prépositions nécessaires). Il s'agit ici de décrire ces constructions de manière plus formalisée et claire visuellement, tout en évitant des codes grammaticaux dont on sait qu'ils font plaisir aux grammairiens et lexicographes (comme en atteste leur présence importante dans les dictionnaires) sans toutefois être vraiment consultés par les apprenants (Harvey et Yuill, 1997).

Les schémas syntaxiques, réutilisés dans la présentation des actants et dérivés syntaxiques (voir plus bas), en replaçant le verbe en contexte avec les variables qui désignent le sujet et les compléments, complètent la définition de la lexie en montrant certains constituants optionnels du verbe. Par exemple, on peut voir dans la figure 5.13 que la définition de la lexie travailler sens 2a ne mentionne que l'emploi intransitif du verbe. Pourtant la construction transitive indirecte travailler pour qqn est possible. Le groupe prépositionnel introduit par pour étant optionnel, il est difficile de le mentionner dans la définition sans alourdir celle-ci. La seule solution reste donc de l'illustrer par les schémas syntaxiques.

Les informations grammaticales pourraient aussi concerner d'autres catégories grammaticales comme les noms et les adjectifs. Par exemple, la construction N de N est fréquente, un apprenant du français se demande souvent si tel ou tel adjectif est postposé ou non, il y a des contraintes de gradation sur certains adjectifs, etc. Ces informations n'ont pas été développées dans ALEXIA, mais il est clair qu'elles devraient figurer dans tout dictionnaire d'apprentissage du français.

Les variations lexicales

Elles concernent principalement les expressions semi-figées. Par exemple, mettre au chômage peut se dire mettre à la porte ou mettre dehors ; avoir la tête de l'emploi peut se dire, dans un autre registre, avoir la gueule de l'emploi. Même s'il s'agit de variantes, les apprenants peuvent parfois les percevoir comme expressions différentes et synonymes. Il convient donc d'en tenir compte dans les relations de synonymie.

Ces variations ont été codées dans la base de données mais elles ne sont pas présentées dans les cartes du dictionnaire. Ce point est aussi à développer.

3.2.4 Les synonymes

Les problèmes auxquels sont confrontés les dictionnaires papier (et leurs équivalents électroniques) concernent tout d'abord l'accès aux synonymes eux-mêmes et ensuite les moyens mis à la disposition de l'apprenant pour saisir les nuances entre les sens exprimés par les différents vocables.

L'accès aux synonymes

Faute de regroupement à un même endroit, les renvois vers les synonymes sont éparpillés dans les articles des dictionnaires papier et il est ainsi plutôt fastidieux de devoir en consulter plusieurs entrées pour discerner tous les différents sens. Il faut alors se tourner vers des dictionnaires spécialisés (Robert des synonymes (Bertaud du Chazaud, 1995), Cambridge Word Routes (McCarthy, 1994), Bénac (1982), Lecointe (1993) ou Macé et Guinard (1990)) ou des thésaurus, qui regroupent les synonymes sous l'entrée centrale au concept et lient les autres vocables vers cet endroit.

Dans ALEXIA, l'ensemble des synonymes est calculé pour chaque vocable en parcourant le réseau (voir ci-dessus) et le système affiche l'ensemble du réseau automatiquement (figure 5.14).

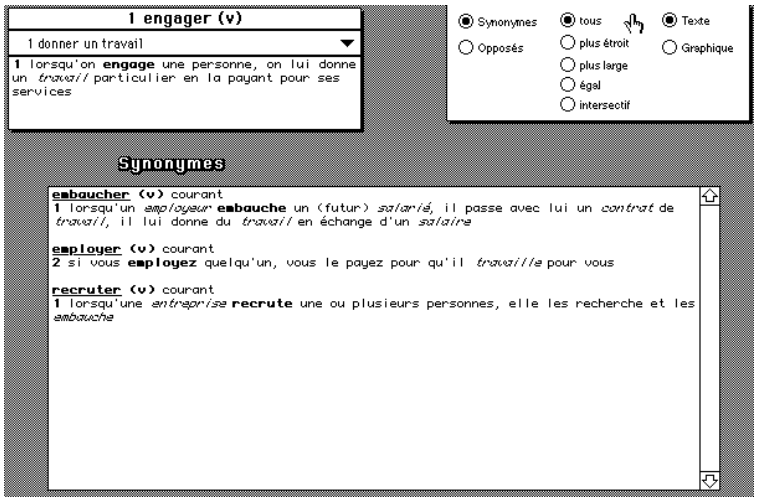


Figure 5.14 : carte de présentation des synonymes d'engager sens 1

Les sens

Bien souvent, toujours à cause du manque de place, le dictionnaire des synonymes papier ne propose pas de définition. Il faut donc se référer à un deuxième dictionnaire. C'est la politique déclarée du Robert qui ne fait que lister les vocables, sans distinction de lexie, en invitant le lecteur à consulter le PR. L'exhaustivité y est certes, mais à quel prix sur le plan de la consultation !

D'un autre côté, on trouve des dictionnaires qui, en regroupant les synonymes, indiquent ce qui les différencie (figure 5.15). Ce sont des informations très utiles qui permettent souvent de déterminer le vocable adéquat.

chef, personne placée à la tête d'un groupe : **dirigeant** (⇧ entreprise, association : *les dirigeants du parti communiste*) ; **leader** (⇧ parti ou groupe d'opinion : *leader de l'opposition*) ; **meneur** (id. ; ⇧ péjor, agitation : *l'on arrêta tous les meneurs de la grève*) ; **patron** (⇧ entreprise, ou fam. : *le patron envisage des licenciements*) ; **employeur** (⇧ offrant du travail) ; ***directeur** (⇧ fonction précise à la tête d'un secteur, d'une entreprise ou d'une administration : *directeur d'école*) ; **responsable** (⇧ insiste sur la charge de responsabilités, plutôt que sur la subordination : *le responsable de l'embauche* ; **animateur** (⇧ insiste sur activité dans les échanges au sein du groupe, idées, etc. : *l'animateur du club de bridge*) ; **commandant** (⇧ militaire) ; v. aussi **patron**. ≈ v. diriger.

Figure 5.15 : synonymes de chef dans Lecointe (1993)

Cependant, dans les deux cas, ces dictionnaires s'adressent à des natifs. Ils présupposent que le lecteur connaît déjà les vocables. De ce fait, il nous semble nécessaire, lorsqu'on s'adresse à un apprenant d'une langue étrangère, d'expliquer les vocables en question en donnant une définition, aussi précise que possible (figure 5.14). L'idéal serait d'ajouter explicitement les différences, mais il n'est pas possible de les calculer automatiquement. C'est pourquoi le système affiche uniquement les synonymes et leur définition. Ces dernières, si elles sont suffisamment précises, doivent permettre de comprendre les nuances. Mel'cuk (1995) préconise d'ailleurs de décrire les entrées en confrontant directement l'ensemble des synonymes délimitant le concept pour arriver au maximum de cohérence et de précision.

Signalons enfin dans la figure 5.14, que la synonymie est établie de lexie à lexie. Ainsi, lorsqu'on fait varier la lexie de départ à l'intérieur du même vocable, via le menu en haut à gauche de la fenêtre, le système recalcule les synonymes.

Les différentes synonymies

Un ensemble de boutons radio permettent de faire varier l'affichage en fonction du type de synonymie souhaitée. Ainsi, l'apprenant a la possibilité de lire uniquement les vocables de sens plus étroit, ou bien plus général. Il a néanmoins la possibilité de faire afficher tous les synonymes.

3.2.5 Les actants et dérivés syntaxiques

Comme expliqué plus haut, nous nous appuyons sur le schéma syntaxique du verbe de la famille lexicale, si elle en possède un, pour spécifier les dérivés et actants. Ils sont assignés à des variables qui décrivent les arguments du verbe (figure 5.16). Les actants circonstanciels sont mentionnés plus bas.

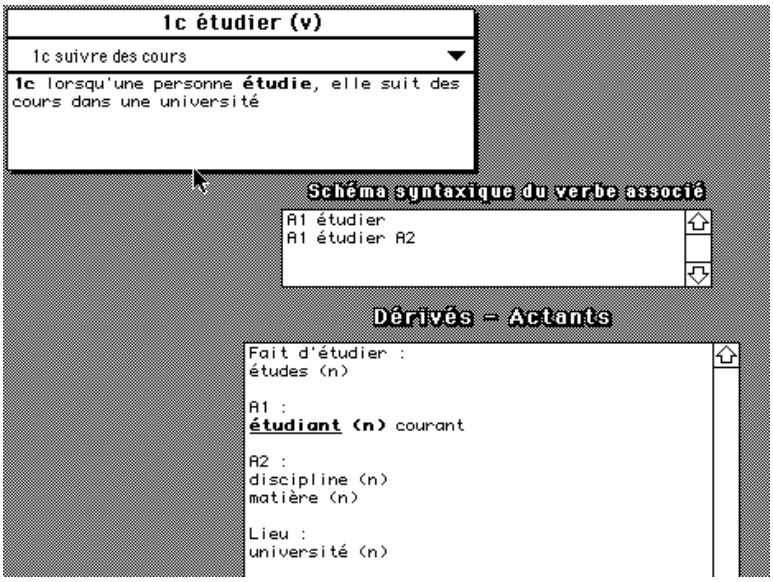


Figure 5.16 : carte de présentation des actants d'étudier sens 1c

Si la famille lexicale ne contient pas de verbe, les actants et dérivés sont alors simplement listés avec leur définition, même si cela peut parfois poser problème (voir plus haut).

3.2.6 Les collocations

La fenêtre concernant les collocations n'a pas été suffisamment traitée, notamment la présentation des informations. Ainsi, pour l'instant, sont affichées dans un ordre arbitraire, l'ordre alphabétique, les collocations (et les co-occurrences) et la fonction lexicale qui les relie à la lexie de base, sans aucun traitement, directement de la base de données lexicales (figure 5.17).

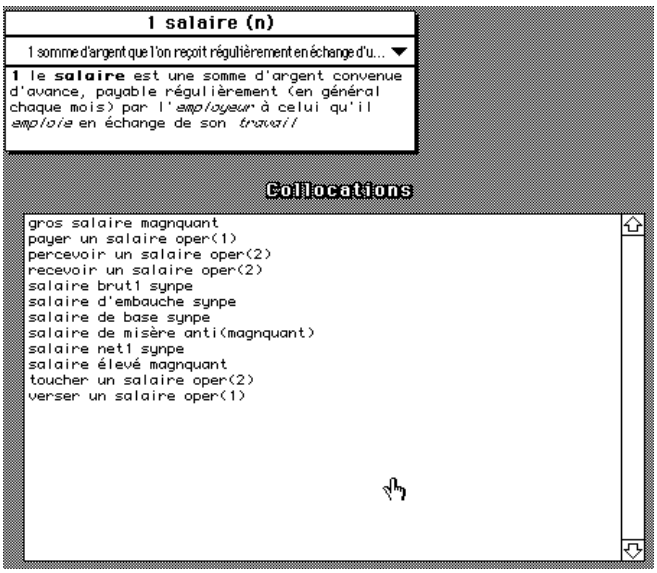


Figure 5.17 : carte de présentation des collocations de salaire sens 1

La principale difficulté est de présenter les informations, notamment les fonctions lexicales, de manière simple et compréhensible par l'apprenant. L'idée que nous comptons adopter est de faire figurer sur la carte une série de boutons, chacun contenant une fonction lexicale, parmi celles répertoriées dans la base de données lexicales. Chaque fonction sera commentée et exemplifiée dans l'aide. En cliquant sur ces boutons, l'utilisateur appliquera en fait la fonction lexicale à la lexie et pourra voir directement le résultat.

Dans le cas de fonctions lexicales standard (ex : Magn, Oper, etc.), le sémantisme expliqué dans l'aide devrait suffire, les fonctions étant suffisamment générales et simples pour être comprises sans trop de problème, bien qu'il y aurait ici objet à mesure. Par contre, concernant les fonctions non standard, plutôt que de rentrer dans une formalisation abstraite, compliquée et peu généralisable, le plus simple consiste à expliquer le lien entre la lexie et la collocation par une définition.

4 Conclusion

Après avoir défini l'unité sémantique de base, la lexie, et précisé les différentes informations utiles aux apprenants, nous avons montré la manière dont nous avons structuré la base de données lexicales du dictionnaire de l'environnement. Prolog s'est avéré un langage efficace pour traiter la structure en réseau, notamment pour les synonymes et les actants.

Dans une deuxième partie, nous avons vu en quoi le dictionnaire d'ALEXIA peut faciliter la consultation et la recherche d'information. L'environnement aide l'apprenant dans le choix du vocable en attirant son attention sur des étapes que l'on néglige trop souvent comme l'homonymie ou les collocations. Il présente les informations de manière sélective par souci d'efficacité et tente d'aider l'apprenant dans la délicate tâche de compréhension des définitions en limitant le plus possible les boucles auxquelles peuvent amener les lectures d'articles, cette opération étant source d'une grande perte de temps et de compréhension. Enfin, il aide l'apprenant à maîtriser les phénomènes d'actance-dérivation et de synonymie, en montrant dans un format approprié, le maximum d'informations possible. Nous allons voir maintenant un autre mode de visualisation des synonymes : la visualisation graphique de réseaux lexicaux.

CHAPITRE 6

Affichage graphique de réseaux lexicaux

Après avoir étudié la manière dont les informations lexicales étaient présentées sous forme textuelle dans le dictionnaire d'ALEXIA, nous présentons dans ce chapitre la partie qui traite de l'affichage graphique de réseaux de synonymes. En effet, dans la carte des synonymes (figure 5.14), la présentation des différents synonymes d'une lexie peut se faire visuellement sous forme d'un graphe généré automatiquement par le système. Nous exposons ici les propriétés de ces graphes et les problèmes rencontrés lors de la génération.

Auparavant, nous étudions différents travaux relatant l'influence du multimédia sur l'apprentissage lexical ainsi que les diverses représentations visuelles existantes de réseaux lexicaux. Des expérimentations aux conclusions contestables nous amènent à définir des critères d'évaluation de qualité de représentations visuelles.

Ce chapitre a été adapté de Chanier et Selva (1998).

1 Influence du multimédia sur l'apprentissage lexical

La théorie du « codage double de l'information » (Chun et Plass, 1997) postule l'existence de deux systèmes de stockage de l'information, l'un verbal et l'autre non verbal. Appliqué au multimédia, ce double système de représentations verbales et visuelles s'avère avantageux en permettant un plus grand stockage de l'information dans les deux systèmes, un codage plus élaboré ainsi qu'une plus grande variété de chemins d'accès. Enfin, il donne l'occasion à l'apprenant de stocker les informations dans le système le plus adéquat en fonction de ses caractéristiques personnelles.

Dans la pratique, la présentation simultanée de textes, d'images et de sons s'est avérée efficace pour l'acquisition. Citons par exemple les travaux de Liu et Reed (1995) qui ont conçu un système hypermédia d'aide à l'apprentissage de vocabulaire autour du video-disque *Citizen Kane*. Tous les différents médias (texte, audio, vidéo) ont soigneusement été intégrés, des aides contextuelles lexicales (parmi lesquelles des réseaux sémantiques et des informations linguistiques de différentes natures) ainsi que des schémas et des outils

d'indexation étaient fournis pour aider l'utilisateur à se déplacer dans les diverses parties du système contenant différents types d'information. Enfin, les utilisateurs avaient la possibilité de s'exercer en effectuant plusieurs activités éducatives.

Expérimenté pendant dix heures sur une période de cinq semaines sur un nombre statistiquement significatif d'apprenants (N = 63), les tests ont permis de conclure que le système était efficace pour l'apprentissage de vocabulaire (plus de 80 vocables). Cependant, ce n'était pas l'objectif principal pour les auteurs qui cherchaient plutôt à étudier les différents profils d'apprentissage (dépendant ou indépendant du contenu). Les résultats ont montré qu'étant donné qu'il n'y avait pas de différence statistiquement significative dans la réussite des épreuves suivant les groupes, les apprenants avaient utilisé de manières très diverses les ressources fournies par le système (Liu et Reed, 1994).

Néanmoins, d'autres expérimentations ne débouchent pas sur les mêmes conclusions. Voici par exemple deux exemples récents, qui ont d'ailleurs été cités à plusieurs reprises dans d'autres travaux.

Tripp et Roby (1994) ont cherché à mesurer les effets des indices visuels et auditifs (métaphores graphiques, ressources audio, graphiques illustratifs, etc.) pour la remémoration de vocabulaire dans un environnement hypermédia. Leurs résultats montrent que le facteur auditif (qui consiste en un vocable prononcé lorsqu'une carte est ouverte) ne contribue pas à l'apprentissage. Ils en concluent donc que l'audio n'est pas nécessaire dans de telles situations d'apprentissage. Cependant, il faut préciser que les 120 vocables étaient présentés seuls, hors contexte, accompagnés uniquement de traductions. De plus, les auteurs n'ont pas pris la peine de contrôler les facteurs personnels.

Une autre expérimentation (Svenconis 1994) visait à évaluer l'influence respective ou combinée sur l'apprentissage de la présentation lexicale de vocables sous forme de listes, de cartes sémantiques et avec ou sans son. Trois variables indépendantes étaient mesurées : la Structure Sémantique, la Méthode de Présentation et la Composante Son. La Structure Sémantique, la manière dont trois groupes de 24 vocables étaient reliés sémantiquement, était divisée en Étroitement Liés (pouvant être définis en d'autres termes dans le même groupe), Modérément Liés et Faiblement Liés (ne pouvant pas être définis en d'autres termes dans le même groupe). La Méthode de Présentation, la manière dont on présentait les vocables aux sujets, était divisée en Carte Sémantique et en Liste de Mots. Sous forme de liste, chaque groupe de 24 vocables était présenté alphabétiquement. Sous forme de carte sémantique, les vocables étaient affichés avec les relations sémantiques entre eux. La carte sémantique pour le groupe Étroitement Liés représentait le graphe d'une famille, où les nœuds étaient les membres et les arcs les relations familiales (figure 6.1). Celle pour le groupe Faiblement Liés représentait un autre regroupement de vocables décrivant une excursion familiale. Les groupes de vocables étaient disposés autour d'un événement et des liens entre les groupes indiquaient l'ordre des évènements.

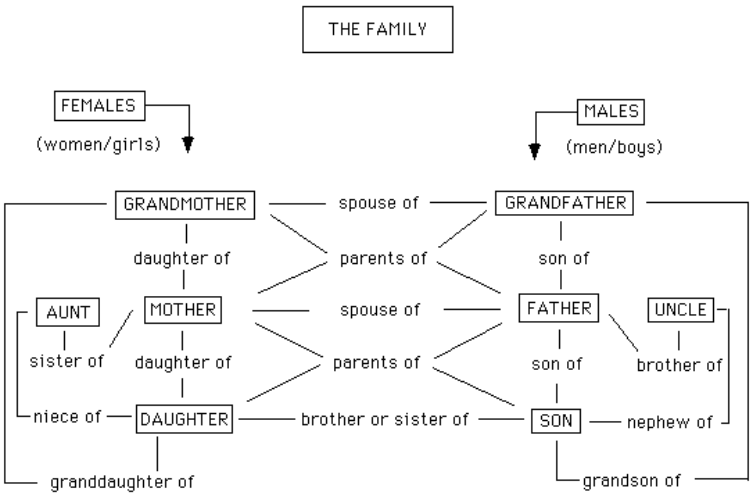


Figure 6.1 : carte sémantique des vocables Étroitement Liés dans le système de Svenconis

Les sujets (12 par groupe) étaient des grands débutants en langue seconde et n'avait jamais utilisé l'ordinateur. Après 15 minutes de cours sur la souris, ils avaient 22 minutes pour « apprendre » 24 vocables. Les résultats des tests ne révélèrent pas de différence significative entre les différentes méthodes. Les meilleurs scores furent ceux de la méthode CarteSémantiqueAvecSon, suivie de près par ListedeMotsSansSon ! Une conclusion rapide (que l'auteur n'a pas faite) pourrait être que l'assemblage de toutes sortes de ressources donne les meilleurs résultats, mais que ce n'est pas peine de se fatiguer à les assembler vu que la classique liste de vocables est tout aussi efficace !

Il y a plusieurs problèmes avec cette expérimentation : premièrement les détails de la procédure (type des apprenants et des matériaux d'apprentissage) montrent qu'elle est pédagogiquement inappropriée ce qui implique que, comme dans l'expérimentation précédente, l'apprentissage ne peut pas vraiment être étudié ; deuxièmement, les facteurs individuels n'étaient pas considérés, empêchant ainsi, même dans de meilleures conditions, une évaluation de l'efficacité relative des médias pour certains apprenants ; troisièmement, trop de facteurs étaient mesurés simultanément. Par exemple, pour répondre à la question intéressante initiallement posée par l'auteur sur le type de carte sémantique qui favorise l'apprentissage, il aurait été plus approprié de monter une expérimentation seulement sur la variable CarteSémantique.

Ces deux exemples s'avèrent donc quelque peu contestables. En effet, si les gains en apprentissage lexical n'ont pas pu être observés, ce n'est pas à cause de la soi-disant inefficacité des ressources multimédias, mais bien parce que les principes pédagogiques et linguistiques en jeu n'étaient pas appropriés pour l'apprentissage. Ceci nous mène à étudier davantage les représentations qui sont proposées à l'apprenant et à dégager des critères pour évaluer leur qualité. Nous nous pencherons plus précisément sur les représentations des informations visuelles utilisées lors de tâches d'apprentissage lexical par ordinateur telles que la compréhension du sens d'un vocable, le choix du vocable adéquat dans un exercice de production ou encore la mémorisation d'un vocable parmi d'autres déjà connus.

2 Critères d’évaluation de la qualité d’une représentation visuelle

Pour toutes ces tâches, le mouvement qu’implique la vidéo n’est que de peu d’utilité puisque l’information visuelle est essentiellement statique, devant mettre en évidence d’un coup d’œil les liens reliant différents vocables. C’est pourquoi nous étudierons les images, une image pouvant être une photo, un dessin ou un diagramme.

Ces représentations doivent être :

- informatives en veillant notamment à ce qu'une trop grande quantité d'information ne nuise pas à la lecture et à la compréhension.
- computationnelles, elles doivent pouvoir se prêter à des traitements informatiques pour rendre différentes opérations (recherche, reconnaissance, inférence) plus faciles à faire et plus conviviales.
- extensibles, en raison de la grande quantité d’informations que contiennent les dictionnaires (ou les bases de données lexicales), on doit pouvoir facilement les générer automatiquement, modifier ou augmenter la quantité d'informations représentées.
- interactives, par opposition à la grande taille des lexiques, les représentations doivent pouvoir fournir des vues restreintes appropriées, et cela de manière interactive.
- versatiles, elle doivent pouvoir s'adapter aux préférences de l'apprenant de manière à pouvoir privilégier une présentation textuelle ou visuelle suivant les cas.

3 Représentations existantes

3.1 Dictionnaires

Les livres peuvent être considérés comme les premiers systèmes hypermédias qui aient jamais existé : information multimédia (texte et images) et organisation hypertextuelle (notes, tables des matières, index, index croisés, etc.). Ceci est encore plus vrai en ce qui concerne les dictionnaires où une lecture linéaire n’est pas du tout celle attendue et où la variété des chemins d’accès aux unités lexicales constituent un des facteurs-clé de l’ouvrage. Comme nous l’avons vu précédemment, les lexicographes anglais ont poussé très loin la description et la présentation des informations lexicales mais celles-ci restent essentiellement d'ordre textuel. En effet, les représentations visuelles restent pour l’instant très classiques. La situation est la même pour les dictionnaires électroniques.

Dans les dictionnaires papier, nous trouvons principalement trois types d’informations visuelles traditionnelles :

- image d’un objet : les dessins d’objets spécifiques comme des outils, des animaux ou des plantes sont souvent plus faciles à comprendre qu’une longue définition. Un dessin est plus informatif qu’un texte, efficace computationnellement (accélère la compréhension et le processus de reconnaissance), mais très limité. De plus, il ne s’applique qu’à des actions ou des noms concrets.
- présentation encyclopédique (comme par exemple la présentation des divers meubles dans une chambre, figure 6.2) : elle peut sembler informationnelle parce qu’elle est le

seul moyen d'accéder aux plus fréquents noms concrets d'un domaine spécialisé de connaissance, mais les objets présentés sont prototypiques. De ce fait, on trouvera uniquement des hyperonymes. Si on cherche des noms d'objets plus spécifiques appartenant à la classe affichée, on devra regarder ailleurs, même si aucun accès n'est prévu depuis l'endroit où l'on est. Ces présentations sont computationnellement limitées et difficilement extensibles à autre chose que des noms concrets. Enfin, la versatilité est aussi réduite car peu de dictionnaires indexent les entrées d'un vocable dans une image encyclopédique.

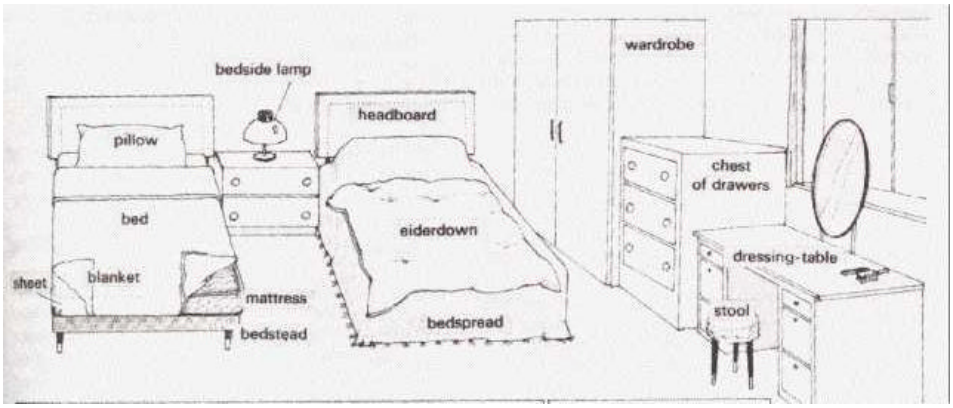


Figure 6.2 : présentation encyclopédique des meubles d'une chambre

- liens sémantiques dans les images encyclopédiques : les relations sémantiques les plus fréquemment rencontrées sont la co-hyponymie (comme les dessins regroupant les différents types de lits, figure 6.3) et la relation partie-tout (affichage des différentes parties d'un corps). Les mêmes commentaires que pour les présentations encyclopédiques s'appliquent ici.

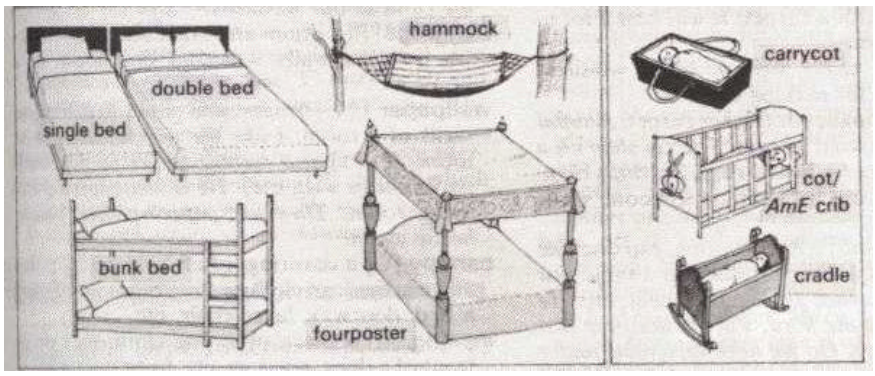


Figure 6.3 : relation de co-hyponymie (divers types de lit)

Il est frappant de constater que très peu de progrès a été effectué dans ce domaine de la lexicographie. On a l'impression qu'il y a un non-dit concernant le fait que les informations visuelles peuvent uniquement être des représentations du monde réel, restreintes et décrivant de ce fait une très petite partie du lexique, considérée comme non linguistique et donc pas

vraiment intéressant. On trouve rarement une représentation visuelle de noms prédicatifs, comme un tableau affichant les relations et les vocables appartenant à une famille humaine. Tout l'effort porte sur la présentation textuelle des synonymes, antonymes, hyponymes, etc.

Ces limitations ne sont pas uniquement dues à la nature du support (encre et papier), car les dictionnaires électroniques, bien que pouvant inclure des traitements automatiques ou du multimédia (prononciation des vocables, par exemple), restent pratiquement au même niveau que les dictionnaires papier concernant les représentations visuelles.

Une raison pour cela pourrait être le fait que la recherche linguistique et psycholinguistique n'est pas assez avancée dans ce domaine.

3.2 Les environnements lexicaux

Avant tout, il convient de se demander quels types de relation sémantique sont les plus intéressants. Les commentaires obtenus par les apprenants lors de l'utilisation du système de Svenconis (Svenconis et Kerst, 1995), peuvent nous mettre sur la voie, même s'ils sont été obtenus à partir d'un nombre limité de sujets (24).

Trois types de cartes sémantiques avaient été dessinées : comme nous l'avons mentionné plus haut, une montrait des vocables étroitement reliés, une deuxième des vocables modérément reliés et une troisième des vocables faiblement reliés. La préférence des apprenants allait au groupe Étroitement Liés (53 %), puis au groupe Modérément Liés (17 %) et enfin le groupe Faiblement Liés (10 %). Le premier groupe (qui affichait la carte de la famille) s'appuyait sur des noms prédicatifs, c'est-à-dire en fait des relations linguistiques, et en conséquence était direct à concevoir. Ce n'est pas le cas du dernier groupe (affichant des évènements autour d'une excursion) qui s'appuie sur une connaissance du monde non linguistique et de ce fait caractéristique de l'image mentale du concepteur. Cette présentation non standard (et aussi non extensible) a pu être en conflit avec la représentation personnelle de l'apprenant (rappelons que le lexique mental a une organisation très individuelle et très différente d'une personne à une autre). On peut donc conclure de cette expérimentation que les relations linguistiques standard semblent être les plus pertinentes pour une représentation visuelle.

Bui (1989) présente un système d'acquisition du vocabulaire en L2, fondé sur l'exploration des liens entre les données multimédias, et présenté comme une extension naturelle des dictionnaires et thésaurus classiques. Dans ce système, appelé HyperLexicon, apparaissent plusieurs nouveaux types de représentations visuelles de relations linguistiques :

- liens analogiques : l'auteur met l'accent sur la nécessité d'aider l'utilisateur à apprendre par analogie. Par exemple pour répondre à des questions comme « le pied est à l'homme ce que le ___ est au cheval », le système peut afficher deux graphes à base de relations partie_tout, l'un pour l'homme, l'autre pour le cheval, et est capable de mettre en évidence sabot lorsque l'utilisateur clique sur pied.
- liens s'applique_à : afin d'illustrer ce qu'elle appelle « proximité sémantique », l'auteur a conçu des graphes aidant l'apprenant à répondre à des questions comme « Quel terme relié à manger s'applique à un cheval ? ». Deux graphes s'appuyant sur des liens *is_a*

affichent les hyponymes d'ingérer et de ChoseVivante. Lorsqu'on clique sur les nœuds manger et cheval, le système peut calculer les relations et afficher le verbe brouter, représenté aussi sur le premier réseau (en utilisant un lien de spécialisation et des contraintes sémantiques). Un problème avec cette représentation est sa computationnabilité : l'apprenant peut facilement faire de fausses inférences comme le fait qu'un cheval broute à chaque fois qu'il ingère quelque chose (ce qui n'est pas le cas dans une étable).

- vues sémantiques paramétrées : ici l'attention est portée sur des relations d'usage (de vulgaire à formel) et sur les connotations positives ou négatives. Par exemple, les adjectifs relatifs à la mort peuvent être ordonnés sur une échelle linéaire allant de crevé (vulgaire) à défunt (formel). Un graphe à deux dimensions (un axe pour l'usage et un autre pour l'antonymie) détermine l'emplacement de verbes exprimant différentes façons d'aimer ou de haïr.

Les deux premières représentations sont informationnellement intéressantes et s'appuient sur des nouveaux types d'interaction, mais elles sont problématiques d'un point de vue computationnel (que veut dire formellement « proximité sémantique ») et sont difficilement extensibles. La troisième est aussi intéressante car elle est un moyen d'afficher différents sens reliés d'adjectifs et de verbes. Mais son extensibilité suppose des échelles linguistiques détaillées. Bui dit qu'elle calcule ses représentations à partir de base de données, y compris celle de WordNet. Mais aucun détail n'est fourni sur la couverture du système (peut-il afficher d'autres représentations que celles faisant partie des démonstrations ?) et rien n'est dit sur l'éditeur hypermédia qui génère les graphes. Bui a peut-être été trop ambitieuse en essayant de mener de front trop de problématiques de recherche.

CAMILLE est un environnement hypermédia pour l'apprentissage du français sur objectifs spécifiques. Il est destiné à un public d'apprenants de niveau intermédiaire et avancé du français langue étrangère (Chanier 1996a ; Lotin *et al.*, 1996). A partir des activités l'utilisateur peut avoir accès à différentes ressources lexicales, un dictionnaire et un ensemble de réseaux lexicaux. Les champs lexicaux, comprenant de nombreux vocables reliés étroitement, ont été choisis en fonction des activités dans le logiciel. Comme il s'agissait ici surtout de rompre avec la description textuelle classique du lexique, les auteurs ont voulu présenter un vocable :

- en tant qu'un ensemble d'éléments extraits de divers ensembles appartenant à différents contextes (un vocable peut être un élément d'une famille lexicale, un élément apparaissant dans plusieurs réseaux ou bien un élément d'information dans les ressources culturelles du logiciel). Grâce aux liens hypertextuels, il est facile de passer d'une présentation graphique à une présentation textuelle et vice versa.
- dont le sens puisse être défini de manière différentielle, c'est-à-dire relativement aux autres vocables.

Les principes à la base de la conception des réseaux de noms, d'adjectifs ou de verbes sont les suivants (figure 6.4) :

- chaque réseau possède un concept central (ténacité, licenciement, persuasion, travail, etc.).

- chaque arc possède une étiquette (dont le nombre est limité) représentant une relation sémantique (antonymie, hyper/hyponymie, quasi-synonymie et une fonction spécifique d'intensification qui relie les vocables possédant divers degré d'intensité). Les trois dernières relations sont représentées sur la figure 6.4.
- les quasi-synonymes sont regroupés dans une ellipse de manière à diminuer la complexité du graphe (réduction du nombre de nœuds et d'arcs) et à aider l'apprenant à faire des inférences : les ellipses entourent les vocables ayant le même sens.
- l'usage des vocables est mis en avant en hiérarchisant les vocables dans trois zones de différentes couleurs correspondant aux registres formel, courant et familier.
- les informations visuelles et textuelles sont reliées : il est possible de passer d'un nœud à son entrée correspondante dans le dictionnaire et vice versa. Une petite fenêtre de commentaire apparaît lorsque l'utilisateur pointe sa souris sur l'étiquette d'un arc expliquant la différence de sens entre les deux nœuds de l'arc.

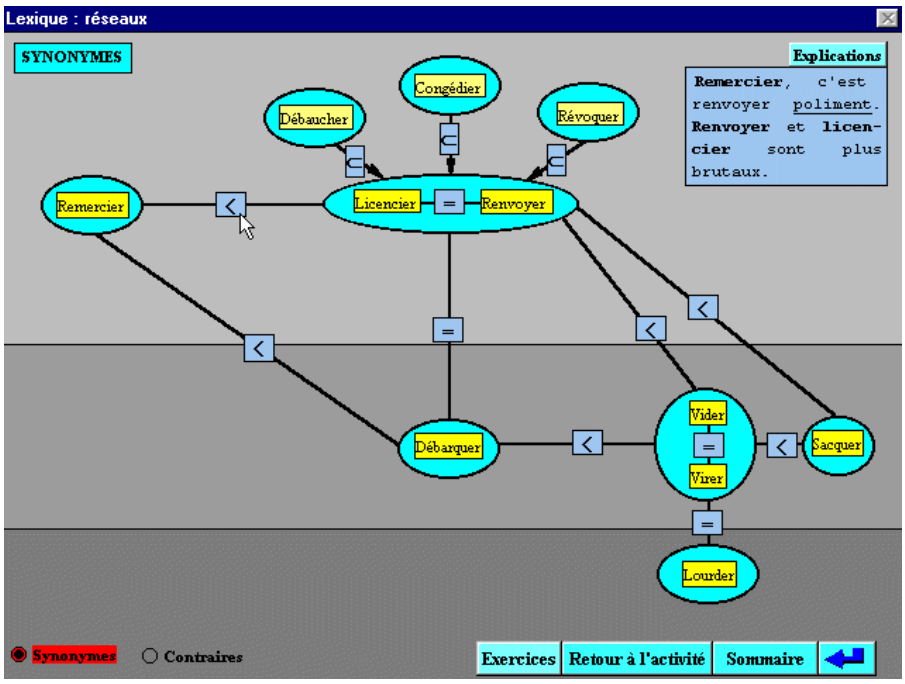


Figure 6.4 : un réseau lexical de synonymes dans CAMILLE
construit autour du thème du licenciement

En observant la figure 6.4, on peut voir que :

- licencier et renvoyer sont quasi-synonymes, de même que vider et virer (dans les acceptions propres à ce thème).
- débaucher, congédier et révoquer sont des hyponymes de licencier (et renvoyer) et ont des sens plus étroits.
- remercier est moins intense (brutal) que licencier et débarquer. De même débarquer et licencier sont moins intenses que vider et virer qui le sont eux-mêmes moins que sacquer, terme le plus violent du réseau.

Enfin, si l'on considère les différents registres, zones marquées par des couleurs plus ou moins sombres, on peut voir que :

- remercier, licencier, renvoyer, débaucher, congédier et révoquer font partie du registre courant.
- débarquer, vider, virer et sacquer appartiennent au langage familier.
- lourder est grossier
- débarquer est quasi-synonyme de licencier mais dans des registres différents, de même que vider et lourder.

Toutes les relations sémantiques sont expliquées dans une fenêtre apparaissant lorsque la souris reste sur le symbole d'un arc.

Les diverses expérimentations menées (évaluations formatives et sommatives au cours desquelles les apprenants ont utilisé le logiciel abondamment – plus de 20 heures par cédérom ; Chanier, 1996b) ont tout le temps reçu des commentaires très positifs sur le logiciel de la part des utilisateurs et aucun problème d'interprétation des réseaux n'a été mentionné. Il faut toutefois préciser dans ces expérimentations que les problèmes lexicaux étaient secondaires et que les mesures n'étaient pas détaillées.

Bien que l'effort ait été mis sur les critères informationnel, computationnel et de versatilité, le principal problème est que ces réseaux ne sont ni extensibles ni modifiables automatiquement, ayant été faits à la main. Nous allons aborder maintenant la génération automatique de réseaux lexicaux en étudiant ce qui a été réalisé au sein de l'environnement ALEXIA.

4 Les réseaux lexicaux d'ALEXIA

Nous commencerons par préciser quelles sont les relations sémantiques traitées par les graphes. Ensuite, nous détaillerons les propriétés des réseaux lexicaux qui sont affichés. Enfin, nous verrons quelles sont les contraintes et limites d'affichage de ces réseaux.

4.1 Regroupement des relations sémantiques traitées par les graphes

Bien que les relations sémantiques puissent être affichées séparément, la carte sémantique d'une lexie sera plus riche et plus globale en regroupant les relations pendant l'affichage suivant leur nature. En effet, actants et dérivés peuvent être affichés simultanément car l'intersection entre l'ensemble des actants et celui des dérivés n'est pas vide (un actant peut être un dérivé).

De la même manière, nous pouvons regrouper les différents types de synonymies. L'hyperonymie qui est une synonymie plus large peut être présentée avec les autres synonymies. Par exemple, si l'on veut les synonymes de directeur, il semble tout à fait normal de lire quelque part chef en vertu du remplacement de directeur par chef dans la phrase :

Ce matin, le directeur nous a expliqué ses projets.

De ce fait, nous pouvons présenter les quasi-synonymes, les hyperonymes, les hyponymes et les synonymes intersectifs d'une lexie simultanément.

Nous avons quatre groupes de relations : les synonymies, l'antonymie, l'actance-dérivation et les fonctions lexicales. Dans l'interface textuelle du dictionnaire d'ALEXIA, toutes ces relations sont présentées. Par contre, seules les synonymies sont traitées par les graphes pour l'instant. D'une part elles sont les plus développées dans la base lexicale. D'autre part, de par certaines propriétés comme la transitivité ou l'égalité, la présentation des différentes synonymies d'une lexie apparaît comme plus appropriée et plus naturelle sous la forme de réseaux.

4.2 Propriétés des graphes

A partir d'une lexie sélectionnée dans le dictionnaire, le système est capable de générer automatiquement un graphe (critère d'extensibilité) montrant toutes les lexies reliées à la première par un type de synonymie. Les nœuds du réseau sont les formes canoniques des lexies suivies de leur numéro de sens dans le dictionnaire. Nous les avons entourées d'une ellipse afin de les mettre davantage en évidence. En fait, chaque ellipse représente un concept sémantique. Les arcs sont des flèches surmontées d'un symbole montrant la relation exprimée (figure 6.5).

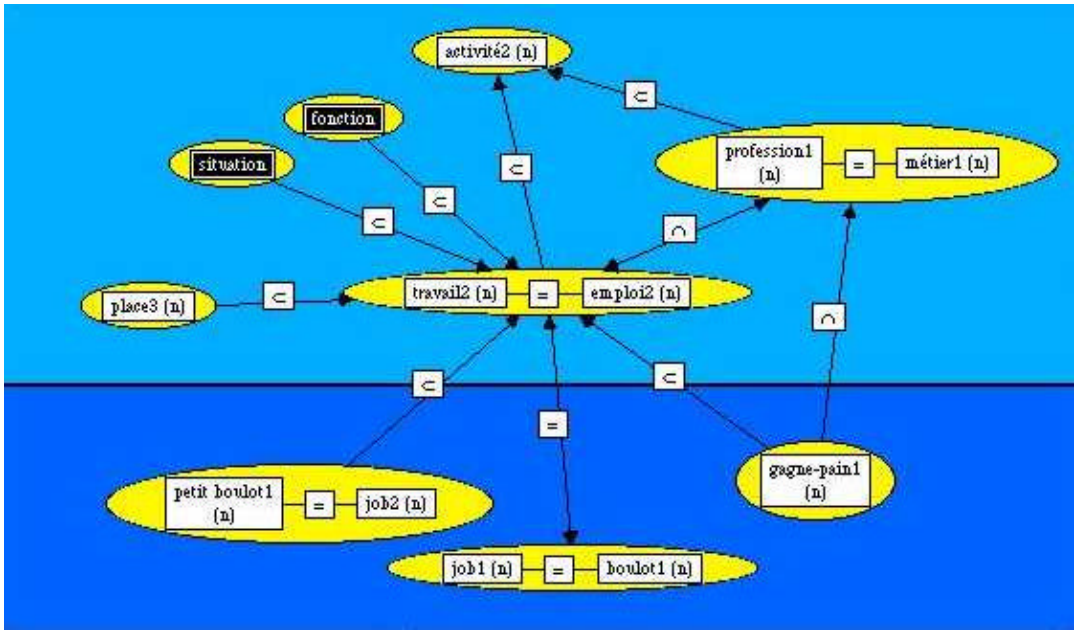


Figure 6.5 : graphe pour emploi

Pour la quasi-synonymie (symbole =) et la synonymie intersective (symbole ∩), l'arc est une double-flèche car ce sont des relations symétriques. L'hyperonymie et l'hyponymie, étant asymétriques et en opposition, ont le même symbole (⊂), car il suffit d'inverser le sens de la flèche pour exprimer l'une ou l'autre relation.

En observant la figure 6.5, on peut voir que :

- travail et emploi sont quasi-synonymes de même que profession et métier

- emploi est une sorte d'activité, c'est-à-dire qu'activité est un hyperonyme d'emploi
- profession est aussi une sorte d'activité
- emploi et profession sont synonymes intersectifs. Ils partagent certains traits sémantiques (relativement au concept de travail) mais ne peuvent pas être considérés comme quasi-synonymes. En effet, en français, profession et métier sont des activités déterminées et définies tandis que emploi et travail ne le sont pas (un emploi de facteur mais *un métier de facteur ou le métier de facteur mais *l'emploi de facteur[8]
- les lexies situation et fonction, hyponymes d'emploi, sont en noir car elles sont seulement citées dans le dictionnaire et non décrites. On touche ici les limites du réseau.

Il est nécessaire d'utiliser des symboles plutôt que des noms ou des paraphrases pour les relations de manière à laisser le graphe clair et à sauvegarder de l'espace. Il y a peu de symboles et des tests (voir Chanier, 1996b pour le logiciel CAMILLE) ont montré que leur interprétation ne posait pas de problèmes aux apprenants. En termes d'efficacité, un symbole s'interprète plus rapidement qu'un nom ou qu'une paraphrase.

Il est important pour un apprenant de connaître facilement le niveau de langue d'une lexie. En effet, le contexte est essentiel dans les relations de synonymies et une lexie, même si la phrase est sémantiquement correcte, peut être trop familière dans un contexte ou trop formelle dans un autre. Nous avons décidé d'établir une gradation dans le graphe en regroupant les lexies dans différentes zones suivant leur registre. Ainsi, les lexies les plus soutenues seront vers le haut du graphe tandis que les plus familières seront vers le bas. Nous avons créé trois zones (correspondant approximativement aux registres soutenu, courant et informel) et mis des couleurs de manière à faciliter leur repérage.

Dans le graphe de la figure 6.5, on peut voir que :

- petit boulot, job, boulot et gagne-pain sont familiers
- job et boulot sont quasi-synonymes
- boulot et travail sont quasi-synonymes mais dans des registres différents

4.3 Taille et limitation dans l'affichage des réseaux

A cause de la taille du réseau, il n'est pas possible de l'afficher totalement et seule la partie autour de la lexie sélectionnée sera visible. Mais où doit s'arrêter cette vue ? A cause de la transitivité de la synonymie, le réseau autour d'une lexie peut être très large et idéalement les frontières sont celles du réseau lui-même. En pratique, il apparaît que l'affichage du niveau 2 des synonymes d'une lexie (c'est-à-dire les synonymes des synonymes d'une lexie) fournit dans certains cas trop d'informations ce qui nuit à la compréhension. De plus la disposition claire de toutes les lexies du réseau les unes en relation avec les autres devient rapidement difficile à établir et les calculs longs à effectuer. L'affichage au niveau 2 demande donc des études de positionnement des lexies et une optimisation des calculs. De ce fait, pour l'instant le système affiche seulement les synonymes directs d'une lexie. Nous pouvons cependant

[8] Suivant que le vocable de base de l'expression est déterminé ou non, l'expression correcte admettra l'article défini ou indéfini.

réserver un traitement de faveur à la quasi-synonymie. En effet, lorsque deux lexies sont quasi-synonymes, elles expriment le même concept et peuvent être regroupées dans une même ellipse. Ce regroupement permet de faire baisser la complexité du graphe tout en affichant plus d'informations (critère d'informationnalité : la représentation doit pouvoir exposer le plus possible d'informations tout en restant claire, c'est-à-dire sans nuire à la compréhension). Par convention, le regroupement n'est pas possible lorsque les lexies n'ont pas le même niveau de langue.

Une représentation informatisée ne doit pas être figée. L'intérêt de l'informatique est justement de pouvoir agir sur les informations et sur leur mode de présentation (critère d'interactivité et de possibilité de traitements automatiques). Nous allons détailler les interactions possibles de l'apprenant avec le système.

5 Interactivité

Les principales interactions sont pour l'instant :

- explications sur chaque partie du graphe : lexies, relations sémantiques, zones de registre,
- déplacement sur le réseau,
- déplacement des flèches et des cercles si le graphe n'est pas clair.

5.1 Explications

L'interaction la plus évidente est la possibilité d'avoir la définition et un exemple d'utilisation d'une lexie lorsqu'on clique sur elle. Ceci est une caractéristique importante car tandis que les définitions et les exemples sont nécessaires à la compréhension, il est inutile de les afficher en permanence sur le graphe (au lieu de les avoir à la demande). Nous pensons ici respecter le critère d'informationnalité (informations et clarté). Une prochaine étape sera la connexion du graphe avec la totalité du dictionnaire. Cela permettra à l'apprenant de choisir entre une représentation graphique et une représentation textuelle des données du dictionnaire (critère de versatilité) et d'avoir facilement à sa disposition toutes les informations disponibles sur une lexie sélectionnée ou sur l'entrée correspondante.

Lorsqu'on étudie des relations synonymiques qui mettent en jeu des vocables proches, il est intéressant de pouvoir comparer les définitions de deux lexies afin de distinguer leurs nuances de sens ou d'emploi. On obtient ces informations lorsqu'on clique sur le symbole d'une relation. Les explications qui sont alors données peuvent aider à saisir les différences entre les deux lexies.

Le système indique aussi à quoi correspondent les différentes zones de registre. Cette possibilité est surtout importante pour les nouveaux utilisateurs qui ne sont pas encore familiers avec le système.

5.2 Déplacement sur le réseau

Il est possible de se déplacer sur tout le réseau en reconstruisant le graphe sur une nouvelle lexie. Supposez que vous vouliez savoir comment on appelle une personne à la tête d'une entreprise.

Vous pouvez commencer par afficher le graphe pour chef (figure 6.6)

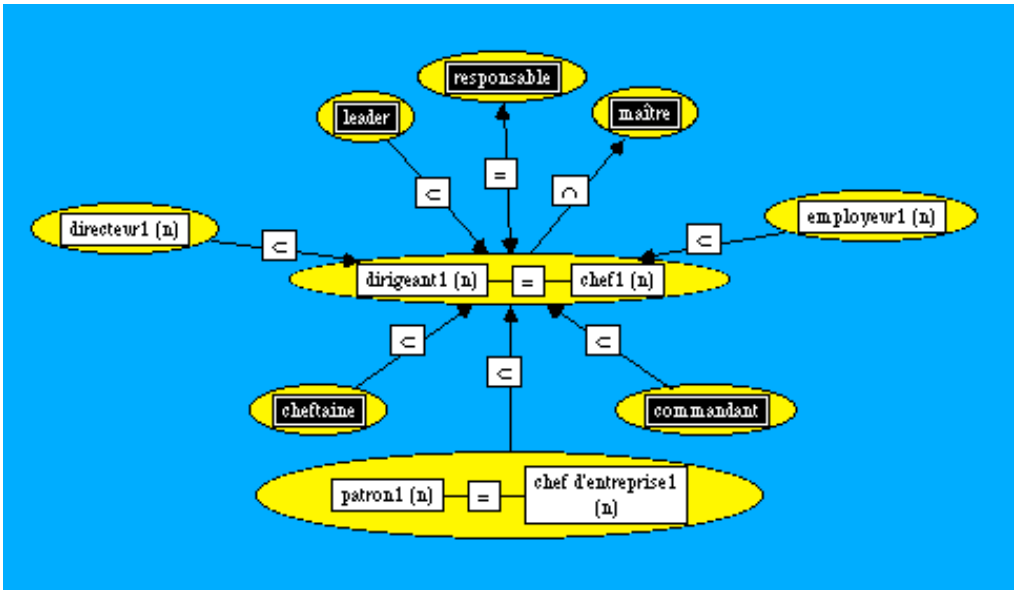


Figure 6.6 : graphe pour chef

En tant qu'hyponyme de chef, vous pouvez lire patron (ou chef d'entreprise), c'est-à-dire ce que vous cherchez. Mais si vous voulez plus de détails (par exemple, y a-t-il d'autres vocables pour patron, des sortes de patron ?), vous pouvez reconstruire le réseau autour de patron (figure 6.7).

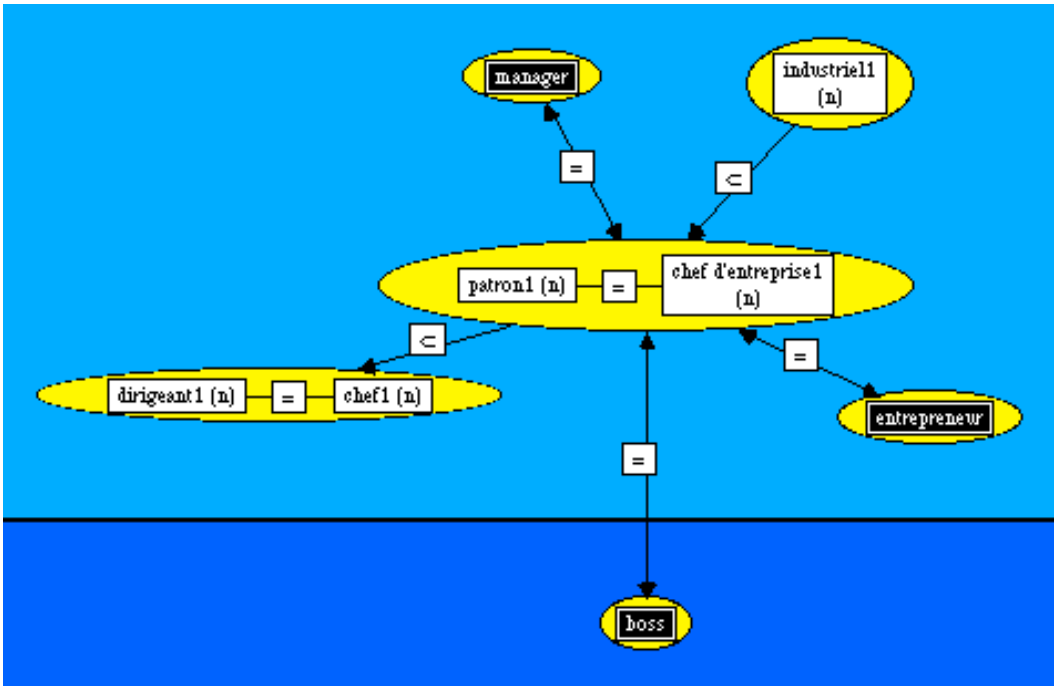


Figure 6.7 : graphe pour patron

Il apparaît alors boss (vocable familier pour patron) qui est peut-être ce que vous cherchiez. Vous découvrez qu'il y a des sortes de patron, comme industriel, hyponyme de patron (figure 6.8).

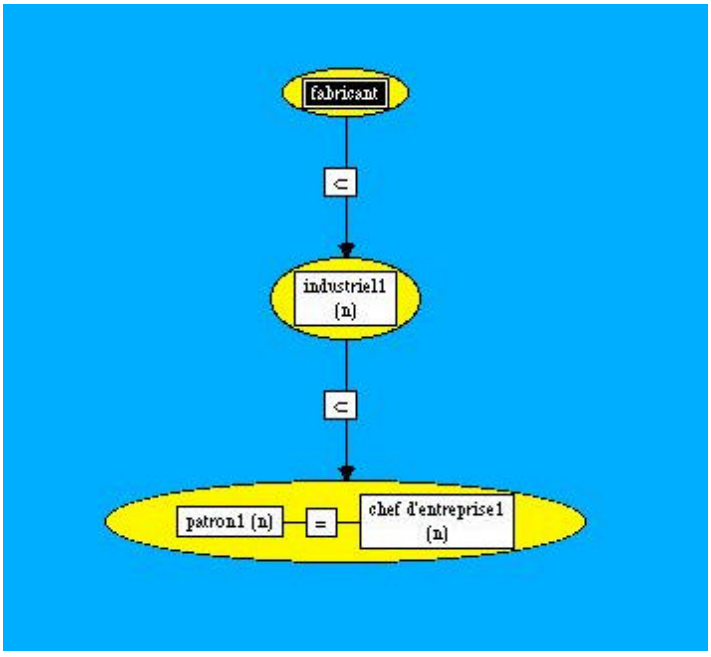


Figure 6.8 : graphe pour industriel

Et sur la figure 6.8 apparaît fabricant.

En production, il est très pratique de pouvoir aller et venir d'une lexie à une autre en fonction du sens que l'on veut exprimer. Pour chaque lexie, l'utilisateur peut voir facilement et

rapidement les sens les plus proches. Dans le cadre d'une utilisation textuelle, la plupart du temps par hypertexte, les synonymes d'un vocable n'ont pas de disposition particulière. Comme les vocables sont suivis de leur définition, il faut plus d'attention pour les séparer du texte qu'un rapide regard sur le graphe.

Cette circulation graphique d'un concept à un concept proche sémantiquement peut se considérer comme une autre manière d'utiliser un dictionnaire. Elle permet de bien saisir le sens d'une lexie et est plus intuitive que le parcours d'une liste alphabétique dans lequel seuls le hasard et les conventions sont intervenus pour disposer les vocables les uns par rapports aux autres. En effet, tout comme le souligne Miller (Miller *et al.*, 1993), ce mode de circulation reflète davantage la structure du lexique mental en faisant intervenir les associations sémantiques entre concepts.

5.3 Glisser et déposer

Dans les cas où beaucoup de lexies et de relations doivent être affichées, il peut y avoir intersection des flèches et superposition de plusieurs nœuds. En effet, la complexité du graphe croît avec la quantité d'items à afficher et une présentation claire demande des algorithmes compliqués et de longs temps de calcul. Pour l'instant, l'utilisateur peut obtenir un présentation claire en manipulant simplement les éléments du graphe par un glisser-déposer.

Nous allons maintenant présenter la méthode que nous avons adoptée pour la génération automatique des graphes. Cette méthode a été élaborée et programmée ex-nihilo par Pichon (1996) car aucun algorithme de calcul de graphe n'était satisfaisant pour pouvoir générer un réseau dont nous avons exposé les caractéristiques ci-dessus.

6 Implémentation informatique

La construction du graphe est précédée d'une étape d'extraction. En effet, il faut déterminer en premier lieu les synonymes de la lexie à afficher. Le programme parcourt la base de données lexicales et calcule les synonymes par réflexivité et transitivité. La complexité du calcul est alors liée au nombre de transitions à parcourir.

L'affichage du graphe s'effectue en cinq étapes par un algorithme qui établit les coordonnées pour chaque lexie. Le but consiste à afficher une lexie L et ses synonymes, ses voisins V_i. Par convention, L est au centre du graphe et les voisins V_i placés autour de lui.

1) Le programme classe les V_i en trois groupes, suivant leur niveau de langue. Un groupe peut être vide.

2) Suivant le nombre de voisins dans chaque registre, le programme calcule le secteur angulaire à réserver pour chaque registre en fonction de constantes pré-définies. Plus un registre contient de voisins, plus le secteur réservé sera grand.

3) Pour chaque registre, le programme calcule le poids de chaque lexie et lui alloue un secteur. Puis, il calcule chaque tendance et ordonne les lexies dans une liste en fonction de cette valeur.

Le poids d'un nœud N (par rapport à un autre, L) est son encombrement. Il dépend de la taille de la lexie (longueur de la graphie), du nombre de quasi-synonymes à afficher dans le même cercle (ou ellipse suivant les cas) et des nœuds qui lui sont liés.

Le poids est donné par la fonction récursive *Poids*. Étant donné V, un voisin, L, la lexie, Niv, le niveau de transitivité (Niveau 1 signifie synonyme d'une lexie, Niveau 2 signifie synonyme d'un synonyme, etc.) et W_1,W_2,..., W_i,..., W_n, les voisins de V :

$Poids(V,L,Niv) = Aire(V)*n^2 + (\Sigma Poids(W_i,V,Niv+1))/(5^{Niv})$

$Poids(V_0,L_0) = Poids(V_0,L_0,1)/1000 + 10$

En examinant cette formule, on peut en déduire que la complexité du calcul du poids des nœuds du graphe est d'ordre n, n étant le nombre de nœuds sur le graphe, les poids des nœuds les plus à l'extérieur étant réutilisés pour ceux plus proches du centre. La formule récursive est prévue pour afficher plusieurs niveaux de synonymies mais comme nous avons juste besoin des synonymes directs de la lexie du centre, le poids est seulement fonction de l'aire du nœud (comme on ne considère pas ses voisins, le facteur multiplicateur *n* est égal à l'unité).

La tendance d'un nœud N (par rapport à un autre, L) est un coefficient heuristique déterminant la direction dans laquelle on conseillerait de tracer N de manière à ce que son arc avec L n'en coupe pas un autre lorsque N est lié avec un nœud dans un autre registre (figure 6.9). Si l'arc doit être tracé plutôt vers le bas, la tendance sera plutôt négative.

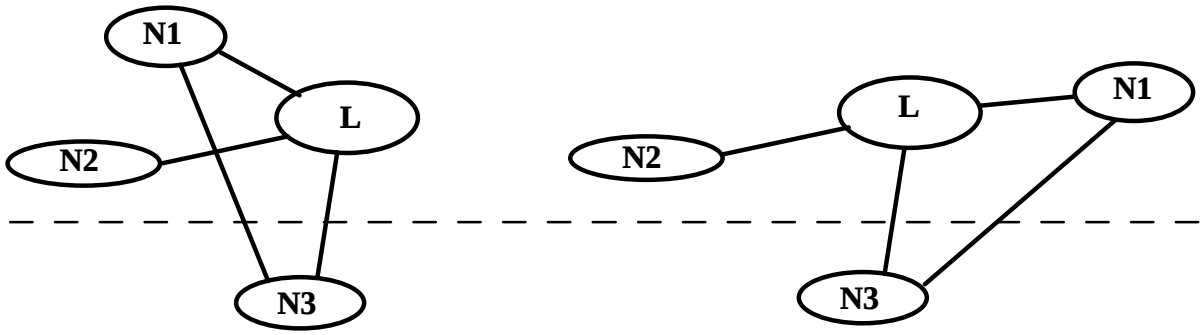


Figure 6.9 : tendance inappropriée pour N1 à gauche, correcte à droite

La tendance est donnée par la fonction récursive *Tend* :

$Tend(V,L,Niv) = TendL(V,L) + (\Sigma Tend(W_i,V,Niv+1))/(5^{Niv})$

$Tend(V_0,L_0) = Tend(V_0,L_0,1)$

où *TendL* est une valeur dépendant du registre de V et de L. Par exemple, TendL(V,L) = 0 si V et L sont dans le même registre mais TendL(V,L) = +50 si V est dans le registre de la partie haute et L dans celui de la partie basse. Les calculs de tendance sont donc utiles lorsque

les nœuds ne sont pas dans le même registre. La complexité du calcul est la même que celle concernant les poids, les deux formules fonctionnant de la même manière.

4) Le programme place les voisins registre par registre autour de L en leur donnant une position angulaire. Le voisin qui a la tendance la plus faible est placé au-dessous de L (angle $-\pi/2$). Puis, on remonte vers le haut à gauche ou à droite de L et on donne une position angulaire à chaque voisin en fonction de la place qu'il reste, des poids et tendances calculés et des liens entre voisins (qui doivent être considérés avant les tendances). Si un des voisins est lié au nœud qui vient d'être placé, il est prioritaire sur les autres pour être placé à côté de lui, indépendamment de sa tendance. Un lien est alors tracé entre les deux nœuds (par exemple entre activité et profession sur la figure 6.5). Il faut donc à chaque placement d'un nouveau nœud examiner si les voisins restants (ceux qui ne sont pas déjà placés sur le graphe) ne sont pas liés. Dans ce cas, la complexité du calcul est donc d'ordre n^2 (il y a (n-1)*(n-2)/2 relations à considérer, *n* étant ici le nombre de voisins de L). Le programme doit placer les voisins alternativement à gauche et à droite de L de manière à ce que le graphe soit équilibré et que les nœuds soient placés harmonieusement autour de L (figure 6.10).

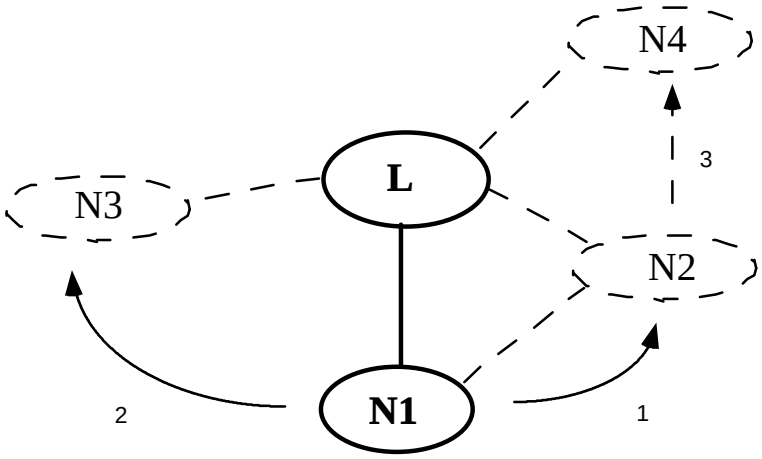


Figure 6.10 : ordre de placement des nœuds autour de L

Les tendances calculées étant des coefficients heuristiques, leur valeur n'est pas toujours adéquate pour éviter les croisements d'arcs. Il est parfois nécessaire de procéder à un réarrangement manuel pour que le graphe soit plus lisible (cas de la figure 6.5 où il a fallu replacer profession et activité).

5) Le programme calcule un rayon pour chaque voisin suivant le nombre de registres qui les sépare. Cette valeur fixée est modifiée si nécessaire de manière à ce que les nœuds ne se recouvrent pas.

Pour plus de détails sur cet algorithme, voir Pichon (1996).

7 Conclusion

Après avoir évoqué la théorie du codage double de l'information qui postule un double système de stockage de l'information, l'un verbal et l'autre non verbal, nous avons examiné diverses expérimentations montrant l'influence positive du multimédia pour l'apprentissage lexical. Nous avons vu que les conclusions contestables de certaines d'entre elles (relayées par d'autres travaux) sur l'inefficacité du multimédia était surtout le fait de principes pédagogiques mal définis ou inappropriés. Cela nous a amené à dégager des critères d'évaluation de la qualité d'une représentation visuelle.

Nous avons ensuite étudié les différentes représentations visuelles existantes principalement dans les dictionnaires et les environnements utilisant des ressources lexicales.

Nous avons ensuite examiné en détail les graphes générés automatiquement par ALEXIA en montrant les principales caractéristiques de cet affichage et en décrivant les possibilités d'interactions. Nous avons rapidement exposé la méthode de positionnement des lexies dans les graphes en soulignant les difficultés d'ordre linguistique et computationnel.

Comme nous l'avons vu, le programme affiche seulement les synonymes directs d'une lexie. Dans certains cas, il peut être intéressant d'en afficher plus en faisant figurer par exemple les synonymes d'ordre 2. Mais la complexité du graphe s'accroît et les contraintes telles que l'intersection des arcs deviennent difficiles à respecter. Elles demandent certainement de nouveaux algorithmes de positionnement des lexies.

Un autre développement possible est l'affichage de relations sémantiques différentes des synonymies telles que l'antonymie, la dérivation ou l'actance. Ceci doit être fait mais la complexité des relations telle que l'antonymie (les antonymes peuvent être exclusifs, peuvent suivre une gradation, etc.) demande davantage de travail linguistique et informatique.

Le multimédia ne doit pas seulement représenter le non verbal. A côté des informations textuelles, le visuel est tout à fait adapté pour décrire les objets concrets du monde réel mais nous semble aussi pertinent pour la présentation d'informations linguistiques plus abstraites telles que des relations sémantiques standard avec les concepts qu'elles relient.

CHAPITRE 7

Préparation des activités lexicales

Les activités lexicales telles qu'elles ont été définies au chapitre 4, exercices de recontextualisation et jeu du Mai, sont construites sur la base de concordances de vocables donnés. Il s'agit en effet, pour chaque vocable, de fournir une phrase qui servira de contexte. Ces phrases sont extraites du corpus.

Le problème doit donc être résolu en trois étapes : il faut d'abord pouvoir repérer chaque vocable dans les textes indépendamment de sa flexion. Il faut donc, dans un premier temps, étiqueter les textes, c'est-à-dire que chaque vocable doit être accompagné de sa forme canonique (ou lemme) et de sa catégorie grammaticale, qui permet de bien identifier le vocable (sans toutefois traiter les cas d'homonymie à ce niveau), ainsi que de codes morphologiques qui rendront compte de la flexion.

Dans un deuxième temps, il faut savoir dans quel texte, dans quel paragraphe ou dans quelle phrase se trouve tel ou tel vocable. La phrase sert de contexte court et le paragraphe de contexte plus large si besoin est. Quant au texte, on pourrait par exemple, choisir un texte automatiquement en retenant celui qui contient le plus de vocable dans un ensemble donné. Cette deuxième opération revient donc à indexer les vocables, un peu comme le font les moteurs de recherche sur Internet, capables de donner en un rien de temps l'ensemble des documents contenant tel ou tel item. La différence cependant est que ces moteurs indexent des occurrences, c'est-à-dire, des graphies (et donc des formes fléchies), tandis qu'ALEXIA doit indexer des vocables, et donc des formes canoniques. Il faut donc réfléchir sur la structure de données permettant des réponses efficaces et rapides.

Enfin, la troisième étape consistera en l’exploration de l’index et la génération des concordances.

1 L'étiquetage des textes

L’étiquetage d’un corpus n’est pas une fin en soi mais bien souvent le préalable à des traitements plus conséquents. Il permet aux mots de changer de statut, passant de graphies à vocables, tout du moins homonymes car il n’est pas possible pour un étiquetage

morphologique de faire des distinctions sémantiques : ainsi, alors que l'étiqueteur fera la différence entre boucher nom et verbe, il ne pourra la faire entre les deux voler, ceux-ci ayant la même catégorie grammaticale.

L'étiquetage consiste à associer à chaque segment de texte (le plus souvent une séquence de lettres, intuitivement un « mot », qui peut comprendre dans le cas des unités polylexicales plusieurs espaces comme travail au noir) une ou plusieurs étiquettes morpho-syntaxiques comprenant la catégorie grammaticale et parfois les codes flexionnels. Certains étiquetages peuvent en outre indiquer le lemme ou des informations contextuelles comme l'appartenance d'un vocable à tel ou tel texte.

Il existe maintenant de nombreux corpus étiquetés, dont il serait difficile de dresser la liste exhaustive (Habert *et al*. (1997, p. 17) en présente quelques uns parmi les plus connus), ceci en raison des méthodes employées pour les constituer et les annoter ou des études linguistiques qui les ont utilisés. La plupart proviennent du monde anglo-saxon et sont en général disponibles pour la recherche universitaire, contrairement aux corpus du français qui ne sont pas dans le domaine public.

L'étiquetage se déroule en trois phases, plus ou moins visibles par l'utilisateur du système : en premier, il y a la segmentation du texte en vocables, ensuite il y a l'assignation à un vocable d'une ou plusieurs étiquettes en fonction des dictionnaires appliqués (plusieurs catégories grammaticales sont possibles) et enfin il y a la levée d'ambiguïtés par l'exploration du contexte. Le résultat final de l'étiquetage dépend donc de nombreux facteurs qui en multiplient les formes, la qualité et le taux d'erreur, à tel point que la comparaison et l'évaluation des différents systèmes d'étiquetage n'est pas chose aisée. Ces points de divergence reprennent les étapes ci-dessus.

L'étiquetage peut être plus ou moins automatisé : il est rare en effet qu'il n'y ait pas d'intervention humaine si l'on tient à un résultat fiable. Il dépend aussi de la segmentation des textes en vocables. Comme le montre Silbserztein (1993), le phénomène est complexe, nécessitant par exemple de traiter des signes délimiteurs ou non tels que l'apostrophe (deux vocables pour l'étude mais un seul pour aujourd'hui), le tiret (délimitant dans vient-il mais pas dans les mots composés comme gagne-pain) ou l'espace (pomme de terre est un seul vocable), et des phénomènes compliqués tels que l'insertion d'un vocable entre l'auxiliaire et le participe passé (J'ai toujours fait...) ou dans des expressions semi-figées (Le phénomène a toujours lieu le vendredi). Enfin, l'étiquetage est fonction du nombre d'étiquettes et de la couverture des dictionnaires appliqués (ce dernier point valant aussi pour la segmentation en vocables). En effet, plus un système contient d'étiquettes ou le dictionnaire appliqué de vocables, plus la description linguistique sera fine, mais plus le taux d'ambiguïtés potentielles sera élevé. Il faut donc trouver un compromis entre un étiquetage juste et une description précise.

Les systèmes actuels utilisent deux grands types de méthodes pour la levée d'ambiguïtés : la désambiguïsation par règles et la désambiguïsation probabiliste (ou stochastique).

La première méthode s'appuie sur le fait que certaines suites de catégories grammaticales sont illicites. Ainsi, un verbe ne peut pas être précédé d'un déterminant. Dans la phrase il le licencie, le ne peut être un déterminant et est donc un pronom. Cependant, il faut parfois explorer le contexte un peu plus loin pour pouvoir désambiguïser des phrases comme il la

trompe. Ces règles sont contenues dans des « grammaires locales » (Silberztein, 1993) qui, en s'appliquant aux textes, éliminent les « chemins » de catégories incorrects. Comme le note Habert *et al*. (1997, p. 167), ces automates ne savent pas traiter les dépendances à longue distance que l'on trouve dans la syntaxe. C'est également le cas pour les systèmes probabilistes.

Ces derniers s'appuient sur la régularité des suites de catégories grammaticales dans les textes. Il s'agit de déterminer en fonction des deux ou trois étiquettes précédant un vocable dans le texte (on parle de bigramme ou de trigramme) les probabilités d'apparition des étiquettes candidates et de garder celle qui a la probabilité maximale. Ces probabilités sont calculées à partir de corpus d'apprentissage préalablement étiquetés et corrigés. Ils doivent être de taille suffisante et sont établis eux-mêmes à partir de corpus plus petits. Pour que le procédé soit fiable, il est néanmoins nécessaire que les corpus d'apprentissage et celui à étiqueter aient des caractéristiques langagières pas trop éloignées.

Brill (1995) résume bien les avantages et inconvénients des deux approches : « [Les] étiqueteurs stochastiques ont bien des avantages sur les étiqueteurs bâtis manuellement, en particulier ils rendent superflue la construction laborieuse de règles manuelles, et saisissent des informations utiles qui peuvent ne pas avoir été remarquées par l'analyste humain. Cependant, les étiqueteurs stochastiques présentent l'inconvénient que les connaissances linguistiques ne sont capturées qu'indirectement, par le biais de grands tableaux statistiques ».

Certaines approches mixtes combinent les avantages des deux méthodes : les erreurs des systèmes probabilistes sont corrigés par des règles. Selon El-Bèze et Spriet (1995), il suffit d'écrire 4 à 5 règles pour corriger environ 50 % des erreurs commises par un système probabiliste.

D'autres approches, dans une perspective d'apprentissage (Brill 1995), tentent de générer des règles ou d'en corriger d'autres à partir de corpus préalablement étiquetés. Le système compare les résultats et propose des règles de correction. Celles-ci sont appliquées, le résultat est à nouveau comparé avec l'étiquetage manuel et ainsi de suite. Le processus s'arrête lorsque le nombre d'erreur générées est supérieur à celui des erreurs corrigées.

En termes de performances, les taux généralement annoncés à l'aide des deux premières méthodes sont de l'ordre de 95 à 98 % d'étiquette justes. Ces pourcentages peuvent paraître bons mais ils doivent être relativisés car, d'une part, ils incluent souvent la ponctuation, qui couvre 10 à 15 % des textes, et d'autre part, seuls environ 40 % des vocables sont ambigus. En effet, Tzoukermann et Radev (1996) trouvent que 58 % des vocables de textes extraits du journal *Le Monde* (200 000 mots) sont non ambigus (25 % présentent deux ambiguïtés et 11 % en présentent trois). Ces chiffres sont à rapprocher des résultats obtenus par l'étiqueteur Cordial (Cordial, 1998, voir plus bas) sur un corpus littéraire (fin 19^{e}) d'environ 2 500 000 mots pour lequel 62 % des vocables ne sont pas ambigus (20 % présentent deux ambiguïtés et 11 % en présentent trois).

Une erreur de 5 % d'étiquetage correspond en moyenne à un vocable sur vingt, soit grosso modo un vocable par phrase. Ceci peut être gênant pour des traitements automatiques ultérieurs (comme l'analyse syntaxique). Néanmoins, certaines ambiguïtés n'ont pas les

mêmes conséquences : on ne peut pas juger sur le même plan une erreur entre un nom et un verbe et une erreur entre un adjectif et un participe passé.

Comme nous l'avons vu précédemment, la multiplicité des méthodes de segmentation et du nombre d'étiquettes font qu'il est difficile de comparer les performances de différents systèmes. En effet, un étiqueteur avec un jeu réduit d'étiquettes ou une segmentation sommaire n'aura pas grand mal à obtenir de meilleurs résultats qu'un autre programme décrivant les données de manière plus fine. C'est l'objet de l'action d'évaluation GRACE (Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation) (Paroubek *et al*., 1997) que de comparer les résultats d'une vingtaine de participants confrontés à l'étiquetage d'un corpus donné du français.

Les problèmes des différents jeux d'étiquettes sont résolus par l'établissement de tables de correspondance entre le jeu de GRACE et ceux des participants. Le jeu de GRACE est le plus fin et une ou plusieurs étiquettes GRACE (idéalement une) correspondent à une étiquette d'un participant. Pour les différences de segmentation, des algorithmes d'alignement de corpus ont été mis au point dans le cas notamment des unités polylexicales (on parle de zones floues lorsque les segmentations entre le corpus GRACE et celui d'un des participants diffèrent), introduisant des caractères fictifs et permettant par la suite de faire porter la comparaison sur les mêmes portions de textes. Le programme extrait automatiquement les morceaux de textes qui n'ont pu être réalignés.

Les résultats des différentes possibilités d'étiquetage (désambiguïsation totale, partielle ou sans désambiguïsation) sont donc comparés avec les étiquettes du corpus de référence et l'évaluation est présentée sur un graphique à deux dimensions où un premier axe mesure la « décision » (selon que plus ou moins de décisions linguistiques ont été prises par le système, c'est-à-dire que la description est plus ou moins fine) et un deuxième montre la « précision » (erreurs ou non d'étiquetage).

Les résultats sont actuellement en cours de validation (GRACE, 1999).

2 L'étiquetage du corpus ALEXIA

Pour donner une idée de ce corpus, voici ses principales caractéristiques. Celles-ci seront détaillées dans le chapitre 9.

Taille : environ 400 textes pour 450 000 occurrences. Environ 19 000 vocables différents selon la segmentation opérée par notre étiqueteur (voir ci-dessous).

Forme : textes ASCII concaténés en un seul fichier, chaque texte contenant les informations suivantes : titre, nom de fichier initial, nombre de mots, vocables-clés, registre, références de provenance. Textes mis au format SGML par les membres du serveur SILFIDE (SILFIDE, 1999).

Langue : Français courant contemporain.

Sources : Livres et presse spécialisée et non-spécialisée contemporaine (1994-1996) sur le domaine du *travail*, de l'*emploi* et du *chômage*.

Auteurs : Thierry Chanier et Thierry Selva ainsi que d'autres membres du LRL dans lequel il a été constitué pour les projets CAMILLE et ALEXIA (1994-1996).

Pour étiqueter le corpus ALEXIA, nous avons utilisé dans un premier temps une version datant de quelques années d'Intex (Silberztein, 1993 et 1994), facile d'accès et fonctionnant sur un PC 486 sous système d'exploitation Next. Nous avons obtenu en sortie, après application de dictionnaires (dont les très complets DELAF pour les vocables simples et DELAC pour les vocables composés) et grammaires locales fournies par le système, des graphes que nous avions l'intention de désambiguïser manuellement. Après quelques milliers de levées d'ambiguïté, nous nous sommes aperçus que le taux de vocables ambigus s'élevait à près de 30 %, soit près d'un vocable sur trois en moyenne. Notre corpus comptant environ 400 000 occurrences, cela faisait près de 130 000 ambiguïtés à lever, un travail considérable, long et fastidieux, pour le peu de moyens humains dont nous disposions. Un peu plus tard, nous avons recommencé le même étiquetage avec la dernière version d'Intex (version 4.0), tournant sur Windows 95, et comprenant un plus grand nombre de grammaires locales. Après comparaison sur quelques phrases tests, les résultats ne nous ont pas semblé meilleurs de manière significative, même si ces nouveaux résultats n'ont pas fait l'objet d'une évaluation comme pour le premier étiquetage. En fait, Intex n'a pas pour but d'étiqueter totalement les textes, mais de lever un taux maximal d'ambiguïté sans erreur. Il s'agit donc de grammaires parfaites, qui ne produisent pas d'erreur mais qui laisse un taux important de vocables non-désambiguïsés. Ce taux n'a pas été évalué sur les dernières versions. Ces résultats s'expliquent par la taille des dictionnaires, visant à l'exhaustivité, et à un grand jeu d'étiquettes.

Nous avons donc utilisé un autre étiqueteur, commercialisé à un prix préférentiel pour la recherche universitaire, Cordial 5 (Cordial, 1998) version université de la société Synapse. Ce produit a pour réputation d'être conforme à l'état de l'art, avec un taux de 5% d'erreur. La figure 7.1 montre le résultat de l'étiquetage sur une phrase du corpus. Chaque code correspond à une étiquette. Il y en a environ 110.

```
Selon l'enquête emploi, la population active en France s'élève à 24 826 000
personnes en mars 1992.

========== DEBUT DE PHRASE ============
Selon      23
l'         15
enquête    26
emploie    103
,          201
la         15
population 26
active     02
en         23
France     26
s'         38
élève      103
à          23
24 826 000 11
personnes  27
en         23
mars       31
1992       31
```

Figure 7.1 : Etiquetage par Cordial 5 (fichier `etiquetage.txt`)

Deux problèmes surgissent au regard de ces résultats : tout d'abord, Cordial ne fait pas figurer les lemmes, ce qui du coup rend impossible tout indexation des vocables des textes. D'autre part, un examen attentif révèle que Cordial 5 corrige les textes en même temps qu'il les étiquette, bien que toutes les options de correction soient désactivées. Cordial est en effet avant tout un correcteur grammatical, sur lequel a été ajouté un module d'étiquetage. Dans l'exemple de la figure 7.1, emploi a été « corrigé » en emploie qui devient du coup verbe (code 103) !

Les textes du corpus n'ont pas été relus manuellement. Il subsiste donc des erreurs et une correction automatique pourrait être la bienvenue. Cependant, Cordial, comme les autres correcteurs, génère des erreurs, et surtout les occurrences obtenues sont différentes de celles des textes. Il devient dès lors impossible d'établir la correspondance entre les textes et le résultat de l'étiquetage.

Concernant le deuxième problème, il a fallu concevoir un petit programme qui aide au repérage des vocables modifiés. Heureusement, leur nombre n'était pas trop grand, à peu près 1 % soit 4 000 occurrences à repérer et à modifier. Les modifications portaient entre autres sur les accents manquants (qui étaient signalés par les mauvais accords entre auxiliaire avoir et participe passé).

Ce bogue semble néanmoins avoir été corrigé sur une version postérieure (5.03).

Le premier problème a été plus ennuyeux. Fort heureusement, nous disposions de l'analyseur morphologique du LIPN qui a permis de retrouver les lemmes à partir d'une occurrence et de la catégorie grammaticale déterminée par Cordial. Il n'y avait d'ailleurs pas forcément accord entre les catégories de Cordial 5 et celles de l'analyseur, les étiquettes étant différentes. Cependant, ces désaccords n'intervenaient que sur des mots outils, qui n'étaient pas fléchis. Dans ces cas, les codes grammaticaux de Cordial ont été laissés pour plus de cohérence avec le reste du texte. Le calcul des lemmes ne s'est donc effectué que pour les mots pleins (noms, verbes et adjectifs). Outre les mots outils, deux autres types d'occurrence n’ont pas pu être traités correctement : les vocables inconnus de l'analyseur, pour lesquels le « lemme » reste identique à l'occurrence, faute de mieux, et, bien sûr, les unités polylexicales. L'analyseur ne contenant aucune expression, celles qui sont reconnues par Cordial restent, comme au cas précédent, sous leur forme fléchie.

Une fois le programme d’étiquetage effectué, il faut nettoyer les résultats, pour lever quelques ambiguïtés sémantiques. En effet, la lemmatisation peut aboutir à deux formes canoniques pour une même occurrence (suis peut avoir comme infinitif être ou suivre, connaissances peut être singulier ou pluriel, etc.). Il faut alors faire le choix manuellement, en fonction du contexte.

3 L’indexation

Le but de cette étape est d’aboutir à un index de tous les vocables des textes. Cet index doit contenir la position en octet de chaque occurrence du vocable dans le fichier des textes (les

textes sont en effet concaténés dans un seul fichier, pour plus de facilité). Il doit en outre être possible de savoir dans quelles phrases, dans quels paragraphes et dans quels textes ces occurrences apparaissent.

Avant l'indexation, il est nécessaire d'opérer un balisage des textes : repérage des mots dans le fichier, marquage des textes, phrases et paragraphes.

3.1 Préliminaire : le balisage des textes

Pour cela, il faut exécuter un programme qui repère chaque occurrence de la figure 7.1 dans les textes, en comptant les caractères un à un et en indiquant le caractère de début de l'occurrence. Par la même occasion, on introduit des marqueurs de début de texte (inscrits dans les textes lors de la concaténation) et de paragraphes (dès qu'un saut de ligne est rencontré). Quant aux phrases, le découpage a été effectué par l'étiqueteur et il suffit de répercuter le marqueur. Cela donne le fichier de la figure 7.2 qui indique, par exemple, que le texte débute à l'octet 1859289 [DTE], la phrase (et le paragraphe) à l'octet 1859301 [DPA] et l'occurrence de Selon à l'octet 1859315.

```
[DTE]_1859289
Emploi_Emploi_24_1859289
[DPA]_1859301
[DPH]_1859301
PRÉSENTATION_PRÉSENTATION_26_1859301
Selon_Selon_23_1859315
l'_l'_15_1859321
enquête_enquête_26_1859323
emploi_emploi_24_1859331
,_,_201_1859337
la_la_15_1859339
population_population_26_1859342
active_actif_2_1859353
en_en_23_1859360
France_France_35_1859363
s'_s'_38_1859370
élève_élever_103_1859372
à_à_23_1859378
24 826_24 826_11_1859380
000_000_11_1859388
personnes_personne_27_1859392
en_en_23_1859402
mars_mars_31_1859405
1992_1992_31_1859410
[DPH]_1859416
À_À_23_1859416
cette_cette_9_1859418
date_date_26_1859424
```

Figure 7.2 : Balisage des textes (fichier `balisage.txt`)

3.2 L'indexation

Le programme suivant est le programme d'indexation proprement dit. Il a deux fonctions. En premier lieu, il repère chaque marqueur de début de texte, de paragraphe ou de phrase, et enregistre leur position dans les textes dans des fichiers correspondants (figure 7.3).

```
00000003
00000023
00000037
00000151
00000255
00000349
00000368
00000387
00000420
00000444
00000490
00000547
00000622
00000652
```

Figure 7.3 : Adresses de début de chaque phrase (fichier `phrases.txt`)

A chaque ligne correspond le début de chaque phrase. La phrase n°2 commence ainsi à l'octet 23 du fichier des textes. Les adresses ont toutes la même longueur pour pouvoir accéder très rapidement, grâce à la fonction `fseek` du langage C (positionne le curseur de

fichier à l'octet désiré), à l'information de n'importe quelle ligne. L'information de la ligne `n` commence donc à l'octet `(n-1)*8` et on lit 8 caractères pour avoir l'adresse entière.

Dans un deuxième temps, le programme classe chaque occurrence rencontrée dans un arbre binaire ordonné avec les informations associées.

Il s'agit ensuite de le lire de gauche à droite et d'écrire cet arbre dans un fichier qui à la forme suivante :

```
Vocable_typecat:(formflec1_codecat pos1_numtexte1_numparag1_numphrase1
pos2…)(formflec2 pos1_... pos2_...)(...)
```

où `typecat` est un nombre représentant un type de catégorie grammaticale (pour différencier ainsi certains homonymes) (nom, verbe, pronom, préposition, etc.), `formflecn` représente une forme fléchie du vocable, `codecat` le code flexionnel attribué par Cordial et `posn`, `numtexten`, `numparagn` et `numphrasen` respectivement la position de `formflecn` dans les textes, le numéro du texte, du paragraphe et de la phrase qui la contient. La figure 7.4 (fichier `index.txt`) montre un extrait de l'index (une ligne a été sautée entre chaque ligne pour plus de lisibilité) :

```
allongement_7:(allongement_24 41553_7_60_555 41652_7_60_556 195276_26_411_1787
432124_56_853_3871 442774_58_871_3958 447906_59_881_3999 639295_78_1339_5777
639487_78_1339_5778 643950_78_1348_5818 1538891_188_2972_14133
1854868_227_3549_17086 2235488_300_4098_19972)

allonger_9:(allonger_100 47897_7_74_599 431734_56_853_3868 645422_78_1351_5833
1429027_174_2788_13151 1436987_175_2793_13224 1515614_185_2934_13962
1555423_192_3009_14281 1609274_199_3126_14792)(allonge_103 442609_58_869_3956
442650_58_870_3957 470871_62_922_4203)(allongent_106
847565_104_1744_7582)(allongée_151 1680065_205_3279_15380)

allouer_9:(allouer_100 1513378_185_2932_13949)(alloués_152
2018697_259_3821_18373)(allouée_151 2356547_324_4225_20949 2356697_324_4225_20950)

allumage_7:(allumage_24 1784958_219_3435_16351 1785854_219_3435_16360)

allumer_9:(allumé_150 872157_106_1786_7829)(allumer_100 1018536_131_2079_9174
1734517_212_3386_15901)(allumant_149 2420626_334_4315_21537)

allumeur_7:(allumeurs_25 649189_79_1362_5862)

allure_7:(allures_27 380188_51_763_3444 1253432_158_2497_11384)(allure_26
494403_64_981_4444 1404866_172_2739_12989 1471068_180_2851_13556)
```

Figure 7.4 : index des mots des textes

Ce fichier comporte 19 276 lignes, ce qui veut dire que le corpus ALEXIA contient environ 19 000 formes canoniques de catégorie grammaticale différente. Ainsi, boucher nom et boucher verbe sont comptés comme deux « mots » différents tandis qu'il n'y a pas de distinction entre voler, dérober et voler, être dans les airs, la distinction étant ici uniquement sémantique.

Comme nous allons le voir par la suite, ce comptage inclut quelques unités polylexicales.

Ce fichier a été obtenu en quatre minutes et demi sur Sun4 Sparc SunOS 5.6 à partir d'un fichier initial d'environ 420 000 lignes.

4 La génération de concordance

En premier lieu, il faut rechercher dans le fichier `index.txt` l'information associée à un vocable. Nous employons ici un algorithme de recherche dichotomique qui procède au maximum à 15 accès fichier ($\log_2(19276) = 14{,}21$), ce qui se fait très rapidement avec les implémentations actuelles de C.

Le problème, c'est que les lignes n'ont pas toutes la même longueur et il n'est donc pas possible d'accéder directement au début de telle ou telle ligne correspondant à une forme canonique donnée.

Il faut donc créer un index intermédiaire qui va repérer, pour chaque forme canonique, le début de la ligne associée dans le fichier `index.txt`, et les écrire de manière régulière dans le fichier `indexsimple.txt` (figure 7.5). Ce fichier est composé de 19 276 lignes et 3 colonnes de largeur constante (en nombre de caractère) : la forme canonique, le type de catégorie grammaticale et la position de la forme dans le fichier index.

```
allocataire                    7  00159755
allocation                     7  00159972
allocations de chômage         7  00163867
allongement                    7  00164165
allonger                       9  00164432
allouer                        9  00164771
allumage                       7  00164914
allumer                        9  00164986
allumeur                       7  00165126
allure                         7  00165173
allusion                       7  00165314
allège                         9  00165407
allègements                    7  00165495
allègent                       9  00165550
allègrement                    1  00165617
alléchant                      0  00165688
allécher                       9  00165832
allées et venues               7  00165882
allégeant                      9  00165964
allégement                     7  00166012
alléger                        9  00166222
allégez                        9  00166455
allégresse                     7  00166503
alléguant                      9  00166577
allégé                         9  00166629
allégée                        9  00166697
allégées                       9  00166768
allô                           5  00166837
alors                          1  00166901
alors même qu'                 4  00171868
alors même que                 4  00171968
```

Figure 7.5 : index intermédiaire

Une fois tous ces fichiers constitués, la génération de concordances se fait de la manière suivante :

1) L'utilisateur entre une forme canonique et une catégorie grammaticale.

2) Le programme effectue une recherche dichotomique sur le fichier `indexsimple.txt` (figure 7.5) et retourne `pos1`, la position de la forme dans le fichier `index.txt` (figure 7.4).

3) Le système accède à l'octet `pos1` et lit dans `index.txt` jusqu'à la fin de la ligne l'ensemble des informations associées.

4) Pour chaque occurrence dans la ligne du fichier `index.txt`, le système lit le numéro de phrase associé, recherche la ligne de même numéro dans le fichier d'adresses des phrases (`phrases.txt`, figure 7.3) et retourne l'adresse `pos2`.

5) Le système lit le fichier des textes à l'adresse `pos2`. Le nombre d'octet à lire (longueur de la phrase) est calculé par la différence entre les adresses de deux lignes consécutives dans le fichier `phrases.txt`.

Ainsi effectuée, la recherche est très rapide et la génération de concordances est quasi-instantanée (figure 7.6). Cet algorithme est moins performant que celui qui utiliserait un automate d'état fini. Néanmoins, le temps de réponse est satisfaisant et la programmation a été moins contraignante.

```
Je voulais savoir si j'avais les capacités à évoluer vers ce type de boulot, et je
me suis inscrit à une ENCP.
Cette partie du boulot m'ennuie, je préfère conceptualiser et créer", reconnaît-il
aujourd'hui.
Vous croyez que c'est drôle de chercher du boulot à 20 ans ? "
J'ai plongé dans le civil , pris des cours de droit commercial , cherché du boulot
, sans en trouver.
Et, pendant ce temps-là, on se tape tout le boulot. "
Ceux qui font mine de ne plus se passionner pour leur boulot sont vraisemblablement
des ratés qui se consolent",ajoute ce contrôleur financier de Saint-Gobain.
Aujourd'hui, il n'a toujours pas de boulot.
Je ne suis pas capable de tenir un boulot", avouait-elle.
J'ai été malade du boulot, raconte Denis Capron, PDG d'Opex.
Un boulot pas très enthousiasmant, mais un boulot tout de même.
Du boulot dans les PAIO
Trouver mieux et plus vite un boulot dans ses compétences - et son salaire et
connaître en permanence l'évolution de son métier.
Un public qui s'élargit Ils sont plus de 3 millions sans boulot, et pour beaucoup,
ça dure.
```

Figure 7.6 : concordances générées pour le vocable boulot

5 Conclusion

Nous avons abordé dans ce chapitre la préparation nécessaire des textes pour la génération de concordances dans le cadre des activités lexicales retenues au chapitre 4 (activités de recontextualisation et jeu du Mai). Le préliminaire à tout travail consiste à étiqueter les textes (segmentation, assignation de catégories grammaticales et levée d'ambiguïtés), de manière à opérer sur des vocables et non pas sur des chaînes de caractères. Après avoir décrit les principales méthodes d'étiquetage, nous avons vu que les résultats obtenus par la plupart des étiqueteurs étaient satisfaisants, même si un « nettoyage » manuel partiel s'avérait le plus souvent nécessaire. Concernant le corpus ALEXIA, l'étiquetage a été réalisé avec Cordial 5.

La génération de concordances passe d'une part par le balisage des textes (phrases, paragraphes et textes) et l'indexation de tous les vocables dans un fichier, et, d'autre part, par

des procédures de recherche exploitant ce fichier. Le balisage a été réalisé en enregistrant les positions de chaque début de phrase, de paragraphe et de texte dans des fichiers spécifiques de manière à connaître les octets que le système doit lire pour l'affichage de concordances. Pour l'indexation, les vocables ont été ordonnés dans un arbre binaire puis « mis à plat » dans un fichier linéaire, où chaque ligne contient toutes les positions des occurrences du vocable considéré dans les textes, ainsi que celles des phrases, paragraphes et textes qui les contiennent. La recherche de positions de vocables dans ce fichier d'index se fait par dichotomie, via un index simplifié pour ne pas être ralenti par les longueurs variables des lignes. Le temps de réponse est tout à fait acceptable.

Nous allons voir maintenant les activités qu'il est possible de générer dans le système ALEXIA à partir de concordances.

CHAPITRE 8

Les activités lexicales dans ALEXIA

Nous avons vu au chapitre 4 les deux types d'activités lexicales, avec leurs variantes, que nous souhaitons proposer aux apprenants dans le cadre de l'utilisation de l'environnement ALEXIA.

Le premier type d'activité concerne des exercices de recontextualisation : à partir d'une sélection de vocables, le système génère des concordances, occulte l'item à chercher et demande à l'utilisateur de le retrouver. Les vocables à retrouver peuvent être liés entre eux par des relations sémantiques ou non et peuvent être affichés ou non. La deuxième activité, le jeu du Mai, reprend le même principe, sauf que les concordances ne seront pas générées à partir d'une sélection de vocables, mais d'un seul, choisi pour sa polysémie. Les phrases affichées concerneront donc ses lexies. Dans les deux cas, l'apprenant pourra compter sur un système d'aide pour lui permettre de réussir l'activité.

Nous discutons dans ce chapitre les caractéristiques de ces deux types d'activités telles que nous souhaitons les concevoir dans le cadre d'ALEXIA, ainsi que les problèmes que vont soulever leur réalisation. Il s'agit ici de spécifications, ces activités n'ayant été implémentées qu'en partie. Néanmoins, il est toujours tenu compte de la faisabilité de leur implémentation.

Nous verrons en effet que, contrairement aux exercices conçus sur le papier, la génération automatique de ces activités à partir de matériaux bruts et authentiques, les textes du corpus, amène à les spécifier de manière précise. En effet, l'enseignant rédigeant les épreuves fait abstraction de beaucoup d'éléments que nous ne pouvons nous empêcher de considérer. L'ALAO impose donc une vue plus générale et plus complète de la conception d'activités.

Les points à étudier et les problèmes à résoudre s'articulent autour de cinq points :

- le choix des vocables/lexies à partir desquels les concordances vont être générées : s'agit-il en effet de vocable ou de lexie, comment ceux-ci ou celles-ci sont-elles déterminées par le système ?
- l'affichage des concordances : les concordances doivent-elles être des phrases ou des contextes ? Que faire si la phrase est trop courte ou trop longue, si le contexte n'est pas suffisant ?

- l’aide proposée : quel type d’aide ? Sera-t-elle la même en fonction de l’activité proposée ou même en fonction de ses variantes ?
- l’acceptation des réponses : demande-t-on, dans le cas d’exercice à trous, de fléchir les formes à retrouver en fonction du contexte ? Si oui, a-t-on les moyens de vérifier automatiquement leur correction ou leur semi-correction ?
- la notation : une réponse est-elle juste ou fausse ? Ou peut-on délimiter des réponses partiellement justes ?

Nous nous proposons maintenant de développer et d’étudier les points ci-dessus.

1 La sélection des vocables ou des lexies

Il n’est pas pertinent de faire effectuer des activités lexicales sur n’importe quels vocables, en particulier ceux sur lesquels l’apprenant n’a pas été exposé et qu’il a donc toutes les chances de ne pas connaître. Notre but n’est pas l’aide à l’apprentissage lexical de toute la langue, mais d’une partie seulement, en fonction du champ notionnel que nous nous sommes fixé. Au cours des activités, l’apprenant teste les connaissances qu’il a acquises et non pas celles qu’il n’a pas.

Les travaux de Goodfellow (1995) ont montré l’importance d’un module de dictionnaire personnel visant à noter et à organiser le vocabulaire en partie connu. En effet, il ne suffit pas de chercher le sens d’un vocable dans un dictionnaire pour le retenir : c’est ce que montrent les travaux en psycholinguistique (voir chapitre 1) et l’expérience avec GLOSSER (2.2.1) où l’utilisateur consulte plusieurs fois les mêmes entrées de dictionnaire.

Nous avons élaboré et implémenté un dictionnaire personnalisé dans ALEXIA (figure 8.1). Celui-ci permet de sélectionner des vocables dans un texte, de les regrouper suivant des caractéristiques communes et de visualiser les groupes constitués. En outre, le module attribut un statut à chaque élément des groupes. Il correspond à la quantité de traitement effectué sur un item et sera utilisé par les activités lexicales pour déterminer ceux sur lesquels elles porteront.

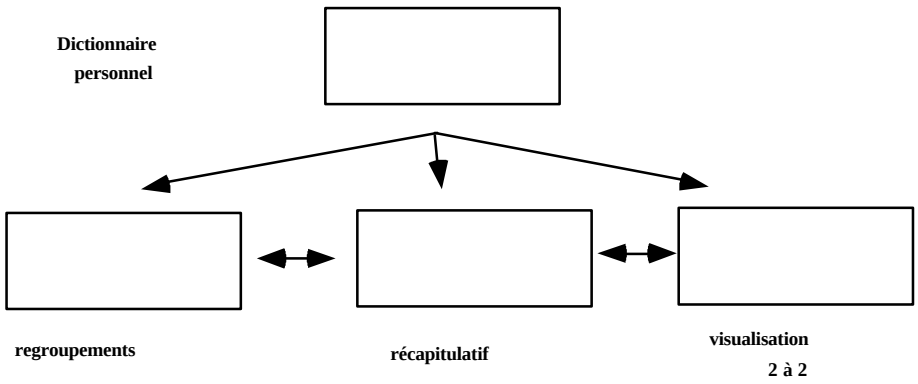


Figure 8.1 : architecture du dictionnaire personnel

1.1 Sélection des vocables dans le dictionnaire personnel

A partir de graphies sélectionnées dans un texte (désambiguïsées et lemmatisées par l'analyseur morphologique de manière à manipuler des vocables, figure 8.2), l'apprenant a la possibilité d'organiser ce vocabulaire dans différents ensembles qui regroupent des vocables partageant une même caractéristique.

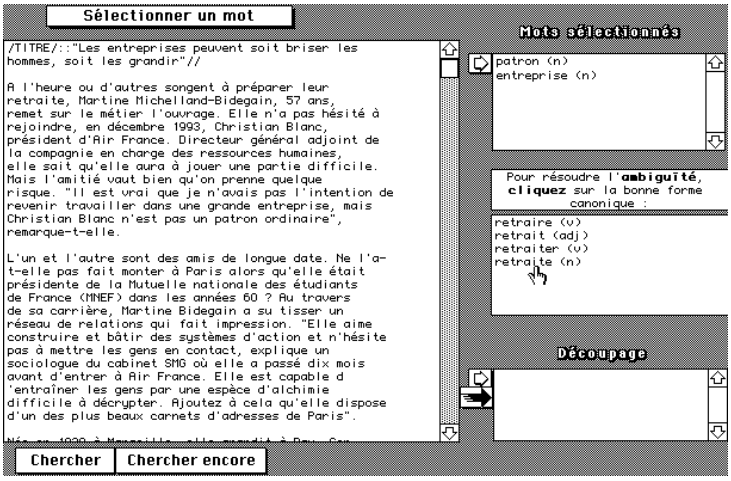


Figure 8.2 : désambiguïsation par l'analyseur morphologique du vocable sélectionné retraite

Un petit traitement de segmentation a été opéré sur les graphies « simples » (ne contenant aucune espace). Ainsi les cas des capitales et de l'apostrophe (sauf sur le s' du verbe pronominal) ont été considérés pour faciliter l'analyse morphologique. Par contre, l'analyseur ne contenant aucune unité polylexicale, celles-ci peuvent être sélectionnées (le cas du retour chariot au milieu de l'expression est traité) mais en aucun cas lemmatisées ou étiquetées. De même pour les graphies ne faisant pas partie du dictionnaire de l'analyseur (comme les abréviations PDG, DRH ou les vocables composés avec trait d'union comme après-midi ou week-end).

1.2 Regroupement des vocables

Lorsque l'apprenant a fini sa sélection, un bouton l'invite à classer les vocables en les regroupant dans différents groupes, selon sa volonté (figure 8.3).

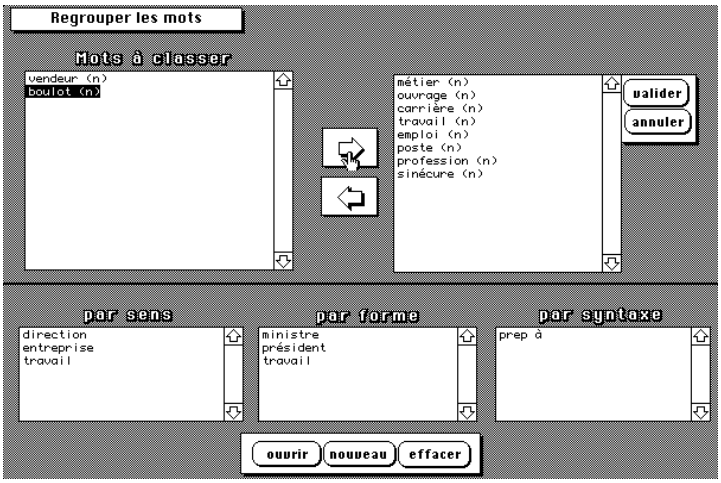


Figure 8.3 : classement du vocable boulot dans le groupe consacré au travail

L'utilisateur a le choix entre trois types de regroupement :

- un regroupement par sens qui réunit des vocables définissant un même paradigme (synonymes, antonymes, etc.). C'est le cas dans la figure 8.3 du groupe *travail (par sens)* dont le contenu est affiché dans la fenêtre en haut à droite.
- un regroupement par forme qui réunit les vocables ayant un similitude de forme comme les dérivés d'une famille lexicale ou, par exemple, partageant un même suffixe. Le groupe *président (par forme)* de la figure 8.3 contient les vocables président, présidente, vice-présidente, présidentielle. Il pourrait comporter présidence ou vice-présidence si ceux-ci avaient été sélectionnés dans les textes.
- un regroupement par syntaxe, qui concerne surtout les verbes. L'apprenant a loisir d'y regrouper, par exemple, les verbes se construisant avec une même préposition. Ainsi le groupe *prep à (par syntaxe)* contient les verbes songer, hésiter, correspondre, s'inscrire et survivre, qui se construisent tous (mais pas exclusivement) avec la préposition à.

Les groupes peuvent être créés, effacés ou bien complétés. Pour les remplir, ou les vider, il suffit de faire passer les items des fenêtres gauche et droite à l'autre au moyen des flèches directionnelles du milieu.

L'apprenant a la possibilité de mettre des annotations (figure 8.4, non géré pour l'instant) sur les vocables qui ont été sélectionnés dans les textes : définitions personnelles, traductions en langue source ou toute information pertinente pour lui. Ces annotations sont sous forme de texte simple. Elles sont libres et donc ni structurées et ni formalisées. Le système ne peut donc que les reciter telles quelles sans pouvoir les exploiter pour d'autres usages. Elles sont donc utiles principalement pour l'apprenant.

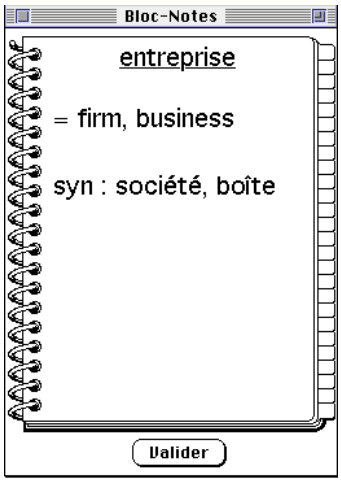
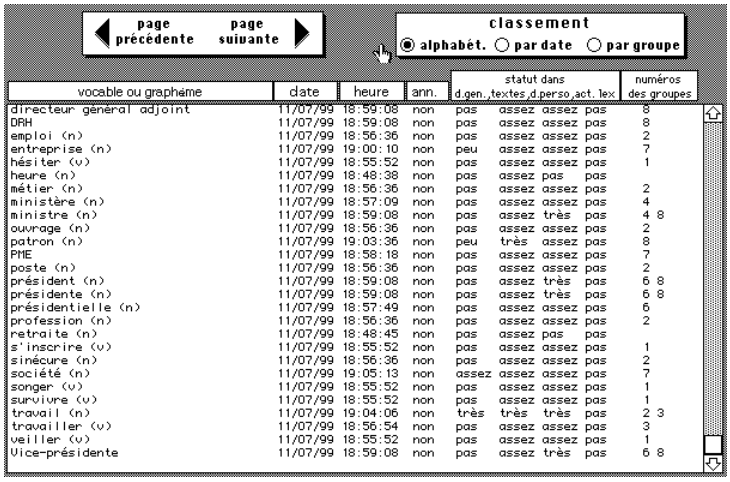


Figure 8.4 : exemple d'annotation d'un anglophone pour entreprise

Par ces regroupements et ces annotations, l'apprenant a donc la possibilité de travailler en profondeur le sens et la forme des vocables qu'il a isolés des textes. Bien plus complet que la simple consultation de l'entrée dans le dictionnaire, ce travail de catégorisation permet de renforcer les liens faibles du lexique mental créés lors de la lecture et de l'exposition aux vocables.

1.3 Visualisation du contenu du dictionnaire personnel

L'apprenant dispose de deux modes de visualisation du contenu du dictionnaire personnel. Le premier est un récapitulatif qui présente une synthèse du travail accompli (figure 8.5)



vocable ou graphème	date	heure	ann.	statut dans d.gen., textes, d.perso, act. lex	numéros des groupes
directeur général adjoint	11/07/99	18:59:08	non	pas assez assez pas	8
DRH	11/07/99	18:59:08	non	pas assez assez pas	8
emploi (n)	11/07/99	18:56:36	non	pas assez assez pas	2
entreprise (n)	11/07/99	19:00:10	non	peu assez assez pas	7
hésiter (v)	11/07/99	18:55:52	non	pas assez assez pas	1
heure (n)	11/07/99	18:48:38	non	pas assez pas pas	
métier (n)	11/07/99	18:56:36	non	pas assez assez pas	2
ministère (n)	11/07/99	18:57:09	non	pas assez assez pas	4
ministre (n)	11/07/99	18:59:08	non	pas assez très pas	4 8
ouvrage (n)	11/07/99	18:56:36	non	pas assez assez pas	2
patron (n)	11/07/99	19:03:36	non	peu très assez pas	8
PME	11/07/99	18:58:18	non	pas assez assez pas	7
poste (n)	11/07/99	18:56:36	non	pas assez assez pas	2
président (n)	11/07/99	18:59:08	non	pas assez très pas	6 8
présidente (n)	11/07/99	18:59:08	non	pas assez très pas	6 8
présidentielle (n)	11/07/99	18:57:49	non	pas assez assez pas	6
profession (n)	11/07/99	18:56:36	non	pas assez assez pas	2
retraite (n)	11/07/99	18:48:45	non	pas assez pas pas	
s'inscrire (v)	11/07/99	18:55:52	non	pas assez assez pas	1
sinécure (n)	11/07/99	18:56:36	non	pas assez assez pas	2
société (n)	11/07/99	19:05:13	non	assez assez assez pas	7
songer (v)	11/07/99	18:55:52	non	pas assez assez pas	1
survivre (v)	11/07/99	18:55:52	non	pas assez assez pas	1
travail (n)	11/07/99	19:04:06	non	très très très pas	2 3
travailler (v)	11/07/99	18:56:54	non	pas assez assez pas	3
veiller (v)	11/07/99	18:55:52	non	pas assez assez pas	1
Vice-présidente	11/07/99	18:59:08	non	pas assez très pas	6 8

Figure 8.5 : récapitulatif du contenu du dictionnaire personnel

Cette synthèse se présente sous forme de tableau où les lignes contiennent les vocables considérés et les colonnes un élément d'information : nom du vocable ou de la graphie (non désambiguïsée par l'analyseur), date et heure de dernière modification, présence ou non d'annotation, niveau de traitement dans les différents modules et enfin le numéro des groupes

dans lesquels l'élément a été classé (ou non). Une petite fenêtre (non encore implémentée) indique la correspondance entre les groupes et leur numéro pour faciliter la compréhension.

Ainsi, sur la figure 8.5, le vocable travail a été modifié pour la dernière fois le 11 juillet 1999 à 19h04, aucune annotation ne lui a été associée, il a reçu une quantité importante de traitement dans les modules dictionnaire général, textes et dictionnaire personnel, n'a fait l'objet d'aucune activité lexicale et a été classé dans les groupes 2 et 3. Les quantités de traitement sont relatives sur l'ensemble des items et sont déterminées de la manière suivante : une variable est assignée à chaque item et contient le nombre d'action par module (dans le dictionnaire général, une action équivaut à la consultation d'un vocable et de ses lexies (un clic sur les définitions pour avoir les informations associées incrémente le compteur), dans les textes, une action équivaut à une sélection, dans le dictionnaire personnel, une action équivaut à un regroupement dans un ensemble et enfin, dans les activités lexicales, la variable indique, à la place du nombre d'actions, le score qu'a reçu l'item au cours des divers exercices). Ensuite, le système parcourt, pour chaque module, l'ensemble des valeurs et note le maximum et le minimum. Il divise la différence en trois et attribut le statut en fonction de l'intervalle dans lequel se classe la valeur de la variable. Exemple : si la valeur maximum parmi les items est 13 et celle minimum est 1, l'intervalle est donc de 4. Une variable ayant pour valeur 7 classera l'item dans l'intervalle 5-9, soit celui du milieu. Les termes *très*, *assez* et *peu* qualifient chacun des intervalles. Si l'item n'a pas été travaillé, sa variable reçoit la valeur *pas*.

Les items peuvent être ordonnés par ordre alphabétique (pour pouvoir les repérer facilement), par ordre chronologique, du plus récent au plus ancien (pour savoir ceux sur lesquels l'apprenant a travaillé récemment) et par ordre de groupe (pour retrouver rapidement le contenu de chacun d'eux).

De cette manière l'utilisateur sait rapidement le degré de travail effectué sur chaque vocable en fonction des modules.

Un deuxième mode de visualisation lui permet d'avoir plus de détails sur chacun des vocables (figure 8.6).

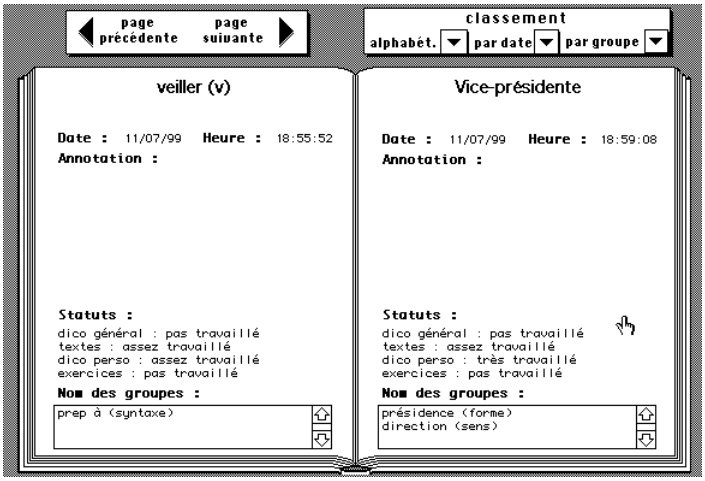


Figure 8.6 : Visualisation deux à deux du contenu du dictionnaire personnel

L'apprenant retrouve les mêmes informations, mais plus détaillées, avec en plus les annotations si celles-ci ont été rédigées. Mais l'intérêt de cette visualisation est de pouvoir comparer le contenu de deux vocables.

De même que dans le mode précédent, les vocables peuvent être rangés suivant l'ordre alphabétique, chronologique ou par groupe.

Après avoir ainsi montré les caractéristiques du module de dictionnaire personnel, revenons aux activités lexicales et à la manière dont les vocables sont sélectionnés.

1.4 Détermination des vocables pour les activités

Ces vocables sont retenus en fonction de leur statut dans le dictionnaire. On peut discerner trois groupes. Il y a, d'une part, les vocables sur lesquels peu de travail a été effectué. En ce sens, il est légitime de penser que l'apprenant ne connaît pas suffisamment d'informations sur eux et qu'il n'est donc pas nécessaire de déclencher des activités. Cela pourrait signifier aussi que l'apprenant connaît déjà bien le vocable et qu'il n'a pas besoin des connaissances des ressources lexicales. Cependant, il est peu probable que ces vocables soient sélectionnés dans les textes, car ils présentent peu d'intérêt du point de vue de l'apprentissage. Nous considérons donc qu'un faible travail implique une connaissance insuffisante et nous écarterons dès lors ces vocables pour les activités. D'autre part, à l'opposé, il y a ceux sur lesquels beaucoup de travail a été effectué et qui ont reçu une bonne note lors d'une activité précédente. Le système se doit donc de les écarter en les considérant comme acquis. Entre ces deux groupes se trouvent les vocables avec un statut intermédiaire, suffisamment travaillés, mais pas encore maîtrisés. Ce sont ceux-là que le système retient et sur lesquels porteront les activités lexicales.

Cependant, le problème ne s'arrête pas là. En effet, les activités peuvent se déclencher sur un ensemble de vocables non forcément reliés sémantiquement (auquel cas, les concordances peuvent être générées sans attendre), mais nous avons vu qu'il était plus intéressant de construire automatiquement ces exercices de recontextualisation sur un ensemble de vocables formant un réseau sémantique. Les relations retenues sont (chapitre 4) la synonymie, l'actance et la dérivation.

Le problème est que ces relations sont établies de lexie à lexie, et non pas de vocable à vocable. Donc comment savoir quelle lexie figure dans une concordance générée à partir d'un vocable ? Il faut donc effectuer un traitement supplémentaire sur le corpus dans ce cas-là : une désambiguïsation sémantique.

Les travaux en désambiguïsation sémantique, ou désambiguïsation lexicale lorsqu'il s'agit d'attribuer un numéro de sens à un segment de texte (Word sense disambiguation), ne sont pas aussi avancés que les travaux d'étiquetage morphologique. L'affaire est bien plus complexe puisqu'elle porte sur le sens des vocables et donc sur la représentation des connaissances. Les corpus annotés sémantiquement existent pour l'instant à l'état embryonnaire puisque l'étiquetage est fait manuellement. Le plus connu est certainement celui fourni sur le site de WordNet (1999) où les mots de classes ouvertes (c'est-à-dire les noms, verbes, adjectifs et adverbes), étiquetés morphologiquement au préalable, sont mis en

correspondance avec les sens de WordNet. Ainsi trois extraits du corpus Brown sont proposés, ce qui représente, sur un total de 680 000 formes, 234 000 étiquettes (ou pointeurs) sémantiques.

Néanmoins, pour en revenir à ALEXIA, deux facteurs spécifiques rendent la tâche moins ardue.

En effet, on peut considérer d'une part que toutes les lexies n'ont pas la même importance car l'exercice sera d'autant plus intéressant que la lexie aura de liens vers d'autres lexies. Autrement dit, le système favorisera la lexie la plus précisément détaillée en calculant parmi toutes celles du vocable celle qui possède le plus de liens dans le réseau considéré. Dès lors, l'objectif de la désambiguïsation consiste à repérer cette lexie et à lui constituer manuellement un ensemble de concordances parmi toutes celles générées à partir du vocable. Un premier objectif pourrait être une trentaine de concordances pour la lexie dominante de chaque vocable.

D'autre part, vu l'orientation du corpus ALEXIA sur le champ notionnel du *travail*, *emploi*, *chômage*, la distribution des lexies est différente de celle d'un corpus plus général. Le corpus ayant servi de base de description lexicographique, les lexies les plus détaillées, celles pour lesquelles le dictionnaire contient le plus de synonymes ou d'actants, sont précisément celles qui sont le plus attestées dans le corpus. C'est le cas par exemple pour les vocables travail, société, cadre, boîte, etc. Dès lors, le parcours et la désambiguïsation sémantique manuelle des concordances se fait bien plus rapidement, l'annotateur devant se contenter de vérifier que le sens de l'occurrence est bien celui attendu, plutôt que de faire un choix à chaque fois entre les différentes lexies possibles.

Par contre, pour le jeu du Mai, l'affaire est moins directe, puisqu'il faut des concordances pour toutes les lexies d'un vocable (on pourra se contenter des principales, en excluant les plus marginales). La tâche est d'autant plus fastidieuse que, contrairement aux activités ci-dessus, l'exercice demande des acceptions que le corpus ne contient pas (c'est le cas du cadre qu'on accroche au mur) et qu'il faudra donc inventer ou récupérer dans un corpus plus général (il en faudrait quatre ou cinq par lexie). Toutefois cette désambiguïsation porte sur des vocables suffisamment polysémiques, ce qui diminue un peu le travail d'annotation. De ce fait, on conviendra que le jeu du Mai est plus long à générer que les activités de recontextualisation.

Le jeu du Mai étant un peu différent des activités de recontextualisation, puisqu'il ne concerne qu'un seul vocable au lieu de plusieurs, les spécifications portant sur l'affichage, l'aide, l'acceptation et la notation ne sont pas les mêmes. Aussi, nous le traiterons à part, au 8.6. Tout ce qui va suivre jusque là concerne donc exclusivement les différentes activités de recontextualisation

2 L'affichage des concordances

A partir de l'ensemble des lexies/vocables (suivant le cas) sélectionnés, le système affiche les concordances et occulte les éléments à retrouver. L'apprenant doit taper la bonne forme dans le champ situé à droite des phrases (figure 8.7). Lorsqu'il valide sa réponse en tapant

return, le système vérifie qu'elle fait bien partie des items de départ et enlève celle-ci de la liste. En cas de modification, ou bien de faute de frappe, l'utilisateur a la possibilité de corriger sa réponse directement dans le champ s'il n'a pas validé, ou bien en cliquant dans le champ avec la touche option enfoncée s'il a validé. Dans ce dernier cas, l'item est replacé dans la liste de départ. Si trop de modifications sont nécessaires, l'utilisateur peut remettre tous les champs à vide en appuyant sur le bouton « Tout effacer ». En appuyant sur le bouton figurant une ampoule éclairée, l'apprenant peut recevoir de l'aide (voir 8.3). Lorsqu'il estime avoir trouvé toutes les réponses, il les valide en totalité avec le bouton « Valider ». Le système met alors en gras celles qui sont justes, établit une note et expose les résultats (voir 8.4). S'il y a des erreurs, l'apprenant peut essayer de les rectifier, jusqu'au sans-faute, mais la note calculée à la première validation reste la même.

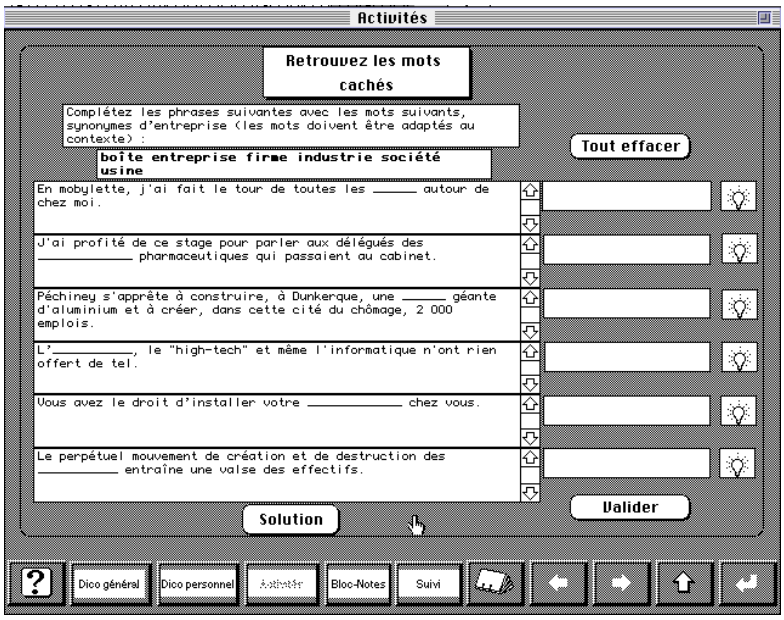


Figure 8.7 : Affichage des concordances dans un exercice de recontextualisation (synonymie)

L'occultation des lexies est facilitée du fait de l'indexation de la forme fléchie avec sa forme canonique dans le fichier des concordances. Il est aussi possible, grâce aux indications de positionnement (en octet) de l'occurrence dans la phrase, de s'affranchir de divers petits traitements préalables pour une analyse de chaque vocable avec l'analyseur morphologique (élimination puis replacement des apostrophes ou des signes de ponctuation). Néanmoins, le problème principal ne se résume pas à ces petites tracasseries informatiques, mais concerne le contexte entourant la lexie à retrouver.

En effet, les concordances affichées sont sélectionnées aléatoirement parmi celles du vocable disponible. Dès lors, le système peut ne pas avoir la main heureuse en retenant des phrases trop courtes, possédant un contexte trop pauvre (comme le choix d'un titre dans le genre Le chômage en France), ou au contraire, des phrases trop longues, dans lesquelles la lexie joue un rôle minime. L'apprenant est alors confronté à une masse de texte inutile. Inutile ? Peut-être pas autant qu'on pourrait le croire. Car c'est peut-être ce contexte qui va lui

permettre de déduire le vocable à retrouver. Dans les deux cas, on s'aperçoit qu'il faut donner à l'apprenant la possibilité de pouvoir gérer le contexte à afficher, en lui permettant de le réduire ou au contraire de l'augmenter, voire carrément de changer la concordance si la phrase s'avère inexploitable. On touche ici le problème crucial de devoir travailler à partir de données réelles et non contrôlées, par rapport à des phrases qui auraient été pensées par un enseignant.

Concrètement, le système affiche un contexte avant et après le vocable de 70 caractères, en s'arrêtant aux limites de la phrases si celles-ci sont atteintes. Au-delà de 70 caractères, la phrase est coupée par défaut. En cliquant sur le champ contenant la concordance, si la phrase initiale est trop longue, le système alterne la phrase affichée par défaut avec la phrase entière. En cliquant sur le champ avec la touche option enfoncée, une nouvelle concordance est générée avec le même vocable.

3 L'aide

L'aide consiste à donner des éléments d'information partiels mais suffisamment significatifs sur la forme ou le sens d'un vocable, de manière à faciliter les inférences et à aiguiller l'apprenant vers la bonne solution. Ces informations sont tirées des connaissances que possède le système, connaissances disséminées en différents endroits : synonymes et définitions (avec le vocable occulté) dans le dictionnaire général, nom du groupe ainsi que les autres éléments en faisant partie pour le dictionnaire personnel, premières lettres ou longueur du vocable. Le registre peut également servir d'aide dans le cas où l'apprenant aurait trouvé le sens effectif de l'item, mais au registre près. Par contre, du fait de leur caractère libre et non contrôlé, les annotations ne peuvent pas servir d'aide : l'apprenant pourrait y trouver des informations, comme une traduction, pouvant le mettre sur la voie immédiatement, sans avoir besoin de faire des inférences.

Les aides doivent être adaptées à l'activité, car il est inutile par exemple de donner les premières lettres d'un vocable à retrouver si celui-ci figure dans l'énoncé (ce serait trop facile) ou bien il est superflu de donner un synonyme dans un exercice portant sur la synonymie (car ils figureront aussi dans l'énoncé). Dès lors, il faut examiner chacune des variantes des activités de recontextualisation et décider quelle aide est appropriée ou non.

- **Activité 1** : Recontextualisation sur vocables non liés (figure 4.1). Le système calcule les vocables du dictionnaire personnel ayant un statut intermédiaire et affiche leurs concordances. Les vocables sont cités dans l'énoncé. L'objectif est de replacer dans les phrases en les fléchissant des vocables qui n'ont pas forcément de liens sémantiques entre eux. Le contexte est donc plus discriminateur, mais les vocables pouvant appartenir à plusieurs champs sémantiques, les connaissances requises sont plus importantes. Les éléments d'aide sont les suivants :
  - synonymie : non, car même si une similitude de sens pourrait faciliter l'inférence, les synonymes proposés n'auront pas forcément le même sens car l'exercice porte sur des vocables et non des lexies.
  - définition : non, pour la même raison que la synonymie

- nom du groupe et autres éléments : oui
- registre : non, car le registre dépend de la lexie et non du vocable (cas de boîte)
- longueur du vocable et première(s) lettre(s) : non, car comme dans tous les exercices où les vocables sont cités dans l'énoncé, ce serait des indices trop immédiats.

On voit ici que peu d'éléments d'aide peuvent être apportés, en raison du manque de désambiguïsation sémantique qui aboutit inévitablement à des contresens, ce qui n'est pas recommandable pédagogiquement. Ce n'est pas le cas avec les autres variantes, qui portent sur des lexies et non sur des vocables, et pour lesquelles les possibilités sont en conséquent plus riches.

- **Activité 2** : recontextualisation sur vocables liés par des synonymies (figure 4.4 et figure 8.7). A partir d'un vocable parmi ceux ayant un statut intermédiaire dans le dictionnaire personnel, le système calcule les lexies synonymes de la lexie dominante et affiche leurs concordances. Les vocables, cités dans l'énoncé, doivent être replacés avec flexion. Ici, la difficulté concerne les nuances respectives des synonymes entre eux, qui sont parfois difficiles à discerner en fonction du contexte.
  - synonymie : non, car le synonyme proposé se trouvera certainement dans l'énoncé
  - définition : oui, car si elle est suffisamment précise, elle permettra peut-être de faire la distinction entre les différentes nuances des vocables du paradigme. Au passage, nous noterons que l'occultation du vocable dans sa définition ne demande pas le recours de l'analyseur morphologique et ne pose pas problème, vu que la place du vocable est codée manuellement dans la base lexicale (sert notamment à mettre en gras l'entrée dans les définitions ; voir par exemple la figure 6.5)
  - nom du groupe et autre éléments : oui, même s'il y a de fortes chances que les vocables synonymes soit classés dans le même groupe. Cette aide peut s'avérer insuffisante.
  - registre : oui, car les indices contextuels indiquant le registre de la phrase peuvent être difficile à discerner (voir première phrase de la figure 4.4)
  - longueur du vocable et première(s) lettre(s) : non, car les vocables sont cités dans l'énoncé.
- **Activité 3** : recontextualisation sur vocables liés par l'actance (figure 4.2). A partir d'un vocable parmi ceux ayant un statut intermédiaire dans le dictionnaire personnel, le système calcule les lexies actants de la lexie dominante et affiche leurs concordances. Les vocables, cités dans l'énoncé, doivent être replacés avec flexion. Cette activité est *a priori* plus facile que celle concernant la synonymie, puisque les vocables sont plus distincts les uns des autres. Néanmoins, l'apprenant doit bien connaître le champ sémantique du vocable de départ.
  - synonymie : oui
  - définition : oui
  - nom du groupe et autres éléments : oui
  - registre : non, les actants partageant *a priori* le même registre

    - longueur du vocable et première(s) lettre(s) : non, car les vocables sont cités dans l'énoncé.
- **Activité 4** : recontextualisation sur dérivés d'une même famille lexicale (figure 4.3). Un vocable est calculé suivant les mêmes caractéristiques que la synonymie et l'actance et le système détermine ses dérivés. Par contre, vu que les vocables sont plus prévisibles que dans les cas des synonymes ou des actants, il n'est pas nécessaire de les faire figurer dans l'énoncé. C'est l'activité orientée le plus vers la production. La difficulté consiste donc dans un premier temps à retrouver les dérivés du vocable de départ et ensuite à les placer dans les phrases avec flexion. Pour cela, l'apprenant pourra s'aider du contexte afin de déterminer les catégories grammaticales des vocables manquants, même si ceci ne sera pas forcément suffisant (deux dérivés peuvent avoir la même catégorie grammaticale).
    - synonymie : oui, même si les vocables à retrouver ont beaucoup de similitude de sens entre eux. Mais cela peu aider à faire la différence entre, par exemple, allocation et allocataire.
    - définition : oui, même si comme plus haut, le problème porte plus sur la forme que sur le sens
    - nom du groupe et autres éléments : oui, sauf s'il s'agit d'un regroupement par forme, auquel cas, les vocables à retrouver auront de grandes chances de s'y trouver et l'exercice sera trop facile à résoudre.
    - registre : non, les dérivés partageant *a priori* le même registre
    - longueur du vocable et première(s) lettre(s) : oui, surtout la longueur du vocable (car les différents vocables à retrouver partagent certainement la même racine, la/les première(s) lettre(s) ne seront pas forcément utiles), les éléments n'étant pas cités dans l'énoncé.

On peut donc résumer les différents types d'aide dans le tableau suivant (figure 8.8) :

	synonyme	définition	groupe	registre	lettres/longueur
Activité 1	N	N	O	N	N
Activité 2	N	O	O ?	O	N
Activité 3	O	O	O	N	N
Activité 4	O	O	O/N	N	O

Figure 8.8 : récapitulatif des différents types d'aide par activité

Les différentes informations sont accessibles dans une fenêtre supplémentaire associée à chacune des concordances où l'utilisateur peut demander l'élément d'aide voulue en fonction de ceux disponibles (figure 8.9).

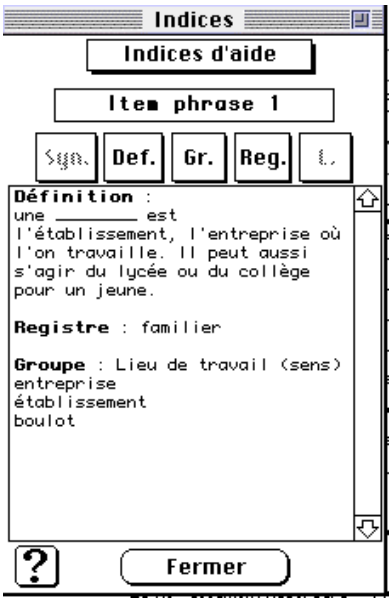


Figure 8.9 : indices d'aide pour l'item de la phrase 1 (boîte, figure 8.7)

Les boutons d'éléments d'aide non disponibles sont désactivés (grisés). Concernant les synonymes, le système en fournit un seul par demande (clic sur le bouton « Syn. »). La priorité sera donné aux quasi-synonymes, puis, s'il n'y en a pas, aux hyperonymes, aux hyponymes et aux synonymes intersectifs. La relation de synonymie est indiquée.

4 L'acceptation

Un des objectifs de ce module d'activités lexicales est de pouvoir adapter son diagnostic en fonction des réponses de l'apprenant. Contrairement aux premiers programmes limités techniquement qui ne pouvaient répondre que par vrai ou faux, les environnements sont maintenant capables de donner des avis plus nuancés. En effet, une réponse n'est pas toujours entièrement fausse et peut se décomposer en plusieurs éléments dont certains s'avèrent justes. C'est le cas par exemple des réponses dont le vocable (ou la lexie) est proche sémantiquement sans être toutefois celui (celle) attendu(e) par le système.

Les réponses partiellement justes sont fonctions des énoncés et donc des activités. Les quatre activités ci-dessus sont concernées par les problèmes de flexion, puisqu'il est demandé à chaque fois d'adapter les vocables aux contextes des différentes phrases. De ce fait, une réponse où le vocable est correct mais mal fléchi ne peut être considérée comme entièrement fausse.

En outre, l'activité 2 sur les synonymes présente quelques spécificités propres à la nature de la relation sémantique en jeu. En effet, comme nous l'avons vu plus haut (4.4), les vocables de départ peuvent être quasi-synonymes et donc s'interchanger entre les phrases. De ce fait, une réponse partiellement juste peut être une réponse :

- où le vocable proposé est sémantiquement proche de celui à deviner (société à la place d'entreprise).
- où le sens est correct mais pas le registre (virer à la place de licencier).

Pour ces deux derniers cas, il est possible, grâce à l'analyseur morphologique et aux réseaux lexicaux du dictionnaire général de savoir si une réponse est proche sémantiquement de celle demandée, en d'autres termes si l'apprenant a répondu par un synonyme à la place de la solution. Il suffit de vérifier, après lemmatisation, que réponse de l'apprenant et solution sont synonymes égaux à ceci près que, l'utilisateur ne pouvant proposer que des vocables et non des lexies, la vérification portera sur les synonymes de toutes les lexies de sa réponse pour voir si parmi elles se trouve la solution.

La problème du registre est un cas particulier de la proximité sémantique vue ci-dessus. Il faudra cette fois vérifier si la réponse de l'apprenant possède un synonyme égal, mais avec un registre différent, qui se trouve être la bonne solution.

Le problème de la bonne (ou non) flexion est plus difficile à trancher, surtout en ce qui concerne les verbes. En l'absence d'analyseur syntaxique dans le système nous ne pouvons compter que sur quatre informations : la forme fléchie et canonique de la bonne réponse et la forme fléchie et canonique (par le biais de l'analyseur morphologique) de la réponse de l'apprenant. Le problème est que même si les deux formes canoniques sont identiques, ce qui indique que l'apprenant a trouvé le bon vocable, mais que sa réponse est fausse (mauvaise flexion), il faut absolument le lui signaler, car il est pédagogiquement peu recommandable de laisser faire des fautes sans au moins les signaler. Or, comment savoir si le vocable est bien fléchi en l'absence d'analyse syntaxique ?

On pourrait penser que la comparaison des deux formes fléchies peut suffire à résoudre ce problème. Ce n'est pas forcément le cas comme le montre l'exemple de la quatrième phrase de la figure 4.3 :

> 4) Mais sur la totalité des entreprises existant en France, seule une sur deux _____ plusieurs salariés.

où le vocable occulté est emploierait dans le texte d'où est extraite cette concordance. Or, en l'absence de contexte plus large, qui n'est d'ailleurs pas traitable par le système, emploit est une réponse tout à fait correcte. De même on peut trouver quantité de phrases où le temps du verbe, ou bien le mode, n'est pas clairement déterminé par les vocables environnants, ce qui autorise plusieurs réponses possibles, toutes plus justes les unes que les autres. La solution adoptée est donc de vérifier que la personne et le nombre coïncident. Dans cet exemple, il suffirait que la solution soit le vocable employer et qu'il soit conjugué à la troisième personne du singulier. Il faut cependant garder à l'esprit que cette vérification n'est pas toujours suffisante. En effet, on peut se trouver confronté à des phénomènes de concordances de temps qui ne permettent pas des flexions à des temps ou des modes approximatifs. Nous espérons cependant que ces cas ne se produiront pas trop souvent. Si ce n'était pas le cas, il faudrait alors se résoudre soit à incorporer un analyseur syntaxique, soit à ne pas demander d'adapter les verbes au contexte en les fléchissant (pour les autres catégories grammaticales, la flexion ne pose pas les mêmes problèmes, puisqu'à chaque fois, une seule solution est possible).

Pour terminer sur ce point, nous abordons les problèmes de fautes de frappe. Il serait dommage d'indiquer comme totalement fausse une réponse contenant une faute de frappe (par exemple boulit à la place de boulot). Dès lors, celles-ci seront indiquées lors de la validation

individuelle de chaque item si elles entrent dans les deux cas suivants : la réponse et la solution diffèrent soit par la permutation de deux lettres ou bien par le remplacement d'une lettre par une autre. Le cas des redoublements de consonnes est plus délicat à traiter, car il peut s'agir soit d'une coquille, soit de l'ignorance de la bonne orthographe. Nous prenons la décision arbitraire de considérer une réponse concernée par cela comme fausse.

5 La notation

Le score obtenu aux activités lexicales doit évidemment tenir compte des réponses partiellement justes suivant les critères déterminés ci-dessus. Il n'est donc pas question de comptabiliser par 1 ou 0 les réponses obtenues mais d'établir un barème en fonction de l'éloignement avec la solution et des possibilités de reconnaissance du système.

La notation est fonction de l'objectif pédagogique de chaque activité. Aussi, voici, activité par activité, des possibilités de barèmes :

- **Activités 1 et 3** : il faut replacer les vocables dans les bonnes phrases en les fléchissant. De ce fait, on peut donner 3 points si la réponse est bonne, 1 point si le vocable est trouvé mais mal fléchi, 0 si le vocable n'est pas trouvé.
- **Activité 2** : il faut tenir compte ici, en plus de la flexion, du fait que les vocables, étant synonymes, peuvent s'interchanger ou correspondre au registre près. De ce fait, on peut donner 3 points par réponse juste ou si un quasi-synonyme a été fourni, 2 points si la réponse est un quasi-synonyme mais avec un mauvais registre, 1 point si le vocable, ou son synonyme, est mal fléchi (0,5 si le registre est incorrect) et 0 point sinon.
- **Activité 4** : les vocables n'étant pas indiqués dans l'énoncé, une partie de l'exercice consiste donc à les produire. De ce fait, on accordera plus d'importance que ci-dessus au fait que le bon vocable ait été fourni. Un barème possible : 3 points pour la bonne réponse, 2 points si le vocable a été trouvé mais mal fléchi et 0 point sinon.

L'utilisation de l'aide rentre également en compte pour la notation. Chaque indice n'a pas la même importance. Ainsi, une définition apparaît comme l'élément pouvant le plus aider. De ce fait, on peut proposer –0,5 point à chaque fois que l'aide est demandée et –1 point s'il s'agit d'une définition.

Les notes concerneront chaque vocable. Elles s'ajouteront à son score dans le module activité lexicale du dictionnaire personnel. Ce barème est pour l'instant indicatif et doit bien sûr être confirmé ou affiné par des données expérimentales qui permettront entre autres de déterminer quelle type d'aide est le plus efficace en fonction de l'activité et quel est celui le plus demandé.

6 Le jeu du Mai

Comme nous l'avons vu plus haut, cette dernière activité est différente des autres car elle ne met pas en jeu un groupe de vocables, mais un seul, décliné dans toutes ses lexies. Dès lors, l'interaction, beaucoup plus simple, diffère quelque peu (figure 8.10) : l'apprenant doit simplement taper le vocable qu'il pense être le bon et valider avec le bouton « Valider ». Comme précédemment, il dispose d'indices d'aide.

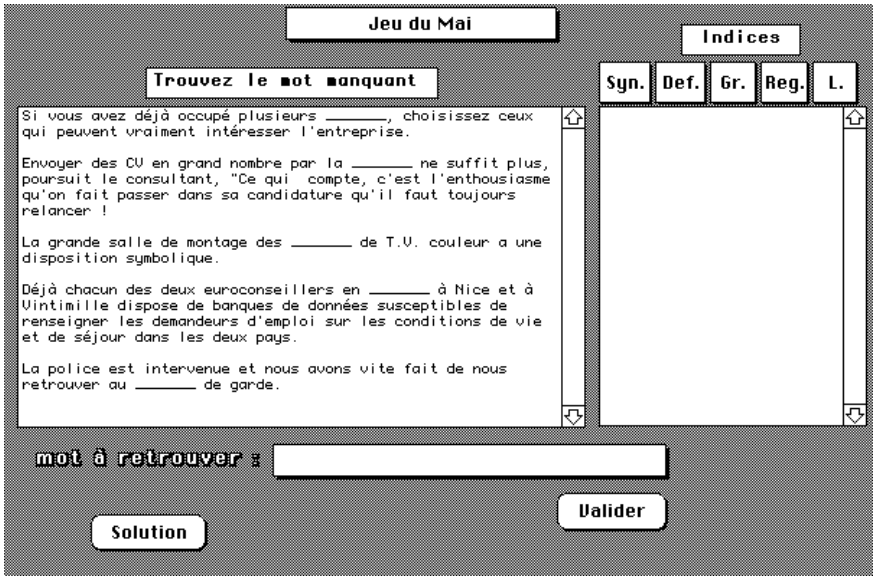


Figure 8.10 : le jeu du Mai (vocable poste)

6.1 L'aide

Tous les indices d'aide sont accessibles pour cet exercice. Cependant, mises à part la longueur du mot et les premières lettres ainsi que les informations propres aux groupes dans le dictionnaire personnel qui concernent le vocable, on peut se demander quel sera le synonyme, la définition ou le registre fourni par le système lorsque l'apprenant demandera cet indice. Ces informations étant liées aux lexies, laquelle choisir ? Comme on ne va pas les afficher toutes en même temps, la seule solution consiste à donner un synonyme ou une définition tiré au sort parmi ceux disponibles. Pédagogiquement, cette solution est pertinente car le synonyme ou la définition sera forcément relié à une des phrases affichées, puisque chaque lexie est concernée. Ceci n'était pas possible dans la première activité de recontextualisation sur des mots non liés sémantiquement : si une définition avait été choisie au hasard, elle pouvait très bien ne pas correspondre du tout à la lexie dans la phrase.

Quant au registre, vu qu'il n'y a pas beaucoup de valeurs possibles (maximum deux), le système peut les afficher d'un coup.

6.2 L'acceptation et la notation

Contrairement aux activités de recontextualisation, il n'est pas nécessaire ici d'adapter le vocable au contexte et donc la flexion n'entre pas en jeu dans l'acceptation. Par contre, il faut tenir compte des proximités sémantiques, notamment des synonymes, même s'il est rare que deux vocables soient synonymes dans toutes leurs acceptions. On pourra donc accepter des synonymes partiels, c'est-à-dire ceux dont certaines lexies ont même signification. Ces réponses pourraient être fournies dans les cas où l'apprenant connaît une partie seulement des lexies.

Par conséquent, un barème possible pourrait être 3 points par bonne réponse, 1 point s'il s'agit d'un synonyme partiel et 0 sinon. L'influence de l'aide est la même que pour les autres exercices, à savoir –1 point pour une définition et –0,5 pour les autres indices. Là encore, cette notation doit être confirmée par des données expérimentales.

7 Conclusion

Nous avons vu dans ce chapitre les caractéristiques et les contraintes des activités lexicales que nous comptons intégrer dans ALEXIA. La première caractéristique, concernant la sélection des items, nous a amenés à parler du dictionnaire personnel de l'apprenant, de son élaboration et de son organisation. Le système attribut des statuts en fonction de la quantité de travail effectué sur chaque item et sélectionne pour les activités ceux qui ont un statut intermédiaire. Une fois cette sélection établie, nous avons constaté que si nous voulions utiliser les relations sémantiques des réseaux lexicaux dans la génération des activités, il nous fallait procéder à une désambiguïsation sémantique. C'est une tâche fastidieuse car elle est manuelle, mais elle peut être réduite de par les caractéristiques du corpus.

Nous avons évoqué ensuite la gestion du contexte dans le cas de concordance trop courte ou trop longue ainsi que le système d'aide qui doit être adapté en fonction de la variante de l'activité en jeu. Enfin, nous avons abordé les problèmes de notation et d'acceptation ou non de réponses partiellement justes, en fonction des connaissances du système et des traitements automatiques qu'il peut effectuer.

De tout ceci, nous retiendrons que la génération automatique d'activités lexicales à partir de matériaux authentiques et bruts demande de considérer un grand nombre de phénomènes dont il n'est pas possible de se rendre compte en composant ces mêmes activités uniquement sur le papier.

CHAPITRE 9

L’environnement ALEXIA

Plusieurs parties de l’environnement ont déjà été étudiées dans les chapitres précédents. Elles ont illustré les réponses que nous apportions aux différents problèmes abordés au cours de ce travail de thèse. Nous nous proposons dans ce chapitre de dresser un récapitulatif du système afin d’avoir une meilleure vue d’ensemble. Nous en profiterons pour décrire deux des modules qui n’ont pas été encore abordés : le module de consultation des textes et le modèle de l’apprenant.

1 Architecture générale

L’environnement ALEXIA est composé de cinq modules, un module de lecture pour le choix et la consultation de textes, un dictionnaire général, un module de construction de dictionnaire personnel, un module d’activités lexicales et un modèle de l’apprenant (figure 9.1). Il comporte en outre deux bases de données, une textuelle et une lexicale. Enfin, il comporte un outil, un analyseur morphologique.

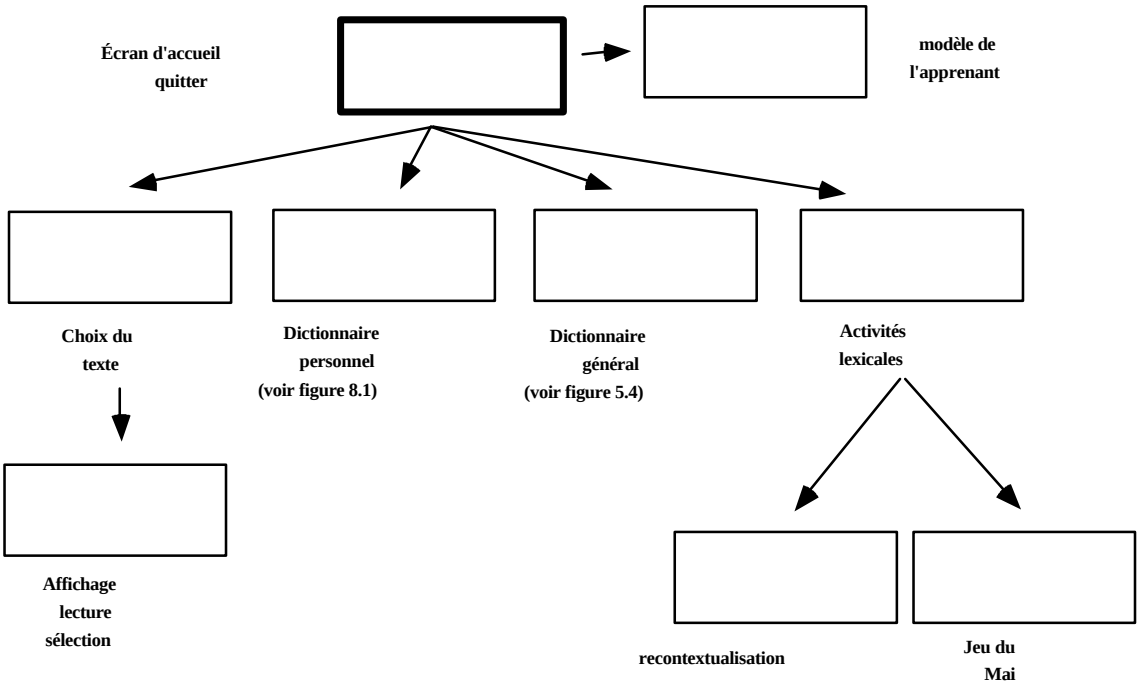


Figure 9.1 : Architecture générale et navigation du système ALEXIA

Trois des cinq modules (le dictionnaire général dans le chapitre 5, le dictionnaire personnel et les activités lexicales dans le 8) ont déjà été abordés. La base de données lexicales et l'analyseur ayant aussi été abordés, nous présenterons ici la base de données textuelles.

2 Base de données textuelles et consultation

La lecture de textes est un moyen naturel pour les apprenants d'être exposés à de nouveaux vocables. Ils tâchent, par différentes stratégies, d'en connaître leur sens et leur fonctionnement syntaxique vis-à-vis des autres vocables pour accroître leur vocabulaire. D'autre part, il faut noter que c'est un moyen direct d'observation de la langue.

Ne voulant pas disperser nos efforts, nous avons choisi d'isoler un champ de langue pour l'étudier plus précisément. Notre préférence s'est portée sur un domaine de la langue courante, par opposition aux domaines de spécialité possédant leur propre terminologie, susceptible d'être maîtrisé par tout natif. N'ayant pas voulu favoriser telle ou telle classe sociale, notre choix s'est porté sur le domaine du travail et de l'activité professionnelle, plus précisément comme il a été dit par ailleurs, sur le champ notionnel du *travail*, de l'*emploi* et du *chômage*. C'est un domaine auquel tout adulte, en principe, est confronté. De par ce choix, l'environnement ALEXIA exclut *a priori* les enfants et il ne peut donc s'agir d'un outil pour l'aide à l'apprentissage de la langue maternelle.

Les corpus du français autres que littéraires (comme la base Frantext, comprenant 180 millions de mots-occurrences issus principalement de la littérature française du 16e siècle à nos jours) récents et libres d'accès n'existent pas. Nous avons donc dû constituer notre propre base textuelle.

L'élaboration du corpus s'est effectuée au sein du LRL et a débuté en 1994 dans le cadre du projet CAMILLE. Le corpus a été ensuite développé pour les besoins d'ALEXIA jusqu'en 1996. Plusieurs membres du LRL y ont participé, parmi lesquels Thierry Chanier, Nathalie Cointe, les CES du laboratoire et moi-même.

Nous avons donc dans un premier temps collecté tout texte portant sur les différents aspects du travail (conditions de travail, formes nouvelles de travail, travail à l'étranger, grèves, etc.), de l'emploi (statistiques, recherche d'emploi, etc.) et du chômage (licenciement, vies de chômeurs, réinsertion professionnelle, etc.). Nous avons voulu être le plus représentatifs possible et avons varié les sources des textes ainsi que leur nature. C'est ainsi que le corpus a été construit à partir d'articles de journaux spécialisés (Les échos) ou généraux (Le Monde, Libération, presse locale), de revues spécialisées (Rebondir, Alternatives économiques, etc.) ou générales (Le Point, l'Express, Télérama, etc.) et (d'extraits) de livres sur le sujet. Nous avons aussi veillé à varier les niveaux de langues en récoltant des textes issus de journaux vendus dans la rue par des associations de SDF (Macadam, Le Lampadaire). Les textes contiennent en outre une partie de langue orale par la présence d'interviews ou de témoignages.

Au total, le corpus est constitué d'environ 400 textes, ce qui représente 450 000 occurrences.

Chaque texte contient, outre son titre et son contenu, les vocables-clés qui identifient le sujet, le nombre de vocables, les références du journal/revue d'où il est extrait, le répertoire dans lequel il a été classé correspondant au type de l'ouvrage et le nom du fichier sous lequel il est enregistré sur le disque.

Ces textes sont tous récents, principalement des années 1994-1996.

L'ensemble des textes a été concaténé en un seul fichier qui a été transmis au serveur SILFIDE (SILFIDE, 1999) afin d'être mis à la disposition de la communauté de chercheurs. En retour, le corpus a été mis sous forme SGML.

3 La consultation et la lecture des textes

L'ensemble des textes a été mis dans une base de données HyperCard sous forme de fiches et accessibles directement depuis ALEXIA.

L'utilisateur peut sélectionner un texte suivant différents critères, qui correspondent en fait aux informations associées à chacun d'entre eux (figure 9.2).

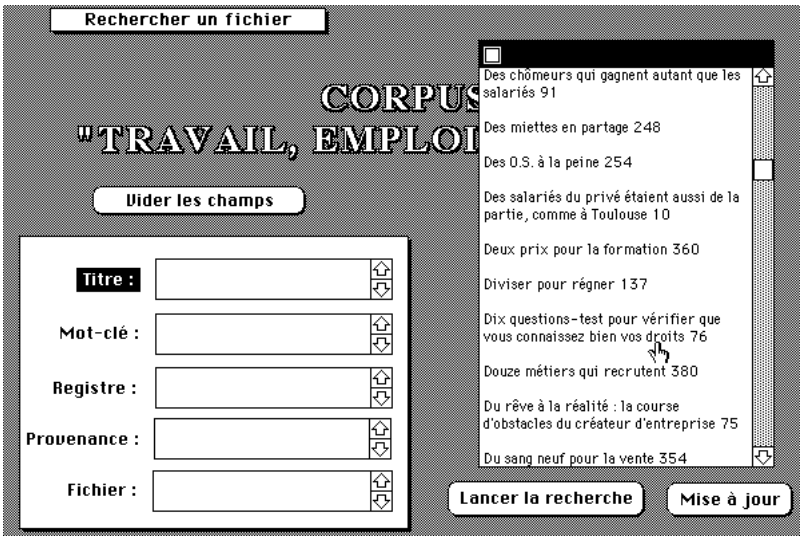


Figure 9.2 : choix d'un texte à partir de différents critères (ici le titre)

Il peut ainsi choisir un texte d'après son titre, d'après son nom de fichier, d'après les vocables-clés, d'après la catégorie de texte et d'après sa provenance. Dans le cas de la sélection par vocable-clé ou par catégorie, comme il peut y avoir plusieurs possibilités (un vocable-clé peut se trouver dans différents textes, tout comme une catégorie regroupe plusieurs textes), une fenêtre demande un choix supplémentaire (figure 9.3).

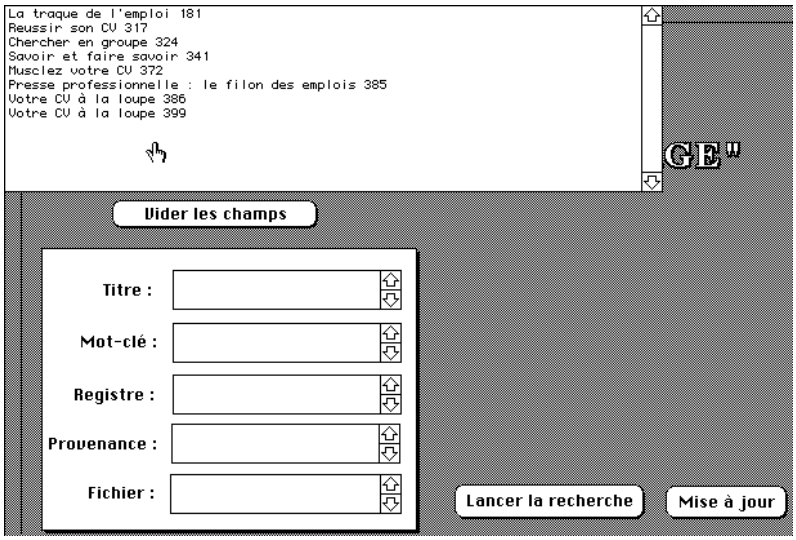


Figure 9.3 : sous-liste des textes ayant cv comme vocable-clé

Une fois le texte sélectionné, le programme cherche l'emplacement du fichier dans le disque dur et l'affiche pour la lecture.

Dans cette carte, l'utilisateur a la possibilité de chercher une occurrence ou bien de retrouver la forme canonique d'une graphie (séparée par deux blancs) en cliquant dessus (la touche option enfoncée de manière à faire la différence avec la sélection de vocables) grâce à l'analyseur morphologique qui recense les différentes possibilités en cas d'ambiguïté (figure 9.4).

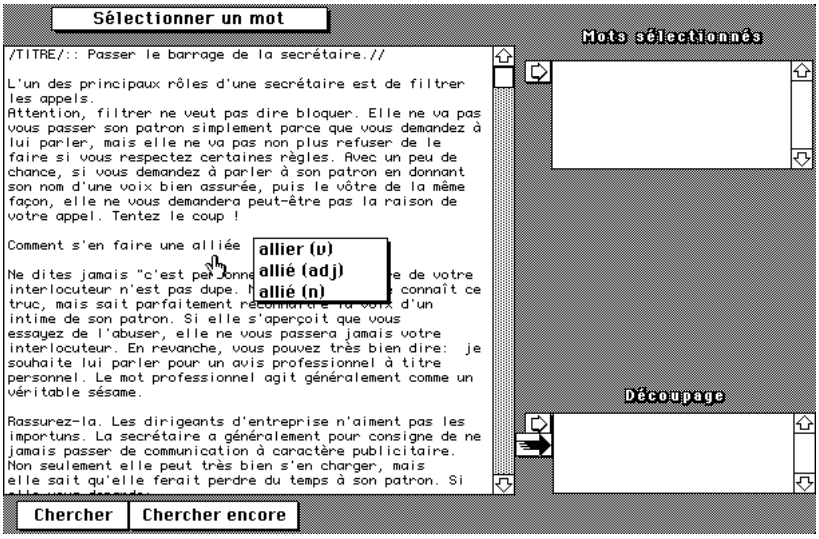


Figure 9.4 : aide de l'analyseur morphologique pendant la lecture

Il a aussi la possibilité de sélectionner des vocables en vue de regroupement dans le dictionnaire personnel comme nous l'avons vu dans le chapitre précédent.

4 Le modèle de l'apprenant

Le module « modèle de l'apprenant » consiste en fait en l'enregistrement de traces des actions qu'effectue l'utilisateur. Il permet de faire une synthèse de l'utilisation du système. Il contient en outre le statut des vocables, correspondant au degré de travail d'un vocable par l'apprenant, données qui sont présentées dans le dictionnaire personnel et dont se sert le système pour déterminer les vocables sur lesquels les activités lexicales sont déclenchées.

Les traces sont organisées par session de travail et enregistrent toute action pertinente effectuée par l'apprenant (figure 9.5): vocable ou lexie consulté, texte choisi, vocables sélectionnés et regroupés, activité effectuée, etc.

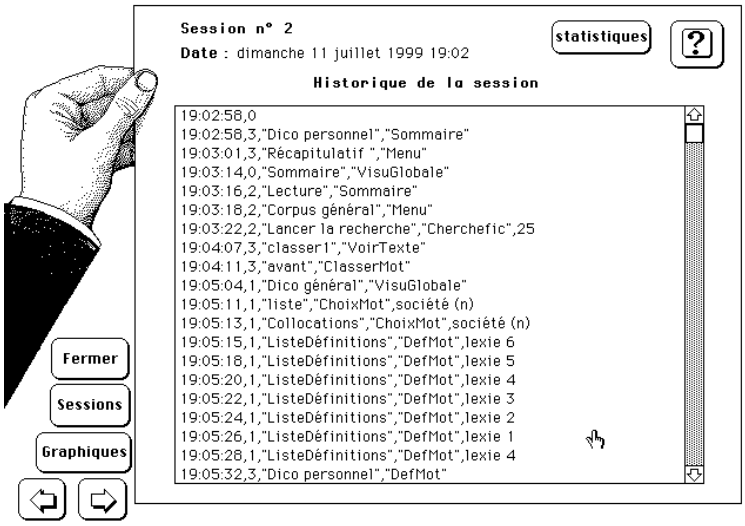


Figure 9.5 : enregistrement de chaque action effectuée

La date (jour et heure) de chaque action est aussi notée ce qui permet de calculer, par différence entre deux dates, le temps passé dans chaque module dans toutes les sessions et d'en donner une synthèse graphique sur la première carte du module (figure 9.6). L'apprenant, ou tout autre personne, peut ainsi se rendre compte de la manière dont est utilisé le système et évaluer si, par exemple, tel ou tel module devrait être plus utilisé.

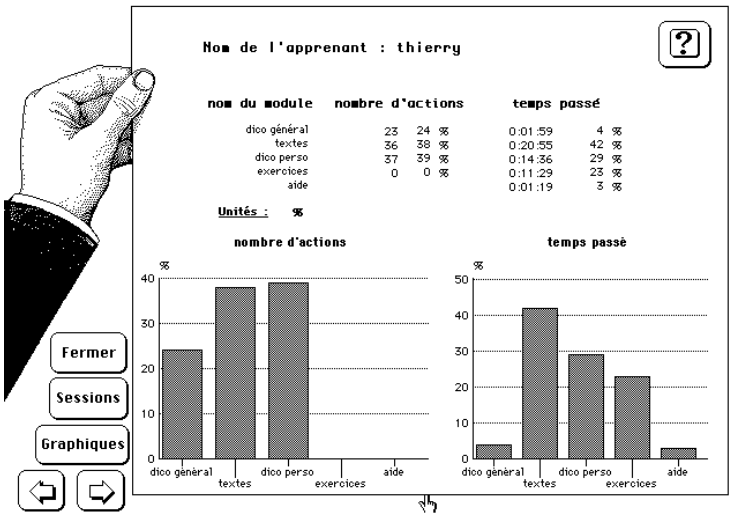


Figure 9.6 : synthèse graphique de l'utilisation du système

CONCLUSION

Le but de ce travail a été de définir et d'étudier les ressources lexicales favorisant un apprentissage en autonomie au sein d'un environnement informatique d'aide à l'apprentissage lexical, ainsi que la manière dont celles-ci s'articulent. Il s'agissait donc d'élaborer un prototype d'environnement destiné à illustrer les problèmes posés par la conception de ces ressources et à montrer les solutions adoptées pour y remédier.

Les modélisations démarquent trois phases lors de l'apprentissage lexical : exposition/compréhension, mémorisation et production/maîtrise du lexique. Elles ont ainsi guidé la conception d'ALEXIA et décidé de ses constituants : un corpus de textes pour l'exposition à de nouveaux vocables, un dictionnaire comme outil d'aide à la compréhension, un dictionnaire personnalisé pour favoriser la rétention et un module d'activités lexicales pour permettre à l'apprenant de s'exercer et d'évaluer ainsi son acquisition. Le schéma type d'utilisation d'ALEXIA commence donc par le choix et la lecture de textes, par l'utilisation du dictionnaire comme outil d'aide, par le regroupement dans le dictionnaire personnel de vocables sélectionnés dans les textes et par des exercices sur ces derniers.

Le corpus d'ALEXIA a répondu à deux exigences. D'une part, il devait circonscrire le cadre d'apprentissage : il n'était en effet pas souhaitable de faire travailler les apprenants sur la langue en général mais au contraire, afin de renforcer l'efficacité de l'apprentissage, de concentrer l'étude sur un domaine suffisamment généralisable. Un champ notionnel du français courant actuel, maîtrisé par tout natif afin de ne pas privilégier telle ou telle classe de la société, a donc été retenu : le domaine du *travail*, de l'*emploi* et du *chômage*. Cette délimitation a aussi été rendue nécessaire par le fait que nous ne pouvions pas préparer de ressources à l'échelle de la langue entière, ressources non disponibles par ailleurs. D'autre part, il devait être représentatif, c'est-à-dire comporter différents registres de langue, d'un registre soutenu comme dans le journal Le Monde à un autre plus familier comme dans les journaux vendus par des sans domicile fixe. Le corpus est à majorité de textes écrits, mais l'oral est tout de même présent à travers des retranscriptions d'interview.

Les résultats des recherches en psycholinguistique montrent que le lexique mental semble être composé de lexies reliées entre elles par des liens de nature sémantique et contextuelle (théorie des toiles verbales). Nous avons tenu compte de ces résultats lors de la modélisation de notre base lexicale en privilégiant une structure sous forme de réseaux pouvant être

accédés en n'importe quel endroit et parcourus. Les problèmes initiaux, en raison des études sur l'utilisation des dictionnaires monolingues, concernaient d'une part l'accès lexical et les stratégies employées par les apprenants pour tirer parti des informations décrites. D'autre part, il s'agissait également de considérer l'évolution que permet le support électronique par rapport au papier au sujet de la nature des informations lexicales et de leur présentation à l'apprenant. L'examen des dictionnaires électroniques existants a montré en effet que ceux-ci ne se différenciaient pas réellement de leur version papier en adoptant, par exemple, la même disposition physique des articles des entrées. Dès lors, nous avons voulu créer un dictionnaire électronique à même de bénéficier des caractéristiques du support électronique et des possibilités de traitements automatiques.

Ainsi, le dictionnaire d'ALEXIA tente de faciliter les problèmes d'accès lexical en mettant l'accent sur les phénomènes d'homonymie et de collocation et sur l'aide à la compréhension des définitions par l'emploi à la fois de définitions abrégées et d'autres plus détaillées et précises, par l'emploi de filtres syntaxiques ainsi que de définitions contextuelles du vocabulaire définitoire. La présentation des articles est aussi abordée par l'adoption d'une disposition régulière et indépendante de l'entrée des différentes informations, par l'affichage simultané, dans le cas de la synonymie, de l'ensemble des lexies synonymes de la lexie de départ, accompagnées de leur définition pour pouvoir distinguer les nuances, par la définition des actants et des dérivés syntaxiques à l'aide d'un schéma syntaxique ainsi que par la présentation des collocations à l'aide de fonctions lexicales provenant de la théorie sens-texte de Mel'cuk. Les possibilités de traitements automatiques, propre aux environnements informatiques, sont clairement illustrées par l'affichage graphique de réseaux lexicaux à base de synonymie, générés automatiquement à partir des informations de la base de données. Ceux-ci ne sont pas statiques mais réagissent aux actions de l'utilisateur lorsque celui-ci clique pour avoir une information sur un des éléments ou désire parcourir le réseau en le faisant se reconstruire sur une autre lexie.

Corpus et dictionnaire électronique ne doivent pas être considérés uniquement sur le plan de la consultation et de l'aide. Ils doivent pouvoir servir de matériaux de base pour générer automatiquement des activités lexicales renouvelables, ce qui est le propre, avec un module d'organisation du vocabulaire personnel, d'un environnement favorisant un apprentissage en autonomie. En fonction de ces matériaux, il s'agissait donc de définir des activités pédagogiques d'un type nouveau, non compatible avec le support papier.

Nous avons choisi des activités de recontextualisation et un exercice proche par la forme : le jeu du Mai. Ces activités consistent, à partir d'un vocable donné, à parcourir la base lexicale pour extraire d'autres vocables, reliés par des relations sémantiques au premier, de générer des concordances et de demander à l'apprenant de retrouver ces vocables, entre-temps masqués. Dans certaines activités, il ne sera pas question de vocables mais de lexies. Elles font travailler les capacités d'inférence de l'apprenant, une des stratégies les plus employées pour l'apprentissage lexical, car il doit s'aider du contexte pour résoudre le problème. D'autre part, elles renforcent les liens établis dans la mémoire de l'apprenant en stimulant les relations sémantiques et les vocables faisant partie du réseau en question. Elles s'appuient donc à la fois sur les réseaux lexicaux de la base de données du dictionnaire et sur le corpus de textes.

Celui-ci doit d'ailleurs être préalablement étiqueté morphologiquement, pour pouvoir travailler non pas à partir de chaînes de caractères (occurrences) mais de vocables, et même sémantiquement, dans une certaine mesure, de manière à pouvoir utiliser les relations sémantiques qui relient les lexies. La génération de concordances peut alors s'opérer à l'aide d'un fichier d'indexation de tous les vocables (ou des lexies suivant le cas) contenus dans les textes.

Les activités de recontextualisation ne se résument pas seulement à la génération automatique de concordances et il est nécessaire de songer à l'interaction et à l'aide que le système peut apporter en fonction des variantes, à la sélection des vocables ou des lexies qui vont être proposés, aux possibilités de correction et d'acceptation des réponses.

Ainsi, les vocables sur lesquels portent les activités sont déterminés par leur statut dans le dictionnaire personnel. Celui-ci est calculé par la quantité de travail effectué dans chacun des modules. Il n'est en effet pas pertinent d'évaluer des connaissances sur des items non travaillés, ni sur des items déjà maîtrisés. L'aide que propose le système est obtenue à partir des différentes informations dans les différents modules : synonymes et définitions du dictionnaire général, membres et noms des groupes du dictionnaire personnel ainsi que les caractéristiques du vocables lui-même.

Concernant le dictionnaire personnel, évoqué ci-dessus, nous nous en sommes tenus aux travaux de Goodfellow qui ont mis en évidence son utilité pour l'organisation et la rétention du vocabulaire personnel. L'apprenant a ainsi la possibilité de l'organiser suivant une structure qui lui est propre et de pouvoir y ajouter des annotations. Le regroupement de vocables possédant une caractéristique commune (par exemple la synonymie) implique un travail de catégorisation et nécessite une bonne compréhension des éléments manipulés. L'effort mental produit favorise ainsi les chances de rétention. En effet, il ne suffit pas d'être exposé une seule fois à un vocable, lors de la lecture d'un texte par exemple, pour le retenir : les liens faibles créés dans le lexique mental par une première exposition doivent être renforcés.

Les perspectives de recherche concernent dans un premier temps l'amélioration de l'existant. Le corpus doit être ainsi étendu de manière à fournir de meilleurs rangs de fréquence pour les vocables qui ne sont pas parmi les plus attestés et à pouvoir observer de plus nombreuses collocations. Concernant le dictionnaire général, outre une extension de la couverture, certaines parties doivent être approfondies, comme par exemple l'antonymie, la présentation des collocations ou des actants. Le dictionnaire personnel n'a pas fait vraiment l'objet de recherches dans ce travail et l'on pourrait se demander si d'autres manières de regrouper le vocabulaire ne sont pas possibles, par exemple des dispositions graphiques de réseaux personnels. Enfin, après l'implémentation des activités définies plus haut, il est tout à fait envisageable d'en incorporer de nouvelles, comme par exemple des exercices à trou graphiques, s'appuyant sur les cartes sémantiques générées par le système.

Un deuxième axe de recherche concerne naturellement l'évaluation de l'environnement avec des apprenants. Ceci permettrait de valider certaines des hypothèses que nous avons posées et d'en formuler d'autres pour pouvoir faire évoluer le système actuel.

BIBLIOGRAPHIE

1 Bibliographie générale

Aitchison J. (1983) : « The mental representation of prefixes », *Osmania Papers in Linguistics* 9-10 (Nirmala Memorial Volume), pp. 61-72.

Aitchison J. (1987) : *Words in the mind*, Oxford, Blackwell.

Alderson J. C. (1984) : « Reading in a foreign language: A reading problem or a language problem? », J. C. Alderson, A. H. Urquhart, London, Longman (Eds.), *Reading in a foreign language*, pp. 1-27.

Allport D. A., Funnell E. (1981) : « Components of the mental lexicon », *Philosophical Transactions of the Royal Society of London* B 295, pp. 397-410.

Anderson R. C., Freebody P. (1981) : « Vocabulary and knowledge », J. T. Gutrie (Ed.), *Comprehension and teaching: Research review*, Newark DE, International Reading Association, pp. 77-117.

Atkins B. T. S., Varantola K. (1997) : « Monitoring dictionary use », *International Journal of Lexicography* 10, n° 1, pp. 1-45.

Atkinson R., Raugh M. (1975) : « An application of the mnemonic keyword method to the acquisition of a russian vocabulary », *Journal of Experimental Psychology - Human Learning and Memory* 104, 125-133.

Ball (1997) : *Concordances and Corpora: Concordancers*, http://www.georgetown.edu/cball/corpora/tutorial3.html.

Bauer, F. S. D., Zaenen, A. (1995) : « Locolex: Translation rolls off your tongue », *actes de la conférence ACH-ALLC'95*, Santa Barbara, CA.

Baxter J. (1980) : « The dictionary and vocabulary behavior : a simple word or a handful », *TESOL Quaterly* 14, pp. 325-336.

Beaudot J. (1992) : *Les fréquences d'utilisation des mots en français écrit contemporain*, Montréal, Les Presses Universitaires de Montréal.

Beck I. L., Perfetti C. A., McKeown M. G. (1982) : « Effects of text construction and instructional procedures for teaching word meaning on comprehension and recall », *Journal of Educational Psychology* 74, pp. 506-521.

Beheydt L. (1987) : « Vocabulary in foreign language teaching methodology », *Dutch Crossing* 32, pp. 3-25.

Béjoint H. (1981) : « The foreign student's use of monolingual English dictionaries : A study of language needs and reference skills », *Applied Linguistics* 2, pp. 207-222.

Béjoint H. (1987) : « The value of the dictionary in vocabulary acquisition », dans A. Cowie (ed.), pp 97-104.

Bénac H. (1982) : *Le Dictionnaire des synonymes*, Paris, Hachette.

Bensoussan M., Sim D., Weiss R. (1984): « The effect of dictionary usage on EFL test performance compare with student and teacher attitudinal expectations », *Reading in a Foreign Language*, 2, pp 262-276.

Bertaud du Chazaud H. (1995) : *Dictionnaire des synonymes*, Les usuels de poche Le Robert, Paris, Le Robert.

Bogaards P. (1988) : « A propos de l'usage du dictionnaire de langue étrangère », *Cahier de Lexicologie* 52, 1988-1, pp. 131-152.

Bogaards P. (1991) : « Dictionnaires pédagogiques et apprentissage du vocabulaire », *Cahiers de Lexicologie* 59, 1991-2, pp 93-107.

Bogaards P. (1994) : *Le Vocabulaire dans l'Apprentissage des Langues Etrangères*, Langues et Apprentissage des Langues, CREDIF, ENS St-Cloud, Hatier/Didier.

Bogaards P. (1995) : « Dictionnaires et compréhension écrite », *Cahiers de Lexicologie* 67, 1995-2, pp. 37-53.

Bogaards (1996) : « Dictionaries for learners of English », *International Journal of Lexicography* 9, 4, pp. 277-320.

Bogaards P. (1998) : « Des dictionnaires au service de l'apprentissage du français langue étrangère », *Cahiers de Lexicologie* 72, 1998-1, pp. 127-167.

Brill E. (1995) : « Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging », *Computational Linguistics*, 21, 4, pp. 543-565.

Browman C. P. (1978) : « Tip of the tongue and slip of the ear: Implications for language processing », *UCLA Working Papers in Phonetics* 42.

Bui, K. P. (1989) « HyperLexicon, a Hypermedia-based Lexicon for Vocabulary Acquisition », dans Goos G., Hartmans J. (eds*), Computer Assisted Learning - 2nd International Conference, ICCAL'89 Proceedings*. Dallas, TX, USA. pp. 14-30.

Catt C. (1992) : « CALL authoring programs and vocabulary development exercises », *Computer Assisted Language Learning* 4 (3).

Chanier T., Colmerauer C, Fouqueré C, Abeillé A., Picard F., Zock M. (1992) : « Modelling lexical phrases acquisition in L2 », *LARS'92 : Second Language Acquisition Research : The state of the art*, Utrecht, Pays-Bas, 19-21 août.

Chanier, T. (1996a) : « Learning a Second Language for Specific Purposes within a Hypermedia Framework », *Computer-Assisted Language Learning (CALL),* vol. 9, 1. pp. 3-43.

Chanier, T. (1996b) : « Evaluation in a project life-cycle: the hypermedia CAMILLE project », *Association for learning technology journal (ALT-J),* vol. 4, 3. pp 54-68.

Chanier T., Selva T. (1998) : « The ALEXIA system: The use of Visual Representation to Enhance Vocabulary Learning », *Computer-Assisted Language Learning* (CALL) 11, 5, pp. 498-521.

Chi M. L. A. (1998) : « Teaching dictionary skills in the classroom », *Actes de huitième congrès international de lexicographie EURALEX'98* (European Association for Lexicography), pp. 565-577.

Chun, D. M. & Plass, J. L. (1997) : « Research on Text Comprehension in Multimedia Environments », *Language Learning & Technology*, vol. 1, 1, http://polyglot.cal.msu.edu/llt. pp. 60-81.

Clarke D. F., Nation I. S. P. (1980) : « Guessing the meaning of words from context: strategies and techniques », *System* 8, pp. 211-220.

Clarke M. (1992) : « Vocabualry learning with and without computers - Some thoughts on a way forward », *Computer Assisted Language Learning* 5 (3), pp. 139-146.

Cobb T. (1997) : « Is there any measurable learning from hands-on concordancing? », *System* 25, 3, pp. 301-315.

Coleman E. B. (1964) : « Supplementary report: On the combination of associative probabilities in linguistics contexts », *Journal of Psychology* 57, pp 95-99.

Collins A. M., Quillian M. R. (1969) : « Retrieval time from semantic memory », *Journal of Verbal Learning and Verbal Behavior* 8, pp 240-247.

Coniam D. (1996) : « Using corpus word frequency data in the automatic generation of english language cloze test », *Actes de Teaching and Language Corpora'96*, Lancaster, pp. 29-43.

Coniam D. (1997) : « A preliminary inquiry into using corpus word frequency data in the automatic generation of english language close tests », *CALICO* 14, 2-4, pp. 15-33.

Cutler A. (1983) : « Lexical complexity and sentence processing », in Flores d'Arcais and Jarvella (1983).

Decoo W. (1993) : « Lexical composition and morpho-syntactic variation in language textbooks - Computer-based approaches as another dimension of CALL, *Computer Assisted Language Learning* 6 (2), pp 123-144.

Deese J. (1965) : *The structure of associations in language and thought*, Baltimore MD, The John Hopkins University Press.

Dokter D. A., Nerbonne J., Schurcks-Grozeva L., Smit P. (1997) : « Glosser-Rug; a user study », http://odur.let.rug.nl/~glosser/Publications/userstudy.ps

El-Bèze M., Spriet T. (1995) : « Intégration de contraintes syntaxiques dans un système d'étiquetage probabiliste », *Traitement Automatique des Langues*, 1,2, pp. 47-66.

Fellbaum C. (1993): « English Verbs as a Semantic Net », article extrait de 5paper.ps, ftp://clarity.princeton.edu/pub/wordnet/5papers.ps.

Fodor J. A. (1981) : *Representations: Philosophical essays on the foundation of cognitive science*, Cambridge MA, MIT Press.

Gairns R., Redman S. (1986) : *Working with words. A guide to teaching and learning vocabulary*, Cambridge, Cambridge University Press.

Gay, G. Mazur, J. (1993) : « The utility of computer tracking tools for user-centered design », *Educational Technology*, 34, 3, pp. 45-59.

Gibbs R. W., Gonzales G. P. (1985) : « Syntactic frozeness in processing and remembering idioms », *Cognition* 20, pp. 243-259.

Gillespie J., Gray G. (1992) : « Hypercard and the development of translation and vocabulary skills », *Computer Assisted Language Learning* 5 (1-2), pp. 3-11.

Goodfellow, R. (1994) *A computer-based strategy for foreign language vocabulary learning*, Unpublished PhD thesis, Institute of Educational Technology, Open University.

Goodfellow, R. (1995) « A Review of Types of Programs for Vocabulary Instruction », *Computer-Assisted Language Learning* 8, 2-3, pp. 205-226.

Goodfellow R., Lamy M.-N. (1997) : « Learning to learn a language - at home or on the Web », *actes d'EUROCALL'97*, Dublin City University, Irlande.

Goodglass H., Baker E. (1976) : « Semantic field, naming and auditory comprehension in aphasia », *Brain and Language* 3, pp. 359-374.

Guillot M.-N., Kenning M.-M. (1994a) : « Electronic Monolingual Dictionaries as Language Learning : a Case Study », *Computers Education*, Vol 23, No 1/2, pp 63-73.

Guillot M.-N., Kenning M.-M. (1994b) : « Le Robert Electronique : a Reassessment of the Case for Dictionary-Based Work », *Computer Assisted Language Learning*, Vol 7, No 3, pp 209-225.

Haastrup K. (1989) : « The learner as word processor », *AILA Review* 6, pp. 34-46.

Habert B., Nazarenko A., Salem A. (1997) : *Les linguistiques de corpus*, Paris, Armand Colin/Masson.

Hammer P., Giauque G. S. (1989) : *The role of cognates in the teaching of French*, New York, Peter Lang.

Hand C. R., Tonkovitch J. D., Aitchison J. (1979) : « Some idiosyncratic strategies utilized by a chronic Broca's aphasia », *Linguistics* 17, pp. 729-759.

Hanson-Smith E. (1993) : « Dancing with concordances », *Computer-Assisted English Language Learning* Journal (CAELL) 4, 2, pp. 40.

Hartmann R. R. K. (1983) : « The bilingual learner's dictionary and its users », *Multilingua*, 2-4, pp. 195-201.

Harvey K., Yuill D. (1997) : « A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing », *Applied Linguistics* 18, 3, pp 253-278.

Hasebrook, J. P., Fezzardi, G. (1996) : « Learning with hypermedia : What users do and how to measure it automatically ». Version longue d'un texte paru dans P Carlson & F Makedon (dir.), Educational multimedia and hypermedia, 1996, *Actes du colloque Ed-Media '96*, Boston, MA, 775.

Haynes M., Baker I. (1993) : « American and Chinese readers learning from lexical familiarization in English texts », T. Huckin, M. Haynes, J. Coady (Eds.), *Second language reading and vocabulary acquisition*, Norwood NJ, Ablex, pp. 130-152.

Herbst T. (1996) : « On the way to the perfect learners' dictionary : a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE », *International Journal of Lexicography* 9, 4, pp. 321-357.

Higgins J., Johns T. (1984) : *Computers in language learning*, Collins Educational.

Hotopf W. H. N. (1980) : « Semantic similarity as a factor in whole-word slips of the tongue », in Fromkin (1980).

Howard D. V., McAndrews M. P., Lasaga M. I. (1981) : « Semantic priming of lexical decisions in young and old adults », *Journal of Gerontology* 36, pp. 707-714.

Hudson R. (1984) : *Invitation to linguistics*, London, Martin Robertson.

Hughes A. (1989) : *Testing for language teachers*, Cambridge, Cambridge University Press.

Hulstijn J. H. (1993) : « When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables », *The Modern Language Journal* 77 (2), pp. 139-147.

Ibrahim A. H., Zalessky M. (1989), « Enquête : l'usage du dictionnaire », *Lexiques*, A. H. Ibrahim (Ed), Paris, Hachette, pp. 24-30.

Ickler T. (1982) : « Ein Wort gibt das andere », *Linguistik und Didaktik* 49-50, pp. 3-17.

Jacoby L. L., Craik F. J. M., Begg J. (1979) : « Effects of decision difficulty on recognition and recall », *Journal of Verbal Learning and Verbal Behavior* 18, pp. 585-600.

Johns T. (1986) : « Micro-concord: a language learner's research tool », *System* 14, 2, pp. 151-162.

Johns T. (1991) : « Should you be persuaded - Two examples of data-driven learning », *English Language Research Journal* 4, pp. 27-45.

Johns T, King P. (eds) (1991) : *Classroom concordancing: English Language Research Journal*, 4, Center for English Language Studies, University of Birmingham.

Kameenui E. JE., Carnine D. W., Fresci R. (1982) : « Effects of text construction and instructional procedures for teaching word meanings on comprehension and recall », *Reading Research Quaterly* 17, pp. 367-388.

Kenning M.-M. (1990) : « Computer assisted Language Learning », *Language Teaching*, Cambridge, Cambridge university Press, pp. 67-76.

Kipfer B. A. (1987) : « Dictionaries and the intermediate student : communicative needs and the development of user reference skills », A. Cowie (ed.), pp. 44-54.

Kukulska-Hulme A. (1988) : « A computerized interactive Vocabulary development system for advanced learners », *System* 16 (2), pp 163-170.

Lakoff G. (1987) : *Women, fire, and dangerous things. What categories reveal about the mind*, Chicago/London, The University of Chicago Press.

Lamy M.-N., Goodfellow R. (1998) : « "Conversations réflexives" dans la classe de langues virtuelle par conférence asynchrone », *Apprentissage des Langues et Systèmes d'Information et de Communication* (ALSIC) 1, 2, pp. 81-99. http://alsic.univ-fcomte.fr/Num2/lamy/defaut.htm

Laufer B. (1989) : « A factor of difficulty in vocabulary learning: Deceptive transparency », *AILA Review* 6 (*Vocabulary acquisition*), pp. 10-20.

Laufer B. (1991) : « How much lexis is necessary for reading comprehension? », P. J. L. Arnaud, H. Béjoint (Eds.), *Vocabulary and applied linguistics*, Basingstoke, Macmillan, pp. 126-132.

Laufer B. (1992) : « Corpus-based versus lexicographer examples in comprehension and production of new words », Tommola H., Varantola K. (eds.), Actes du colloque Euralex'92 Part I, pp. 71-76.

Laufer B. (1997) : « The lexical plight in second language reading: words you don't know, words you think you know and words you can't guess ». J. Coady, T. Huckin (Eds), *Second Language Vocabulary: a rationale for pedagogy*, Cambridge, Cambridge University Press, pp 20-34.

Laufer B., Hadar L. (1997) : « Assessing the effectiveness of monolingual, bilingual and 'bilingualised' dictionaries in the comprehension and production of new words, *The Modern Language Journal* 81, 2, pp. 189-196.

Laurier M, Renié D. (1999) : « Observation des apprentissages par l'analyse de traces informatiques : études de cas avec le cédérom *Camille- L'acte de vente* » , à paraître dans les actes de CREAL99, Université d'Ottawa, Ottawa, mai 1999.

Lecointe J. (1993) : *Dictionnaire des synonymes et des équivalences*, Les Usuels de Poche, Paris, Librairie Générale Française.

Leech G, Candlin C. N. (1986) : *Computers in English language teaching and research*, London, Longman.

Levelt W. J. M. (1989) : *Speaking. From intention to articulation*, Cambridge MA, MIT Press.

Liu, M. & Reed, W. M. (1994) : « Relationship between the learning strategies and learning styles in a hypermedia environment », *Conference of Assocation for Educational Communications and Technology (AECT)*, Nashville, February. 15p

Liu, M. & Reed, W.M. (1995) : « The Effect of hypermedia assisted instruction on second language learning », *Journal of Educational Computing Research*, 12. pp 159-174.

Lomicka, L. L. (1998) : « "To gloss or not to gloss": An investigation of reading comprehension online », *Language Learning & Technology, 1*(2), pp. 41-50 [Disponible à http://polyglot.cal.msu.edu/llt/]

Lyman-Hager M. A., Davis J., Burnett J., Chennault R. (1993) : *Une Vie de Boy*, Interactive Reading in French, Borchardt, Johnson (Eds.), pp. 93.

McCarthy M. (1994) : *Cambridge Word Routes*, Cambridge, Cambridge University Press.

Macé P., Guinard M. (1990) : *Dictionnaire des synonymes*, Paris, Nathan.

Marslen-Wilson W. D., Tyler L. K. (1980) : « The temporal structure of spoken language understanding », *Cognition* 8, pp. 1-71.

Marslen-Wilson W. D., Tyler L. K. (1981) : « Central processes in speech understanding », *Philosophical Transactions of the Royal Society of London* B 295, pp. 317-332.

Meara P. (1989) : « Matrix models of vocabulary acquisition », *AILA Review* 6, pp. 66-74.

Meara P., Jones G. (1988) : « Vocabulary size as a placement indicator », dans Grunwell P. (dir.), *Applied linguistics in society*, London, CILT, pp. 80-87.

Meara P., Jones G. (1989) : *Eurocentres vocabulary test 10 KA*, Eurocentres, Zurich.

Mel'cuk I. (1992) : *Dictionnaire Explicatif et Combinatoire du français contemporain, Recherche lexico-sémantique III*, Montréal, Les Presses de l'Université de Monréal.

Mel'cuk I., Clas A., Polguère A. (1995) : *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, Duculot.

Miller G. & al (1993): « Introduction to WordNet : An On-line lexical Database », article extrait de 5paper.ps, ftp://clarity.princeton.edu/pub/wordnet/5papers.ps.

Milton J., Meara P. (1995) : « How periods abroad affect vocabulary growth in a foreign language », *ITL Review of Applied Linguisitcs* 107/108, pp. 17-34.

Mothe J.C. (1975) : *L'évaluation par les tests dans la classe de français*, Paris, Hachette/Larousse.

Müllich H. (1990) : *« Die Definition ist blöd ! » Herübersetzen mit dem einsprachigen Wörterbuch. Das französische und englische Lernerwörterbuch in der Hand der deutschen Schüler*, Tübingen, Niemeyer.

Murrell G. A., Morton J. (1974) : « Word recognition and morphemic structure », *Journal of Experimental Psychology* 102, pp. 963-968.

Nation I. S. P. (1983a) : *Learning and teaching vocabulary*, Wellington NZ, Victoria University.

Nation I. S. P. (1983b) : « Testing and teaching vocabulary », *Guidelines* 5 (RELC supplement), pp. 12-24.

Nation I. S. P. (1990) : *Learning and teaching vocabulary*, New York, Newbury House.

Nerbonne J., Smit P. (1996) : « Glosser-RuG: in Support of Reading », COLING96, Groningen, The Netherlands.

Nerbonne J., Dokter D., Smit P. (1998) : « Morphological Processing and Computer-Assisted Language Learning », *Computer-Assisted Language Learning* (CALL) 11, 5, pp. 543-559.

Nesi H. (1996) : « For future reference ? Current english learners' dictionaries in electronic form », System 24, 4, pp. 537-546.

Nesi H., Meara P. (1991): « How using dictionary affects performance in multiple-choice EFL tests », *Reading in a Foreign Language*, 8, pp 631-643.

Noblitt, J. S., Bland, S. K. (1991) : « Tracking the learner in computer-aided language learning », dans B. F. Freed (dir.), *Foreign Language Acquisition Research and the Classroom* , pp. 120-132, Lexington, Heath.

Oxford R., Crookall D. (1990) : « Vocabulary learning: Critical analysis of techniques », *Revue TESL du Canada* 7 (2) , pp. 9-30.

Paribakht T. S., Wesche M. (1997) : « Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition », J. Coady, T. Huckin (Eds), *Second Language Vocabulary: a rationale for pedagogy*, Cambridge, Cambridge University Press, pp 174-200.

Paroubek P., Adda G., Mariani J., Rajman M. (1997) : « Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de Parties du Discours pour le Français », *actes des 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF* (FRANCIL), Avignon, France.

Perkins K., Brutten S. R., Pohlmann J. T. (1989) : « First and second reading comprehension », *RELC Journal* 10 (2), pp. 1-9.

Pichon, S. (1996) : *Les relations sémantiques dans l'apprentissage lexical en langue seconde: Visualisation graphique et interaction dans un dictionnaire électronique*, mémoire de DEA, Laboratoire de Recherche sur le Langage, Université Clermont-Ferrand II.

Rabetifia M. (1994) : *Extraction automatique des collocations et des expressions terminologique*, Mémoire de DEA, Université Paris-IV.

Renié D. (1999) : « La trace informatique : un nouvel outil pour le formateur de langue seconde ou étrangère », à paraître dans Duquette, L. et Laurier, M. (Eds.*). L'enseignement-apprentissage de la L2 dans des environments multimédias*, Montréal, Les Éditions Logiques.

Rey A. (1989), « Le français et les dictionnaires aujourd'hui », *Lexiques*, A. H. Ibrahim (Ed), Paris, Hachette, pp. 6-17.

Schouten-van Parreren M. C. (1986) : « Nieuwe perspectieven op de didactiek van de woordenschatverwerving », *Levende Talen* 416, pp. 618-625.

Schweigert W. A., Moates D. R. (1988) : « Familiar idiom comprehension », *Journal of Psycholinguistic Research* 17, pp 281-296.

Sérasset G. (1997) : « Informatisation du Dictionnaire Explicatif et Combinatoire », *Actes de la quatrième conférence annuelle sur le Traitement Automatique du Langage Naturel (TALN'97)*, Grenoble, 12-13 juin, pp. 194-198.

Siegel M., Misselt A. (1984) : « Adaptative feedback and review paradigm for computer-based drills », *Journal of Educational Psychology* 18, pp 294-309.

Silberztein M. (1993) : *Dictionnaires électroniques et analyse automatique de textes : Le système Intex*, Paris, Masson.

Singleton D. (1994) : « Le rôle de la forme et du sens dans le lexique mental en L2 », *Acquisition et Interaction en Langue Étrangère* (AILE) 3, pp. 3-27.

Stahl S. (1983) : « Differential word knowledge and reading comprehension », *Journal of Reading Behaviour*, 15, pp. 33-50.

Stanners R. F., Neiser J. J., Painton S. (1979) : « Memory representation for prefixed word », *Journal of Verbal Learning and Verbal Behavior* 18, pp. 733-743.

Sternberg R. J. (1987) : « Most vocabulary is learned from context », M. G. McKeown, M. E. Curtis (Eds.), pp. 89-105.

Stevens V. (1991) : « Concordance-based vocabulary exercises: a viable alternative to gap-fillers », dans *Classroom concordancing: English Language Research Journal*, 4, Johns T, King P. (eds), pp. 47-63, Center for English Language Studies, University of Birmingham.

Stevens V. (1993) : « Concordances as enhancements to language competence », *TESOL Matters* 2, pp. 6-11.

Sussex R., Cumming G., Cropp S. (1994) : « A tools-based environment for discovery-oriented CALL - Cognitive, pedagogical and ergonomie issues for interactive learning », *Computer Assisted Language Learning* 7 (2), pp. 133-149.

Svenconis, D. J. (1994) : *An investigation into the teaching of second-language vocabulary through semantic mapping in the hypertext environment.* PhD Dissertation. The Catholic University of America, Wahsington, DC.

Svenconis, D. J. & Kerst, S. (1995) : « Investigating the teaching of second-language vocabulary trhough semantic mapping in a hypertext environment », *CALICO*, vol. 12, 2-3. pp 33-57.

Swinney D. A., Cutler A. (1979) : « The access and processing of idiomatic expressions », *Journal of Verbal Learning and Verbal Behavior* 18, pp. 523-534.

Taft M., Forster K. I. (1975) : « Lexical storage and retrieval of prefixed words », *Journal of Verbal Learning and Verbal Behavior* 15, pp. 607-620.

Tomaszczyk J. (1981) : « Issues and developments in bilingual pedagogical lexicography », *Applied Linguistics* 2, pp. 287-296.

Tréville M.-C., Duquette L. (1996) : *Enseigner le vocabulaire en classe de langue*, Paris, Hachette.

Tribble C. (1990) : « Concordancing and an EAP writing programme », *Computer-Assisted English Language Learning* Journal (CAELL) 1, 2, pp. 10-15.

Tripp, S. & Roby, W. (1994) : « The Effects of various information resources on learning from a hypertext bilingual lexicon », *Journal of Research on Computing in Education*, 27. pp 92-103.

Tzoukermann E., Radev D. R. (1996) : « Using word class for part-of-speech disambiguation », dans Ejerhed E., Dagan I. (eds.) : *Fourth Workshop on Very Large Corpora*, Copenhague, Danemark.

Ulijn J. M., Strother J. B. (1990) : « The effect of syntactic simplification on reading EST texts as L1 and L2 », *Journal of Research in Reading* 13, pp. 38-54.

Van Roey J. (1990) : « Work in progress: A parallexicon of English-French faux-amis », M. Snell-Hornby (Ed.), pp. 161-169.

Verlinde S. (1994) : *Dictionnaire contextuel du français économique, Tome B : le commerce - Exercisier*, Louvain, Garant.

Zöfgen (1994) : *Lernerwörtebücher in Theorie und Praxis. Ein Betrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*, Tübingen, Niemeyer.

2 Publications ALEXIA

Chanier T., Colmerauer C, Fouqueré C, Abeillé A., Picard F., Zock M. (1992) : « Modelling lexical phrases acquisition in L2 », *LARS'92 : Second Language Acquisition Research : The state of the art*, Utrecht, Pays-Bas, 19-21 août.

Chanier T., Fouqueré C., Issac F. (1995) : « AlexiA : Un environnement d'aide à l'apprentissage lexical du français langue seconde », *Conférence Environnements Interactifs d'Apprentissage avec Ordinateur* (EIAO'95), pp 79-90, Eyrolles, Paris.

Chanier T., Selva T. (1998) : « The ALEXIA system: The use of Visual Representation to Enhance Vocabulary Learning », *Computer-Assisted Language Learning* (CALL) 11, 5, pp. 498-521.

Issac F. (1997) : *Analyse syntaxique et apprentissage des langues*, thèse de doctorat, Université Paris-Nord.

Selva T., Issac F. (1996) : « Représentation et utilisation de connaissances dans un système d'aide à l'apprentissage lexical », *actes du 2e Colloque Jeunes Chercheurs en Sciences Cognitives*, Giens, pp 192-201.

Selva T., Chanier T. (1997) : « Traitement automatique pour la représentation graphique de réseaux lexicaux en apprentissage des langues », *Actes du colloque FRAnche-Comté Traitement Automatique des Langues* (FRACTAL), Besançon, pp 361-371.

Selva T., Chanier T. (1998) : « Apport de l'informatique pour l'accès lexical dans les dictionnaires pour apprenants : projet Alexia », *EURopean Association for Lexicography* (EURALEX'98), Liège, Belgique, pp. 631-642.

3 Logiciels

Brunet E. (1994) : *Hyperbase* version 2.5, UFR Lettres, Nice, France. http://ilex.cc.kcl.ac.uk/toronto/1001h/HyperBase/hyperbase.html

Cordial (1998) : *Cordial 5*, correcteur grammatical de la langue française, diffusé par Synapse Développement, Toulouse, France. http://www.synapse-fr.com.

Descamps J.-L., Boyzon-Fradet D., Mochet M. A. (1996) : *Le jeu du Mai*, diffusé par le CNDP, Paris. http://www.cndp.fr/lettres/lemai/ljdm.htm.

Lotin P., Pothier M., Chanier T (1996) : *CAMILLE Travailler en France*, progiciel d'apprentissage du français sur objectifs spécifiques, 2 cédéroms (module 1 : A la recherche d'un emploi ; module 2 : L'acte de vente), LRL, Université Clermont 2, diffusé par CLE International, Paris.

http://lib.univ-fcomte.fr/RECHERCHE/P7/Camille/CAMILLE.html.

Miller G. A. (1995) : *WordNet - a Lexical Database for English* , Université de Princeton, NJ. http://www.cogsci.princeton.edu/~wn.

Scott M., Johns T. : *MicroConcord*, Oxford University Press, diffusé par Athelstan, La Jolla, CA. http://www.athelstan.com

Silberztein (1994) : *système Intex*, LADL, Université Paris 7. http://www.ladl.jussieu.fr/intex/index.html

4 Sites Internet

Collins (1999) : *Cobuild Home Page*, consulté en juillet 1999 : http://titania.cobuild.collins.co.uk

GRACE (1999) : *GRACE home page at LIMSI-CNRS*, consulté en août 1999 : http://m17.limsi.fr/TLP/grace/

La Passerelle (1999) : *The Half-Baked Interactive Tests*, consulté en juillet 1999 : http://www.lapasserelle.com/lm/pagespeciales/half.baked/halfbakedtests.index.html.

Longman (1999) : *Longman Dictionaries Home Page*, consulté en juillet 1999 : http://www.awl-elt.com/dictionaries/.

SILFIDE (1999) : *Serveur Interactif pour la Langue Française, son Identité, sa Diffusion et son Etude*, consulté en juillet 1999 : http://www.loria.fr/Projet/Silfide/

WordNet (1999) : WordNet – a Lexical Database for English, consulté en octobre 1999 : http://www.cogsci.princeton.edu/~wn/

ANNEXE A

Liste des vocables du dictionnaire général

En **gras** le vocable principal de la famille lexicale.

activité (n)
être en activité (v)
déborder d'activité (v)
taux d'activité (n)
actif (adj)
actif (n)
vie active (n)
population active (n)
allocation (n)
allocation chômage (n)
allocations familiales (n)
allocation logement (n)
allouer (v)
allocataire (n)
appointements (n)
atelier (n)
besogne (n)
aller vite en besogne (v)
abattre de la besogne (v)
boîte (n)
bosser (n)
bosseur (n)
bosseur (adj)
bosser comme un fou (v)
boulot (n)
petit boulot (n)
du bon boulot (n)
bûcher (v)
bûcher (n)
bûcheur (n)
bureau (n)
cadre (n)
jeune cadre dynamique (n)
encadrer (v)
encadrement (n)
cadrer (v)
charges (n)
charges sociales (n)
charges patronales (n)
charges salariales (n)
chef (n)
chef d'entreprise (n)
chômage (n)
durée du chômage (n)
chiffres du chômage (n)
chômage de longue durée (n)
assurance chômage (n)
chômage technique (n)
mettre au chômage (v)
pointer au chômage (v)
taux de chômage (n)
chômer (v)
ne pas chômer (v)
chômé (adj)
chômeur (n)
chômeur de longue durée (n)
condition (n)
conditions de travail (n)
congé (n)
congédier (v)
contrat (n)
contrat du siècle(n)

contrat à durée déterminée (n)
contrat à durée indéterminée (n)
contrat emploi solidarité (n)
contracter : passer un accord (v)
contracter : attraper (maladie) (v)
contracter : réduire le volume (v)
corvée (n)
coût (n)
à prix coutant (adv)
coûter (v)
coûteux (adj)
croissance (n)
croître (v)
croissant (adj)
croissant (n)
débaucher (v)
débauche (n)
débauché (adj)
débauche (n)
dénicher (v)
dénicheur (n)
destituer (v)
destitution (n)
directeur (n)
directeur (adj)
directeur général (n)
directeur du personnel (n)
diriger (v)
direction (n)
directive (n)
durée (n)
embaucher (v)
embauche (n)
emploi (n)
demandeur d'emploi (n)
offre d'emploi (n)
plein emploi (n)
retrouver un emploi (v)
avoir la tête de l'emploi (v)
employer (v)
employé (n)
employeur (n)
engager (v)
entreprise (n)
libre entreprise (n)
petites et moyennes entreprises (n)
établissement (n)
étudier (v)
étudiant (n)
firme (n)
formation : enseignement (n)
formation : fait de prendre ou d'avoir une certaine forme (n)
formation professionnelle (n)
formation continue (n)
formation sur le tas (n)
gagne-pain (n)
honoraires (n)
impôt (n)
imposer : faire payer (v)
imposer : obliger (v)
indemnité (n)
indemnité de chômage (n)
indemnité de déplacement (n)
indemnité journalière (n)
indemnité de licenciement (n)
indemnité parlementaire (n)
indemniser (v)
indemnisation (n)
industrie (n)
industrie de pointe (n)
industrie lourde (n)
industrie légère (n)
industrialiser (v)
industrialisé (adj)
industriel (n)
industriel (adj)
industriellement (adv)
intérim (n)
intérimaire (adj)
intérimaire (n)
job (n)
licencier (v)
licenciement (n)
licenciement économique (n)
licenciement sec (n)
métier (n)
ouvrier (n)
patron (n)
patronat (n)
personnel (n)
place (n)
poste : endroit (n)
poste : qui traite le courrier (n)
poste : appareil (n)
profession (n)
profession libérale (n)
professionnel (n)
professionnel (adj)
professionnellement (adv)

professionnalisme (n)
professionnaliser (v)
qualification (n)
qualifier (v)
qualifié (adj)
qualificatif (adj)
recruter (v)
recrutement (n)
recruteur (n)
recrue (n)
rémunération (n)
rémunérer (v)
rémunérateur (adj)
renvoyer (v)
revenu (n)
RMI (n)
Rmiste (n)
salaire (n)
salaire brut (n)
salaire net (n)
SMIC (n)
smicard (n)
salarié (n)
salarié (adj)
salariat (n)
salarial (adj)
masse salariale (n)
société (n)
social (adj)
social (n)
travailleur social (n)
climat social (n)
conflits sociaux (n)
protection sociale (n)
sécurité sociale (n)
syndicat (n)
syndical (adj)
travail
marché du travail
travailleur (n)
travailleur (adj)
travaillé (adj)
travailler (v)
travailler au noir
se crever à la tâche
usine (n)

ANNEXE B

Format des entrées lexicales

Le premier argument de chaque prédicat Prolog est l'entrée à décrire. Pour éviter les confusions, il faut distinguer l'entrée polysémique de l'entrée désambiguïsée. Dans le code, l'entrée polysémique est un symbole (cas des prédicats `entree` et `derive`), tandis que l'entrée désambiguïsée est une fonction avec un argument qui est le numéro du sens (fonctions citées dans entree). Tous les autres prédicats ont une entrée désambiguïsée. Dans la description qui suit l'entrée est symbolisée par X.

Mis à part dans les prédicats entree et entreesec, A représente l'expression ou le mot relié à l'entrée. Dans le code, c'est une fonction s'il s'agit d'une entrée décrite ailleurs. Sinon il s'agit d'une liste de mots formant l'expression.

entree

définition : donne les caractéristiques de l'entrée (vocable) propres à tous les sens.

forme : `entree(X,cat:A,gn:B,chaine:C,hom:D,freq:E).`

exemple : `entree(travail,cat:n,gn:1,chaine:"travail",hom:"",freq:496).`

A, catégorie de l'entrée (trait cat), prend les valeurs suivantes : `n` (nom), `v` (verbe), `adj` (adjectif), `adv` (adverbe), `p` (préposition), `det` (déterminant), `pron` (pronom), `prop` (proposition), `cl` (clitique), `np` (nom propre).

B, genre et nombre de l'entrée,nombre. Par convention

1 : masculin, peut se mettre au pluriel (ex : travail)

2 : féminin, peut se mettre au pluriel (ex : activité)

3 : masculin singulier invariable (ex : plein emploi)

4 : féminin singulier invariable (ex : libre entreprise)

5 : masculin pluriel invariable (ex : honoraires)

6 : féminin pluriel invariable (ex : allocations familiales)

7 : masculin ou féminin, qui peut se mettre au pluriel (ex : allocataire)

8 : masculin ou féminin singulier invariable

9 : masculin ou féminin pluriel invariable

10 : singulier masculin, pluriel féminin (ex : amour, délice, orgue).

D, chaîne, indique la graphie du mot

E, chaîne, est une précision sémantique qui permet de distinguer les homonymes

F, nombre (trait freq), indique le nombre d'occurrence de l'entrée dans un corpus d'environ 2 millions de mots.

nc indique un nombre non connu.

entreesec

définition : donne les caractéristiques d'un sens de l'entrée (lexie).

forme :

```
entreesec(X,minidef:A,def:B,formes:C,contraintes:D,var_lex:E,reg:F,exemple:G).
```

exemple : `entreesec(pointchom('1'),mini:"se signaler à l'ANPE",def:"lorsqu'une personne pointe au chômage, elle va à l'ANPE et enregistre son passage, ce qui montre qu'elle cherche vraiment du travail2",formes:[[n1,'V',au,'N2']],contraintes:[['N2',nombre,sing],['N2',det,+]],var_lex:[[au,chômage], [[l,'ANPE']]],reg:2,exemple:[74]).`

A, chaine, est la mini-définition (trait `mini`), c'est-à-dire la définition abrégée du lexie.

B, chaine, définition de l'entrée (trait `def`).

C, liste de listes de catégories (trait `formes`), définit la construction syntaxique de l'entrée en explicitant les différentes constructions possibles. En cas d'ambiguité sur la catégorie (deux mots ont la même catégorie), celle-ci est indicée dans un ordre croissant à partir de 1.

D, liste de listes de trois éléments (trait `contraintes`), définit les contraintes sur la construction ou les variations syntaxiques de l'entrée. Chaque élément de liste est un triplet composé de :

- la catégorie (et par suite le mot) sur laquelle porte la contrainte.
- le trait TAG qui exprime la contrainte
- la valeur du trait TAG, liste de listes de deux éléments, définit les variations lexicales des mots (constituants) de l'entrée. Chaque élément de liste est un doublet composé de :
- la catégorie (et par suite le mot) sur laquelle porte la variation.
- une liste de mots qui peuvent remplacer le mot en question.

exemple d'utilisation de l'entrée en contexte, est un pointeur indexé sur le corpus de textes.

E, liste de liste de deux éléments (trait `var_lex`), indique les variations que peut subir l'entrée. Le premier élément est le mot sur lequel porte les variations, le deuxième une liste de mots que l'on peut substituer au premier. D est une liste de liste car l'entrée peut subir plusieurs variations.

F, registre de l'entrée (trait `reg`), est un nombre allant de 1 à 5.

G, exemple d'utilisation de l'entrée en contexte (trait `exemple`), est une liste de nombre faisant référence à des phrases extraites du corpus, contenues dans un fichier spécifique. La liste est vide s'il n'y a pas d'exemple.

actant

définition :indique les actants de l'entrée.

forme : `actant(X,A,cat:B,act:C)`

A=actant$_B$(x), A est l'actant de type B de x (trait `act`), ou A=B(x) (en remplaçant B par une de ses valeurs)

ex : entreprise=`lieu`(travail))).

Il y a deux « sortes » d'actances : les actances simples, exprimées par des numéros d'argument, ainsi que des actances circonstancielles telles que le lieu, le moyen, le résultat, etc.

B, exprimant la relation d'actance entre A et X, prend les valeurs suivantes :

- 1, 2, 3, etc. s'il s'agit d'une actance simple, c'est-à-dire qu'il s'agit du premier, du deuxième, du troisième, etc... du mot.

- Pour les actances circonstancielles :

- `lieu`, indique le lieu où se déroule l'action

ex : entreprise=`lieu`(travail).

- `instrument`, indique le moyen (inanimé) par lequel s'accomplit l'action

ex : outil=`instrument`(travail).

- `resultat`, indique le résultat de l'action

ex : produit=`resultat`(travail)).

- `temps`, indique le temps pendant lequel se déroule l'action

ex : durée du travail=`temps`(travail)).

C, est une catégorie grammaticale.

cooc

définition : donne une collocation dans laquelle figure l'entrée ainsi que le lien sémantique qui les relie.

forme : `cooc(X,A,sem:B,freq:C).`

B est une fonction lexicale au sens large (trait `sem`), c'est-à-dire qu'en plus des fonctions lexicales définies en tant que telles (comme `magn` ou `oper`$_i$), il peut s'agir d'actants. On a alors la relation A=B(x). De plus, ces fonctions peuvent se composer pour former des relations plus complexes (`magn` (`agent`)). Outre les actants (voir plus haut), ces fonctions sont :

- `magn` (magnitude), indique l'intensification de l'action

ex : travailler d'arrache-pied=`magn`(travailler).

- `bon`, donne une idée de louange

ex : ouvrage de première main=`bon`(ouvrage).

- `singulier`, un élément d'une multitude

ex : bateau=`singulier`(flotte).

- `pluriel`, ensemble régulier d'unité

ex :flotte=`pluriel`(bateau).

- `anti`, donne l'opposé de son argument

ex :travailler au noir=`anti(legal`(travailler)).

- `oper`$_i$, verbe sémantiquement vide qui prend le pronom impersonnel (*il*)ou le nom du i^{ème} actant de la situation C_0 comme son sujet grammatical, et le mot-clé C_0 comme son complément d'objet principal

ex : verser (une allocation)=`oper`$_1$(allocation), toucher (une allocation) = `oper`$_2$(allocation).

- $\texttt{func}_i$, verbe sémantiquement vide qui prend C_0 comme son sujet grammatical et, dans le cas où la situation C_0 à des actants, le nom du $i^{ème}$ actant de C_0 comme son complément d'objet principal

ex : régner=$\texttt{func}_0$(silence))

- `syneq`, `syninter`, `synpe`, `synpl` : synonyme équivalent, intersectif, plus étroit, plus large

- `antoeq`, `antointer`, `antope`, `antopl` : antonyme équivalent, intersectif, plus étroit, plus large

C est un argument optionnel (trait `freq`). Sa présence permet de différencier la co-occurrence de la collocation. Etant donné que la collocation a sa propre entrée dans laquelle figure sa propre fréquence, il est inutile de la préciser ici. La présence de cet argument indique donc une co-occurrence, l'absence une collocation. Il indique le nombre d'occurrence de la co-occurrence dans le corpus de textes. `nc` indique que le nombre n'est pas connu.

derive

définition : donne tous les dérivés morphologiques d'une entrée appartenant au même groupe sémantique.

forme : `derive(X,A,cat:B,arg:C).`

Par convention, on exprimera toujours les dérivés par rapport au verbe du champ sémantique, ce qui a effet de clarifier la représentation en ce qui concerne les arguments.

A est le dérivé

B est la catégorie du dérivé (trait `cat`). Prend les mêmes valeurs que le trait `cat` dans entree.

C, nombre (trait `arg`), définit l'argument avec les conventions suivantes :

- 1 indique le sujet (ou premier actant) de l'action exprimé par le verbe du champ sémantique

ex : travailleur=`derive`(travail,`cat:n,arg:1`)).

- 2 indique le premier complément (ou deuxième actant) de l'action.

- 3 indique le deuxième complément (ou troisième actant) de l'action, etc...

- 0 fait référence à l'entrée elle-même

ex : travail=`derive`(travailler,`cat:n,arg:0`).

semantique

définition : exprime la relation sémantique entre deux entrées.

forme : `semantique(X,A,sem:B).`

x et A sont les deux entrées.

B est la fonction lexicale qui lie les deux entrées : A=B(x), du même type que celles définies dans `cooc`.

ANNEXE C

Détails sur l'implémentation

1 Communication entre applications

L'environnement ALEXIA a été implémenté en plusieurs langages : HyperCard pour l'interfaçage, Prolog pour la base de données lexicales et la génération des graphes et C pour l'analyseur morphologique, l'indexation du corpus et la génération de concordances. Cette distribution hétérogène a toutefois l'avantage de tirer parti des points forts et des facilités de programmation de chaque langage, appropriés pour les différents traitements dans l'environnement.

Cette différenciation n'est toutefois possible que si les possibilités de communication entre applications sont fiables, performantes et faciles à utiliser. Ces conditions ont été remplies sous la plate-forme Macintosh au moyen des AppleEvents. Nous décrivons dans cette annexe les codes de programmation qui ont permis l'échange d'informations entre applications. Les AppleEvents sont toujours envoyés par HyperCard. Il s'agit d'un message dont la valeur doit être traité par Prolog ou C qui retournent un résultat à HyperCard. Il n'y a pas de communication entre Prolog et C.

1.1 HyperCard

Les AppleEvents sont des évènements sous forme de message envoyés par une application à une autre. Ils sont composés de deux attributs, la classe et la valeur. Ces attributs sont fixés par HyperCard à `misc` (miscellaneous) pour la classe et `eval` (evaluation) pour la valeur.

Comme dit plus haut, HyperCard envoie une requête à Prolog ou C et en reçoit le résultat. L'instruction HyperCard est :

`request` *`expression`* `from program id` *`IDApplication`*

ou bien

`request` *`expression`* `from` *`program`*

où *`expression`* est une expression susceptible d'être évaluée par l'application cible, *`IDAplication`* la signature de l'application, c'est-à-dire "SIGM" pour Prolog et *`program`* le nom d'un exécutable (des exécutables indépendants de l'application qui les a créés ont pu être générés par C, ce qui n'a pas été le cas pour Prolog, ne disposant pas de la plate-forme complète à cause de son coût). L'application ou l'exécutable doivent être actifs au moment où l'AppleEvent est envoyé et leur nom doivent figurer dans le Finder. *`expression`* est une chaîne de caractères construite par HyperCard contenant une série d'attributs-valeurs.

Le résultat de la requête est simplement stocké dans la variable locale `it` d'HyperCard. D'après ce que nous avons pu observer, si l'application cible n'est pas active ou, pour une raison ou une autre, ne répond pas, `it` reçoit la valeur `empty`.

1.2 Réception et réponse par Prolog

Le message est reçu par le prédicat `ae_receive/4`.

`ae_receive` est de la forme

```
ae_receive(Classe,Valeur,AppleEvent,Reponse)
```

où `Classe` et `Valeur` sont la classe et la valeur de l'AppleEvent reçu (il faut donc les unifier avec `misc` et `eval` dans notre cas), `AppleEvent` est l'AppleEvent reçu et `Reponse` l'AppleEvent renvoyé.

Ensuite, il faut récupérer les paramètres de l'AppleEvent par le prédicat `ae_get_param/2` qui est de la forme

```
ae_get_param(AppleEvent,Parametres)
```

et une fois le traitement accompli, on renvoie la réponse, qui doit être une chaîne, par `ae_put_param/2` qui est de la forme :

```
ae_put_param(Reponse,Parametres)
```

On passe le retour dans `Parametres`.

Cela donne, pour ALEXIA, le code suivant :

```
ae_receive(misc,eval,AppleEvent,Reply):-
      ae_get_param(AppleEvent,bytes('TEXT',Chaine)),
      traite_message(Chaine,Reponse),
      ae_put_param(Reply,bytes('TEXT',Reponse)).
```

1.3 Réception et réponse par C

Les choses sont moins simples en C où la déclaration et le traitement des AppleEvents sont effectués par une série de fonctions. Cependant, le principe reste le même que pour Prolog, une variable, `string`, reçoit une chaîne de caractères, le traitement est effectué, et la réponse, `rep`, est retournée de même sous forme de caractères (le code, qui ne traite que des AppleEvents, nous a été fourni par Fabrice Issac) :

```
#include <AppleEvents.h>
#include <GestaltEqu.h>

#define NewEventHandlerProc(x) (EventHandlerProcPtr)x

void InitAEStuff(void);
void DoHighLevel(EventRecord *AERecord);
```

```
static pascal OSErr HandleOapp (AEDescList *aevt, AEDescList *reply, long refCon);
static pascal OSErr HandleQuit (AEDescList *aevt, AEDescList *reply, long refCon);
static pascal OSErr HandleOdoc (AEDescList *aevt, AEDescList *reply, long refCon);
static pascal OSErr HandlePdoc (AEDescList *aevt, AEDescList *reply, long refCon);

static pascal OSErr AESimpleHandler(AEDescList *aevt, AEDescList *reply, long
refCon);

/* definition d'un  apple event pour HyperCard */

#define kSimpleEvent 'eval'
#define kSimpleClass 'misc'

EventRecord gERecord;
Boolean gDone;

main()
{

gDone = false;
InitAEStuff();
     do {
        WaitNextEvent(everyEvent, &gERecord, 30, nil);
        switch (gERecord.what) {

            case nullEvent:
                /* no nul processing in this sample */
                break;
            case keyDown:
                 gDone=true;
                /* don't care */
                break;
            case kHighLevelEvent:
                DoHighLevel(&gERecord);
                break;

        }
    } while (gDone != true);

   ExitToShell();
}


void InitAEStuff(void)
{
      OSErr  err;
      long   result;

      err = Gestalt(gestaltAppleEventsAttr, &result);
      if (err == noErr) {
            (void)AEInstallEventHandler(kCoreEventClass, kAEOpenApplication,
NewAEEventHandlerProc(HandleOapp), 0, false);
            (void)AEInstallEventHandler(kCoreEventClass, kAEOpenDocuments,
NewAEEventHandlerProc(HandleOdoc), 0, false);
            (void)AEInstallEventHandler(kCoreEventClass, kAEPrintDocuments,
NewAEEventHandlerProc(HandlePdoc), 0, false);
            (void)AEInstallEventHandler(kCoreEventClass, kAEQuitApplication,
NewAEEventHandlerProc(HandleQuit), 0, false);

            (void)AEInstallEventHandler(kSimpleClass, kSimpleEvent,
NewAEEventHandlerProc(AESimpleHandler), 0, false);
      }
}/* end InitAEStuff */



void DoHighLevel(EventRecord *AERecord)
{
```

```
    AEProcessAppleEvent(AERecord);
}/* end DoHighLevel */


pascal OSErr HandleOapp (AEDescList *aevt, AEDescList *reply, long refCon)
{
#pragma unused (aevt, reply, refCon)
      return noErr;
} /* NotHandled */


pascal OSErr HandleOdoc (AEDescList *aevt, AEDescList *reply, long refCon)
{
      #pragma unused (reply, refCon)
      return noErr;
} /* HandleOdoc */


pascal OSErr HandlePdoc (AEDescList *aevt, AEDescList *reply, long refCon)
{
#pragma unused (aevt, reply, refCon)
      return errAEEventNotHandled;
} /* HandlePdoc */


pascal OSErr HandleQuit (AEDescList *aevt, AEDescList *reply, long refCon)
{
#pragma unused (aevt, reply, refCon)
      gDone = true;
      return noErr;
} /* HandleQuit */


/* les informations qui transitent sont des chaines de caracteres */
/* string    : chaine qui recoit les info d'HyperCard                    */
/* rep       : chaine qui contiendra ce qui est envoye a HyperCard */

pascal OSErr AESimpleHandler(AEDescList *aevt, AEDescList *reply, long refCon)
{
char string[255],rep[32000];
 DescType type_e;
 Size taille_e;

/* pour recevoir 40 caracteres */
 OSErr myErr = AEGetParamPtr(aevt, keyDirectObject, typeChar, &type_e, string, 40,
&taille_e);

/* traitement de string et construction de rep */

/* pour envoyer 32000 caracteres */
 myErr = AEPutParamPtr(reply, keyDirectObject, typeChar, rep, 32000);
 return(noErr);
}
```

2 Particularités de la programmation de l'interface sous HyperCard

HyperCard est un langage d'interface évolué à base d'objets (boutons, champs, cartes, piles) et de scripts dans ces objets qui, par le biais de commandes et de fonctions, s'envoient des messages. Une hiérarchie (figure C.1) est imposée et par défaut, les messages remontent la hiérarchie du bas vers le haut.

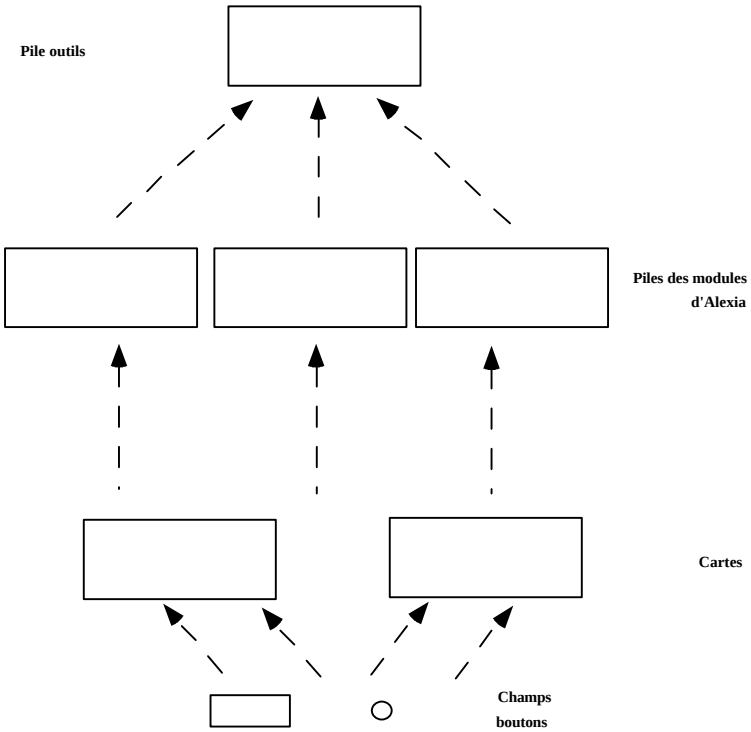


Figure C.1 : Circulation des messages par défaut
dans la hiérarchie HyperCard

Ainsi, si une commande ou une fonction est propre à un objet, elle est programmée dans cet objet (le plus souvent ce sont des commandes locales à des boutons ou de champs). Par contre, si elle agit sur plusieurs objets, elle est programmée au niveau de la hiérarchie le plus bas qui supervise tous les objets. Chaque module d'ALEXIA étant une pile différente, une pile outils contenant toutes les commandes et fonctions générales de l'environnement est donc placée au-dessus des modules et au sommet de la hiérarchie (en fait il y en a deux car, pour des raisons propres au Macintosh, le script associé à une pile, comme à tout autre objet d'ailleurs, ne peut dépasser 32 Ko et cela était insuffisant).

Nous avons veillé à regrouper l'ensemble des variables globales et leur initialisation à un seul endroit, la pile du sommaire de l'application, qui est accédée au lancement de l'environnement. Cette pile contient notamment les valeurs des chemins situant l'ensemble des fichiers sur le disque dur, de manière à faire des modifications en un seul endroit si l'environnement ALEXIA venait à être déplacé.

La navigation dans l'application est assurée par un script commun aux nombreux boutons de l'environnement. Ils envoient un message à une commande dans la pile outils qui s'appuie sur une hiérarchie, sous forme d'arbre binaire, décrite dans un champ de la pile sommaire (figure C.2).

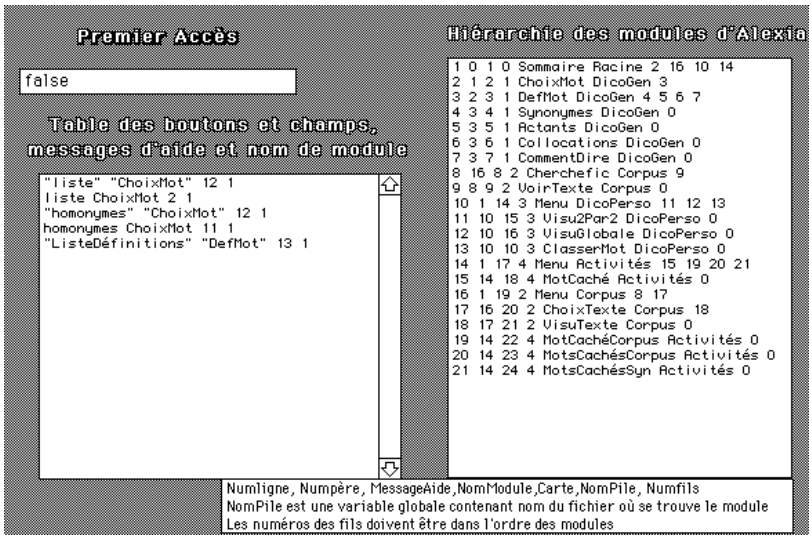


Figure C.2 : hiérarchie des modules d'ALEXIA

Ainsi, si un module devait être ajouté, enlevé ou changé de place, il suffirait de modifier quelques lignes de cette hiérarchie plutôt que le script de tous les boutons concernés.

3 Avantages et désavantages de l'utilisation de Prolog

Comme nous l'avons vu plus haut, Prolog s'est avéré pratique pour la déclaration des données, aussi bien pour la description des vocables et de leur lexies, reliés par un identificateur et dont les informations propres ont été codées dans la structure des faits Prolog, que pour établir les diverses relations entre lexies, les différentes propriétés linguistiques telles que la réciprocité ou la transitivité étant prises en compte par les règles Prolog de parcours du réseau. L'implémentation s'avère performante car le temps de réponse est correct sur un PowerMac 7600/120 correspondant à la technologie de 1996 (instantané pour les synonymes directs, 1 seconde pour l'ensemble des synonymes). Cependant, le dictionnaire ne contenant que 200 vocables et 400 lexies, comment se comporterait-il s'il devait couvrir la langue entière ? Nous ne pouvons pas le savoir *ad hoc*, mais il va sans dire que les temps de réponses devraient s'allonger et que le système demanderait de grosses capacités en mémoire vive, ce qui s'explique par le fait que toutes les données sont dans la mémoire vive (pas d'accès dans des fichiers sur le disque dur).

Nous pouvons avoir une petite illustration de ce qui pourrait se passer avec l'implémentation Prolog de WordNet (WordNet, 1999) qui est disponible gratuitement sur Internet. Le réseau lexical WordNet couvre l'ensemble de la langue anglaise ce qui représente d'après leurs auteurs (Miller *et al.*, 1993, pp. 2) 95 600 vocables différents (51 500 mots simples et 44 100 collocations) organisés en 70 100 *synsets* ou concepts.

Concrètement, les concepts ne sont pas nommés mais représentés par un identificateur dans le réseau Prolog. Ils sont définis par les lexies qui leur sont reliées :

```
s(100342842,1,'job',n,1).
s(100342842,2,'employment',n,2).
s(100342842,3,'work',n,3).
```

Dans cet exemple, le concept `100342842` est défini par les sens (supposés connus) 1 de job, 2 d'employment et 3 de work. Bien que cette organisation nécessite un niveau supplémentaire d'information, il a le mérite de rendre inutiles les règles de transitivité (et de transitivité étendue pour les relations d'hyper/hyponymie) telles que définies dans l'implémentation Prolog du réseau d'ALEXIA.

L'ensemble des fichiers Prolog contenant les données de WordNet occupe près de 15 Mo sur le disque dur dont plus de 5 Mo réservés aux seuls *synsets*. Lorsqu'on exécute le programme sur les synonymes (les différentes relations sémantiques, organisées en différents fichiers, peuvent être compilées indépendamment les unes des autres), le temps de réponse est assez réduit pour les synonymes directs. Par contre, dès que l'on aborde, par exemple, les relations d'hyponymie, celui-ci peut s'allonger considérablement en fonction de l'éloignement des concepts dans le réseau. Signalons en outre que les ressources mémoires sollicitées sont importantes : 20 Mo de mémoire vive rien que pour les synonymes.

Ces caractéristiques sont typiques de Prolog qui ne pratique aucun accès sur le disque dur. Pour pallier cela, WordNet comprend aussi une application exécutable dont les temps de réponse (à partir de données contenues dans des fichiers sur disque dur et non plus en mémoire vive) sont tout à fait corrects quelles que soient les requêtes (y compris par exemple l'hyponymie) et dont les ressources nécessaires en mémoire sont bien plus modestes. Il en résulte une application bien moins lourde et de ce fait plus transportable (WordNet a été implémenté sous Mac et PC ainsi que station Unix).

De tout cela on peut conclure que Prolog est utile pour définir et implémenter une base de données lexicales sous forme de réseau, mais que la consultation et l'utilisation de cette base requière un autre type de langage autorisant un accès séquentiel et des accès disque (C, par exemple). Il faut donc programmer un utilitaire qui permette de générer une sorte de base de données relationnelles à partir de réseaux lexicaux (et vérifier en même temps la cohérence des réseaux).

Index des termes utilisés

Table des matières détaillée